A person looking at a digital screen

AI-generated content may be incorrect.

**Detecting Early Indicators of Mental Health Conditions in Social Media**

*Semantic, Temporal and Ethical Perspectives*

**Data Science Project**

**Zoya Aamir**

*Word Count: 10,098*

Table of Contents

[Abstract 4](#_Toc199510232)

[List of Figures 5](#_Toc199510233)

[List of Tables 6](#_Toc199510234)

[1. Introduction 7](#_Toc199510235)

[1.1 Background of the Study 7](#_Toc199510236)

[1.2 Research Problem 7](#_Toc199510237)

[1.3 Research Aim, Objectives and Questions 8](#_Toc199510238)

[1.3.1 Research Aim 8](#_Toc199510239)

[1.3.2 Primary Objective 8](#_Toc199510240)

[1.3.3 Secondary Objectives 8](#_Toc199510241)

[1.3.4 Research Questions: 8](#_Toc199510242)

[1.4 Research Contribution and Significance 9](#_Toc199510243)

[1.4.1 Academic Contribution 9](#_Toc199510244)

[1.4.2 Societal and Practical Contribution 9](#_Toc199510245)

[1.5 Dissertation Structure 9](#_Toc199510246)

[2. Literature Review 10](#_Toc199510247)

[2.1 Introduction 10](#_Toc199510248)

[2.1.1 Mental Health, Social Media and AI Research 10](#_Toc199510249)

[2.1.2 Key Research Questions 10](#_Toc199510250)

[2.1.3 Research Approach and Structure 10](#_Toc199510251)

[2.2 Theme 1: Mental Health and Social Media 10](#_Toc199510252)

[2.2.1 Psychological Theories Explaining Social Media Impacts 10](#_Toc199510253)

[2.2.2 Types of Social Media Engagement 13](#_Toc199510254)

[2.2.3 Cyberbullying and Online Risks 13](#_Toc199510255)

[2.2.4 Gender and Neurodiversity in Social Media’s Impact 14](#_Toc199510256)

[2.2.5 Identified Research Gaps 14](#_Toc199510257)

[2.3 Theme 2: AI and Mental Health Detection 15](#_Toc199510258)

[2.3.1 The Role of NLP in Mental Health Detection 15](#_Toc199510259)

[2.3.2 NLP Methodologies in Mental Health Classification 17](#_Toc199510260)

[2.3.3 Comparison of Studies and Methodology 18](#_Toc199510261)

[2.3.4 Challenges and Research Gaps 19](#_Toc199510262)

[2.4 Theme 3: Ethical Considerations 20](#_Toc199510263)

[2.4.1 Privacy and Data Protection 20](#_Toc199510264)

[2.4.2 Algorithmic Bias and Fairness 20](#_Toc199510265)

[2.4.3 Explainability and Transparency 20](#_Toc199510266)

[2.4.4 Ethical Implementation of AI 21](#_Toc199510267)

[2.5 Conclusion 22](#_Toc199510268)

[3. Methodology 23](#_Toc199510269)

[3.1 Research Philosophy and Approach 23](#_Toc199510270)

[3.2 Research Design and Strategy 23](#_Toc199510271)

[3.3 Data Collection and Preprocessing 24](#_Toc199510272)

[3.3.1 Sources and Acquisition 24](#_Toc199510273)

[3.3.2 Cleaning and Transformation 25](#_Toc199510274)

[3.4 Model Selection and Justification 26](#_Toc199510275)

[3.5 Training, Testing and Validation 27](#_Toc199510276)

[3.5.1 Classical Models 27](#_Toc199510277)

[3.5.2 MentalBERT 27](#_Toc199510278)

[3.6 Post-Processing and Calibration Layer 27](#_Toc199510279)

[3.7 Cross Platform Generalisability 28](#_Toc199510280)

[3.8 Ethical Considerations 29](#_Toc199510281)

[3.8.1 Best Practices Applied: 29](#_Toc199510282)

[4. Findings and Analysis 31](#_Toc199510283)

[4.1 Platform Signals and Mental Health Context 31](#_Toc199510284)

[4.2 Temporal Patterns and Emotional Timing on Twitter 31](#_Toc199510285)

[4.3 Engagement Without Empathy: The Misreading of Support 34](#_Toc199510286)

[4.4 Lexical Signatures of Distress and Model Interpretability 36](#_Toc199510287)

[4.5 Emotional Overlap and Escalation Paths 39](#_Toc199510288)

[4.6 Reddit vs Twitter and the Training Illusion 41](#_Toc199510289)

[4.7 Summary and Model Design Implications 44](#_Toc199510290)

[5. Evaluation and Discussion 46](#_Toc199510291)

[5.1 Critical Evaluation of Models and Methodology 46](#_Toc199510292)

[5.2 Comparative Model Evaluation 46](#_Toc199510293)

[5.3 Error Analysis: Why Models Fail 46](#_Toc199510294)

[5.4 Domain Shift and Platform Tensions 47](#_Toc199510295)

[5.5 Study Limitations 48](#_Toc199510296)

[5.6 Objective Evaluation Summary 48](#_Toc199510297)

[6. Conclusions and Future Work 49](#_Toc199510298)

[6.1 Core Contribution Summary 49](#_Toc199510299)

[6.2 Practical and Technical Takeaways 49](#_Toc199510300)

[6.3 Ethical Futures: Responsible AI for Distress Detection 49](#_Toc199510301)

[6.4 Future Research and System Redesign 50](#_Toc199510302)

[References 51](#_Toc199510303)

[Appendix A 54](#_Toc199510304)

[Section A: Data Types, Sources and Collection 54](#_Toc199510305)

[Section B: Attributes and Variables 56](#_Toc199510306)

[Appendix B: Project Code Repository (GitHub) 58](#_Toc199510307)

[Appendix C: Power BI Dashboard Snapshots 59](#_Toc199510308)

[Appendix D: Ethics Form A 63](#_Toc199510309)

[Appendix E: Project Proposal 68](#_Toc199510310)

# Abstract

The rise of social media as a site for emotional expression has created new opportunities for early mental health detection using artificial intelligence (AI). This study investigates how natural language processing (NLP) techniques can identify signs of depression, anxiety, and suicidal ideation within social media content. The primary objective was to evaluate the effectiveness of AI models in classifying psychological distress, while also addressing critical ethical challenges.

A domain-adapted transformer model, MentalBERT, was fine-tuned on Reddit data and benchmarked against classical classifiers using both Reddit and Twitter datasets. Comparative performance analysis revealed that MentalBERT significantly outperformed baseline models, achieving a macro F1 score of 0.76 and demonstrating strong sensitivity to subtle and euphemistic expressions of distress. Post-processing strategies, including fairness adjustments and contextual error analysis, were employed to improve ethical alignment.

Ethical considerations were integrated throughout the project, including data anonymisation, explainability mechanisms, and safeguards against algorithmic bias. External validation of the model was conducted by Adil Majeed, a data scientist at Intellytics, who confirmed its technical robustness and ethical viability for use in sensitive domains.

This research contributes a scalable and ethically grounded framework for digital mental health detection, highlighting the importance of context-aware systems in clinical triage and digital health platforms. It advocates for responsible AI that prioritises interpretability, inclusivity, and user dignity in preventative care settings.

# List of Figures

[Figure 1: Code Snippet of Data Cleaning Process 26](#_Toc199509565)

[Figure 2: Post-Processing Code Snippet 28](#_Toc199509566)

[Figure 3: Power BI Temporal Trends Dashboard 32](#_Toc199509567)

[Figure 4: Twitter Density Heatmap 33](#_Toc199509568)

[Figure 5: Power BI Engagement and Support Dashboard 34](#_Toc199509569)

[Figure 6: Emotional Probabilities with and without Engagement 35](#_Toc199509570)

[Figure 7: Sentiment Distribution 35](#_Toc199509571)

[Figure 8: Top TF-IDF Trigrams in Suicidal vs Depressive Tweets 37](#_Toc199509572)

[Figure 9: Radar Chart for Sentiment 37](#_Toc199509573)

[Figure 10: Word Count vs Sentiment 38](#_Toc199509574)

[Figure 11: Label Co-Occurrence Heatmap 39](#_Toc199509575)

[Figure 12: Transition Probability 40](#_Toc199509576)

[Figure 13: Sankey-Style Transition Frequencies 40](#_Toc199509577)

[Figure 14: Word Count Distribution by Label 41](#_Toc199509578)

[Figure 15: Word Count Distribution by Label 42](#_Toc199509579)

[Figure 16: Lorenz Curve 42](#_Toc199509580)

[Figure 17: Top Diverging Terms 43](#_Toc199509581)

[Figure 18: Average Distress Score by Month 59](#_Toc199509582)

[Figure 19: 2016 Temporal Trend 59](#_Toc199509583)

[Figure 20: 2015 Temporal Trend 60](#_Toc199509584)

[Figure 21: Dashboard 1 60](#_Toc199509585)

[Figure 22: Dashboard 2 61](#_Toc199509586)

[Figure 23: Dashboard 3 61](#_Toc199509587)

[Figure 24: Dashboard 4 63](#_Toc199509588)

# List of Tables

[Table 1: Dissertation Structure 9](#_Toc199509537)

[Table 2: Theory Analysis 11](#_Toc199509538)

[Table 3: Comparison of Literature 12](#_Toc199509539)

[Table 4: Gender and Impact 14](#_Toc199509540)

[Table 5: Research Gaps within Literature 14](#_Toc199509541)

[Table 6: Framework Comparisons 16](#_Toc199509542)

[Table 7: Comparison of Studies and Methodologies 18](#_Toc199509543)

[Table 8: Research Gaps and Relevance 19](#_Toc199509544)

[Table 9: Ethical Concerns and Solutions 21](#_Toc199509545)

[Table 10: CRISP-DM Framework Application 23](#_Toc199509546)

[Table 11: Overview of Datasets Used 24](#_Toc199509547)

[Table 12: Overview of Model Characteristics and Justifications for Use 26](#_Toc199509548)

[Table 13: Post-Processing Calibration Tools 28](#_Toc199509549)

[Table 14: Observations and Possible Implications 33](#_Toc199509550)

[Table 15: Observations and Possible Implications 36](#_Toc199509551)

[Table 16: Platform Feature Analysis 44](#_Toc199509552)

[Table 17: Final Observation Summary 45](#_Toc199509553)

[Table 18: Model Performance Comparison 46](#_Toc199509554)

[Table 19: Classification Error Types in Mental Health Detection Models 47](#_Toc199509555)

[Table 20: Research Objectives Status 48](#_Toc199509556)

[Table 21 Future Improvements 50](#_Toc199509557)

[Table 22: Reddit Dataset Description Table 54](#_Toc199509558)

[Table 23: Twitter Depression Dataset Description Table 54](#_Toc199509559)

[Table 24: Behavioural Tweets Dataset 55](#_Toc199509560)

[Table 25: Reddit Attribute Table 56](#_Toc199509561)

[Table 26: Twitter Depression Attributes 56](#_Toc199509562)

[Table 27: Twitter Suicide Attributes 57](#_Toc199509563)

[Table 28: Twitter Behavioural Attributes 57](#_Toc199509564)

# Introduction

****1.1 Background of the Study****

Mental health conditions such as depression, anxiety, and stress-related disorders affect nearly one billion individuals globally (Mental Health Foundation, 2022). These conditions are frequently underdiagnosed and undertreated, particularly among young people, marginalised groups, and those with limited access to care. Early detection is essential for improving outcomes, reducing suicide risk, and alleviating strain on healthcare systems. However, traditional diagnostic methods such as clinical interviews and self-reported questionnaires are increasingly viewed as inadequate. They are time-consuming, resource-intensive, and often miss early signs of distress due to stigma, accessibility issues, and underreporting (Andrade et al., 2014).

As a result, social media sites like Twitter and Reddit have developed into important conduits for mental health signals.. Users often share personal thoughts and emotional struggles in ways that may not surface in clinical environments. Studies have shown that online language patterns, including the frequent use of first-person pronouns, negative emotional words, and fragmented sentence structures, can indicate psychological distress (Guntuku et al., 2019). This led to the development of the discipline of digital mental health, which uses online data and computer techniques to identify risks early.

Advancements in natural language processing, especially through transformer-based models such as BERT, have enabled large-scale, context-aware analysis of unstructured text. These models show improved capacity for detecting complex emotional cues and identifying at-risk individuals based on their online behaviour (Shatte, Hutchinson and Teague, 2019). Despite these strengths, challenges persist. Many models struggle with generalisation across platforms and populations, demonstrate cultural and linguistic bias, and lack transparency in their predictions. These limitations can result in ethical risks, especially when consent is assumed based on data accessibility or when misclassifications go unchecked (Moreno et al., 2013).

To be both effective and responsible, AI systems in mental health must be designed with cultural awareness, interpretability, and ethical safeguards. Beyond predictive performance, such systems must prioritise accountability, respect for user autonomy, and alignment with the complex realities of mental health expression in digital environments.

****1.2 Research Problem****

Although prior work has demonstrated that social media data contains rich signals of psychological distress, many existing AI-based detection tools remain flawed in critical ways. Traditional models such as support vector machines and logistic regression rely heavily on handcrafted features and often perform poorly in recognising nuanced or context-dependent expressions of distress. Even state-of-the-art deep learning models, while improving in accuracy, often rely on generic pre-trained embeddings and datasets that lack diversity and domain specificity, limiting their effectiveness across populations and contexts (Guntuku et al., 2017).

One key limitation of current systems is their inability to distinguish between clinically relevant distress and non-pathological expressions such as venting, sarcasm, or satire. For instance, a user expressing frustration after a long day may be flagged as depressed, while someone expressing suicidal ideation metaphorically may go undetected. This ambiguity is particularly problematic in mental health contexts, where false positives and negatives can lead to serious consequences. Furthermore, most models are trained on Western-centric datasets and fail to capture cultural or linguistic variation in how distress is communicated. This has been shown to reduce classifier performance and raise fairness concerns, particularly across gender and racial groups (Aguirre, Harrigian & Dredze, 2021). Because of this, these models frequently show poor generalisability and have the potential to exacerbate already-existing healthcare inequities.

The moral consequences of implementing such systems present yet another obstacle. Mental health is a highly sensitive domain, and algorithmic decisions must be transparent, explainable, and accountable. Many existing tools function as ‘black boxes’ with little insight into how predictions are made. This opacity limits their usability in clinical settings and erodes trust among users and practitioners. Additionally, few studies have incorporated feedback from mental health professionals or domain experts during model development. Designing AI models that are not just technically sound but also interpretable, culturally sensitive, and morally sound is therefore imperative.

****1.3 Research Aim, Objectives and Questions****

### 1.3.1 Research Aim

To develop and evaluate AI models for detecting early signs of mental health conditions in social media posts, enabling timely interventions and improved care outcomes.

### 1.3.2 Primary Objective

* Identify linguistic and behavioural markers of mental health conditions using social media data.

### 1.3.3 Secondary Objectives

* Develop AI models to classify mental health-related content on social media.
* Evaluate the ethical implications of using personal online data for psychological assessment.
* Recommend ways AI tools can be integrated into preventative mental healthcare services.

### 1.3.4 Research Questions:

1. What linguistic and behavioural markers in social media content are associated with mental health conditions?
2. How can AI models be optimised to detect early signs of mental distress from social media data?
3. What are the ethical considerations in using AI for mental health analysis via social media?
4. How can AI tools be practically implemented in mental healthcare platforms?

****1.4 Research Contribution and Significance****

### 1.4.1 Academic Contribution

MentalBERT, a domain-adapted transformer model trained on mental health-related Reddit content, advances digital mental health by combining advanced natural language processing techniques with ethically informed post-processing methods to mitigate misclassification risks. This research contributes to ongoing discourse in computational psychiatry, responsible AI, and digital ethics by operationalising psychological constructs like depression and anxiety into measurable features for machine learning (Calvo et al., 2020). It also addresses key gaps in existing literature, including the treatment of cultural and linguistic diversity, the interpretability of opaque models, and the role of user agency in digital mental health tools. Additionally, it examines the linguistic expression of psychological distress across platforms such as Reddit and Twitter, an area that remains underexplored in current scholarship.

### 1.4.2 Societal and Practical Contribution

From a societal perspective, the work offers practical insights for digital health platforms, public health policymakers, and mental health organisations integrating AI into their services. By supporting the development of accurate and interpretable models, it aids early identification of at-risk individuals and enables timely, personalised care. The emphasis on ethical principles such as data minimisation, transparency, fairness audits, and user feedback helps ensure responsible system design. Validation by Adil Majeed, Data Scientist at Intellytics, also further affirms its relevance for use in sensitive real-world settings.

****1.5 Dissertation Structure****

The following chapters are structured to reflect both the technical and ethical dimensions of the research:

Table 1: Dissertation Structure

| Chapter | Details |
| --- | --- |
| Chapter 1: Introduction | Establishes the research context, motivation, aim, and contribution. |
| Chapter 2: Literature Review | Critically evaluates existing work on social media mental health signals, AI classification methods, and ethical challenges. |
| Chapter 3: Methodology | Details the dataset, model architecture, ethical review, and evaluation strategy. |
| Chapter 4: Findings and Analysis | Presents and interprets classification results, platform-specific patterns, and key insights. |
| Chapter 5: Evaluation and Discussion | Assesses findings in relation to existing literature, ethical frameworks, and research objectives. |
| Chapter 6: Conclusions and Recommendations | Summarises core contributions, limitations, and directions for future development. |

# Literature Review

## ****2.1 Introduction****

### 2.1.1 Mental Health, Social Media and AI Research

The growing prevalence of mental health disorders, including anxiety and depression, has become a critical global concern. Traditional mental health assessments, reliant on clinical evaluations and self-reports, often fail to detect early psychological distress in daily online interactions (Mansoor & Ansari, 2024). Social media, where users regularly express their emotions and experiences, serves as a rich data source for analysing mental health trends. By integrating artificial intelligence (AI) with social media analytics, researchers can unlock new opportunities for early detection and intervention.

### 2.1.2 Key Research Questions

This review explores three guiding questions: (1) How can AI detect and classify mental health indicators in social media posts? (2) What psychological and linguistic patterns signal distress online? (3) What ethical challenges arise when analysing user-generated content? Supporting the rationale for developing scalable, ethically responsible NLP systems for early mental health detection.

### 2.1.3 Research Approach and Structure

Interdisciplinary frameworks integrating psychological theory, computational AI techniques, and ethical considerations in digital health form the basis of the discussion. It begins by examining the impact of social media on mental well-being, followed by an exploration of AI’s potential to detect mental health signals and the key challenges involved in its implementation.

****2.2 Theme 1: Mental Health and Social Media****

Social media significantly impacts adolescent mental health, offering both support and risks. As people explore identity and peer connections, they become more sensitive to online experiences. While social media fosters social support, it also exposes users to cyberbullying, social comparison, and body dissatisfaction (Karim et al., 2020; Nesi, 2020). Understanding these effects is key to using AI for early mental health detection and intervention.

### 2.2.1 Psychological Theories Explaining Social Media Impacts

#### 2.2.1.1 Problematic Social Media Use (PSMU) and Uses & Gratifications

**The PSMU Theory** associates excessive social media use with compulsive behaviours and mental distress (Shannon et al., 2022) yet lacks clear thresholds differentiating problematic from typical use. **Uses and Gratifications Theory** (Katz et al., 1973) offers an alternative view, suggesting individuals engage with social media for identity exploration, validation, and entertainment. This implies that social media’s impact is shaped not only by usage patterns but also by underlying motivations and susceptibility to addiction.

#### 2.2.1.2 Social Comparison Theory and the Cognitive-Behavioural Model

**Social Comparison Theory** (Festinger, 1954) states that adolescents measure self-worth by comparing themselves to online portrayals, potentially leading to self-esteem issues, anxiety, and body dissatisfaction (Shannon et al., 2022; Karim et al., 2020). However, distinguishing constructive (motivational) from harmful (self-deprecating) comparisons remains a gap in the literature. **The Cognitive-Behavioural Model** (Beck, 1976) further explains that negative self-comparisons reinforce maladaptive thought patterns (e.g., catastrophising, overgeneralisation), contributing to compulsive social media engagement and social withdrawal.

#### 2.2.1.3 Comparison and Critical Analysis of Theories

This theoretical interplay offers a nuanced view of digital influences on adolescent mental health, though each has limitations affecting their explanatory power (see Table 2 below).

Table 2: Theory Analysis

| Psychological Theory | Justification for Relevance | Limitations in Explaining Behaviour |
| --- | --- | --- |
| Problematic Social Media Use (PSMU) | Directly links excessive social media use to psychological distress, relevant for identifying behavioural patterns indicative of mental health risks. | Lacks clarity in distinguishing between beneficial versus harmful levels of engagement, overlooking individual differences such as personality and resilience. |
| Social Comparison Theory | Clearly addresses how adolescents assess self-worth through comparisons online, relevant for understanding body image issues and anxiety. | Fails to capture contextual factors (e.g., cultural influences, offline relationships) affecting individual susceptibility to harmful comparisons. |
| Objectification Theory | Particularly relevant for explaining gender disparities in mental health impacts due to emphasis on appearance-based interactions common on social media. | Limited in scope, does not adequately address why certain adolescents resist or internalise objectification more strongly, overlooking protective psychological factors. |
| Cognitive-Behavioural Model (CBM) | Describes how negative self-perceptions on social media reinforce maladaptive thought patterns (e.g., catastrophising, overgeneralisation), increasing vulnerability to anxiety and depression. | Overlooks external factors (e.g., peer influence, cultural context) that may moderate cognitive distortions, assuming distress arises purely from individual thought processes. |
| Uses and Gratifications Theory (U&G) | Explains why individuals engage with social media, such as identity exploration, validation, and entertainment, helping distinguish intentional vs. compulsive use. | Lacks a clear framework for differentiating harmful overuse from normal engagement, making it difficult to assess mental health risks effectively. |

#### 2.2.1.4 Comparison and Critical Analysis of Literature

As shown below, studies use different methods, each with strengths and limitations. Longitudinal studies (Kelly et al., 2018) offer strong evidence but face self-report bias, while meta-analyses (Shannon et al., 2022; Karim et al., 2020) provide broad insights yet lack consistency. A key limitation is their reliance on observational data, limiting causal inference.

Table 3: Comparison of Literature

| Study | Methodology | Size | Key Findings | Limitations |
| --- | --- | --- | --- | --- |
| Kelly et al. (2018) | Large-scale longitudinal study using multivariate regression and path models. | 10,904 participants | Increased social media use linked to higher depressive symptoms, particularly in girls. Mediating factors include sleep disruption, self-esteem, and online harassment. Tracks social media use overtime. | Self-reported data may introduce bias; causality cannot be established |
| Karim et al. (2020) | Systematic review of 50 studies focusing on the relationship between social media use and mental health | 50 studies | Social media has both positive and negative effects; some users experience emotional support, while others suffer from increased anxiety and depression, detecting real-time emotional shifts | Varied methodologies across studies limit direct comparability |
| Sadagheyani & Tatari (2021) | Review of 50 studies, synthesising evidence on both positive and negative effects of social media | 50 studies | Dual nature of social media facilitates peer support and mental health awareness but also contributes to cyberbullying, body dissatisfaction, and stress | Secondary data reliance introduces potential selection biases |
| Shannon et al. (2022) | Meta-analysis of 18 studies, examining correlations between problematic social media use and mental health outcomes | 9,269 participants | Statistically significant correlation between problematic social media use and depression (r=0.273), anxiety (r=0.348), and stress (r=0.313) Aggregating multiple studies | Does not account for individual differences in engagement patterns |
| Nesi (2020) | Commentary and review-based study analysing both risks and opportunities of social media for adolescent mental health | N/A (Review) | Highlights both positive and negative impacts of social media, including social support, identity exploration, and increased risks of depression and body image concerns | Limited data, relies on secondary sources; does not provide direct evidence |

#### 2.2.1.5 Findings

Kelly et al. (2018), found that over five hours of daily social media use significantly increased depressive symptoms compared to moderate use (1–3 hours). However, the study overlooks individual differences like personality traits and offline support, limiting its ability to isolate social media as a direct cause. While multivariate regression accounted for socioeconomic status, it did not consider psychological moderators such as resilience and pre-existing mental health conditions. Integrating psychological profiling in future research could provide a more comprehensive understanding of these findings.

Karim et al. (2020) and Sadagheyani & Tatari (2021) highlight social media's dual impact, where benefits like social support and identity exploration contrast with risks such as cyberbullying and self-comparison-induced anxiety. They find that passive use (e.g., mindless scrolling) is linked to psychological distress, while active engagement (e.g., meaningful interactions) supports well-being. Shannon et al. (2022) reinforces these findings with a meta-analysis of 18 studies, showing strong correlations between habitual social media use and mental health risks (depression r = 0.273, anxiety r = 0.348, stress r = 0.313). Yet, most studies fail to distinguish adaptive from maladaptive engagement, making it difficult to interpret. This highlights the need for AI-driven behavioural analysis to move beyond self-reported data and objectively classify engagement patterns in mental health research.

### 2.2.2 Types of Social Media Engagement

Building on these findings, the distinction between active and passive social media engagement is crucial in understanding mental health outcomes. Active engagement, involving meaningful interactions such as commenting and content creation, has been linked to greater well-being and social connectivity (Karim et al., 2020)​. In contrast, passive consumption, including prolonged scrolling and observing others’ curated content, correlates with higher levels of anxiety and depressive symptoms​. However, direct comparisons are hampered by different studies' varying definitions of engagement kinds. AI-driven behavioural analysis is a viable method to fill these gaps by objectively categorising engagement patterns and evaluating their effects on mental health.

### 2.2.3 Cyberbullying and Online Risks

Cyberbullying has severe psychological consequences, with victims facing twice the risk of developing depression compared to non-victims (Sadagheyani & Tatari, 2021). Research predominantly focuses on victims, with limited exploration into the motivations behind perpetrators. AI-based behavioural analysis could offer insights into cyberbullying patterns, aiding prevention strategies.

### 2.2.4 Gender and Neurodiversity in Social Media’s Impact

Table 4: Gender and Impact

| Gender | Impact |
| --- | --- |
| Girls | More likely to engage in image-based interactions, making them susceptible to body image concerns and social validation pressures (Karim et al., 2020). |
| Boys | More likely to engage in gaming and networking activities, which have a less direct impact on mental well-being. |

As shown in Table 4, Gender differences are also evident, with girls more prone to body dissatisfaction and validation pressures, while boys experience greater cyberbullying effects (Kelly et al., 2018). Additionally, neurodivergent adolescents (e.g., ADHD, ASD) exhibit distinct engagement patterns that ADHD users may struggle with impulsive use, whereas ASD individuals may rely on online spaces for communication (Williams et al., 2023). This suggests that interventions should be tailored not only by gender but also by cognitive profile.

### 2.2.5 Identified Research Gaps

As previously discussed, while digital platforms offer valuable insights into mental health expression, existing literature reveals ongoing gaps, including the challenge of distinguishing adaptive from maladaptive engagement, the dominance of Western-centric datasets, and limited understanding of behaviours like cyberbullying that can exacerbate psychological distress. Traditional methods, which rely heavily on self-reports and observational data, often overlook real-time expressions of psychological distress and fail to capture the complexity of digital behaviour. These drawbacks highlight the necessity of algorithms that transcend disconnected psychological theories. AI-driven models provide scalable and objective tools for analysing engagement patterns, linguistic signals, and emotional cues on social media.

By translating complex psychological constructs into measurable features, AI can help close these gaps and support early, personalised intervention. These gaps and the potential of AI-driven approaches are summarised in Table 5, as shown below. These challenges highlight the computational methodologies explored in Theme 2.

Table 5: Research Gaps within Literature

| Research Gap No. | Research Gap | Link to Study Objectives | AI Application & Relevance |
| --- | --- | --- | --- |
| 1 | Unclear distinction between adaptive and harmful social media use | Essential for defining clear behavioural criteria for online mental health. | Enables AI models to accurately classify patterns of healthy versus harmful social media interactions. |
| 2 | Limited differentiation in types of social comparison behaviours | Helps identify critical behavioural markers associated with psychological distress. | Enhances accuracy of NLP models by distinguishing harmful social comparisons from positive interactions. |
| 3 | Insufficient understanding of cyberbullying perpetrators' motivations | Supports deeper behavioural understanding crucial for preventative interventions. | Strengthens AI's predictive analytics capability by incorporating motivations behind harmful online behaviours. |
| 4 | Neglect of individual psychological moderators (personality, resilience) | Enables personalised assessment, enhancing the effectiveness of interventions. | Improves AI precision and adaptability by accounting for individual variations affecting mental health outcomes. |

## ****2.3 Theme 2: AI and Mental Health Detection****

The need for early mental health intervention has accelerated AI adoption in detecting distress signals from social media (Teferra et al., 2024). Computational Psychiatry (Montague et al., 2012) models mental health through AI-driven behavioural analysis, offering a scalable alternative to traditional diagnosis. However, as noted earlier, challenges such as data imbalance, ethical concerns, and limited contextual understanding persist (Malgaroli et al., 2023).

### 2.3.1 The Role of NLP in Mental Health Detection

#### 2.3.1.1 Sentiment Analysis and Psychological Linguistics

Sentiment analysis, a widely used NLP technique, classifies text as positive, negative, or neutral based on linguistic cues. LIWC-based models (Pennebaker et al., 2001) detect depressive tendencies through language patterns but struggle with sarcasm and contextual sentiment shifts (Malgaroli et al., 2023). Recent advancements have attempted to incorporate context-aware models, such as Hierarchical Attention Networks (HANs), which improve sentiment detection by considering conversational context (Adel et al., 2024). However, sarcasm detection and cultural variations in sentiment expression remain major limitations.

#### 2.3.1.2 Machine Learning and Transformer-Based Models

Traditional natural language processing (NLP) models for mental health detection primarily rely on supervised machine learning techniques, such as Support Vector Machines (SVMs), Random Forest, Decision Trees, and Naïve Bayes classifiers (Teferra et al., 2024; Cao et al., 2024; Owen et al., 2024). While these models have demonstrated effectiveness in structured datasets, they struggle with the unstructured, context-rich, and noisy nature of social media content—one of the most dynamic sources for mental health signals (Malgaroli et al., 2023). Their dependence on static feature engineering also restricts adaptability to evolving linguistic trends, making them less suited for capturing subtle, nuanced expressions of psychological states. This limitation directly reflects Research Gap 1, where traditional models fail to reliably distinguish between adaptive versus maladaptive distress signals in natural discourse.

The emergence of Transformer-based models such as BERT and RoBERTa has significantly advanced the field by addressing many of these challenges. By using self-attention mechanisms, these models can capture long-range dependencies and subtle semantic nuances in text, leading to improved detection of sentiment, intent, and psychological distress (Adel et al., 2024). For example, BERT has achieved classification accuracies of 91% for six specific mental health conditions and 87.85% for broader disorder categories, outperforming traditional machine learning methods (Adel et al., 2024). These results highlight the value of context-aware learning, particularly when dealing with figurative language, idiomatic expressions, and emotional subtext, which are often present in mental health discourse.

Though, despite their technical superiority, Transformer-based models come with notable challenges. Their high computational demands make real-time application difficult, and their reliance on large, labelled, predominantly Western-centric datasets raises concerns about fairness and cultural generalisability. This relates directly to Research Gap 2, where models trained primarily on English-language or Western datasets risk underperforming in non-Western or multilingual settings, potentially leading to biased or inaccurate assessments.

To overcome these restrictions, recent studies have investigated multimodal AI systems that combine textual, visual, and behavioural cues. This integration enhances the accuracy and robustness of distress detection by going beyond linguistic data alone (Owen et al., 2024). Such approaches show particular promise in identifying emotional ambivalence and subtle affective states that text-only models may overlook. However, significant challenges persist, especially in relation to dataset bias, privacy concerns, and the lack of transparency in deep learning models.

These issues have led to increasing interest in Explainable AI (XAI) frameworks and culturally adaptive NLP architectures aimed at improving trust and promoting equity in mental health applications (Malgaroli et al., 2023; Teferra et al., 2024). Moving forward, research must focus on refining hybrid models and diversifying datasets to support reliable, context-aware predictions in real-world settings. This is essential for improving both the generalisability and interpretability required for successful clinical adoption.

Table 6: Framework Comparisons

| Framework | Strengths | Limitations |
| --- | --- | --- |
| LIWC & Sentiment Analysis) | Strong empirical foundation, interpretable results | Lacks contextual awareness, struggles with sarcasm and slang​ |
| SVM & Traditional ML | Effective with structured datasets, requires fewer computational resources | Poor performance on social media data, relies on manual feature extraction​ |
| Transformer-Based Models | Context-aware, high accuracy in classification | Computationally expensive, requires large labelled datasets​ |

A comparative analysis highlights the strengths and limitations of AI in mental health detection (Table 5). Traditional models like SVMs and Decision Trees are efficient but struggle with unstructured social media data, requiring manual feature engineering (Cao et al., 2024). In contrast, Transformer-based deep learning models achieve higher accuracy in sentiment classification (Adel et al., 2024).

### 2.3.2 NLP Methodologies in Mental Health Classification

#### 2.3.2.1 Feature Engineering and Linguistic Markers

Many NLP-based mental health detection models rely on feature engineering, where linguistic, temporal, and contextual attributes are extracted from text. Common features include:

* **Lexical features** (word frequency, sentiment polarity).
* **Syntactic structures** (sentence complexity, pronoun usage).
* **Temporal patterns** (post frequency, engagement levels).

Malgaroli et al. (2023) argue that lexical markers contribute more to classification accuracy than audio or visual cues​. Nonetheless, feature engineering is often domain-specific, requiring manual refinement and extensive preprocessing to remove noisy, irrelevant content.

#### 2.3.2.2 Deep Learning for Text Classification

Deep learning has transformed mental health detection by automating feature extraction and analysing complex linguistic patterns beyond traditional machine learning. Unlike SVMs, Random Forest, and Naïve Bayes classifiers, which rely on handcrafted features, deep learning models learn contextual dependencies directly from data, making them highly effective for processing unstructured social media text (Adel et al., 2024).

Transformer-based models, such as BERT and RoBERTa, outperform earlier models by leveraging self-attention to capture long-range dependencies (Cao et al., 2024). BERT achieves state-of-the-art accuracy, reaching 91% across six mental health conditions and 87.85% for eleven disorders (Adel et al., 2024). Their ability to detect nuanced sentiment shifts enhances distress detection beyond conventional methods.

Deep learning models continue to face challenges such as overfitting, particularly when working with imbalanced datasets in which depressive language is underrepresented, leading to classification biases (Cao et al., 2024). To counteract this, researchers have implemented techniques like weighted cross-entropy and focal loss functions to better handle class imbalances (Teferra et al., 2024). Despite these improvements, Transformer-based models still demand substantial computational resources and large volumes of labelled data, which limits their scalability in real-time clinical settings (Owen et al., 2024).

Recent studies have explored multimodal AI systems that integrate text analysis with image recognition and behavioural tracking to improve classification accuracy and minimise bias (Adel et al., 2024). Additionally, Hierarchical Attention Networks (HANs), which account for conversational context, have demonstrated improved performance over traditional Transformers in mental health classification tasks (Owen et al., 2024).

These advancements mark a shift toward context-aware models capable of integrating multimodal and longitudinal data for more nuanced detection. While such progress improves diagnostic precision, it also introduces pressing ethical concerns related to transparency, algorithmic bias, and user privacy (Malgaroli et al., 2023). The reliance on social media data further complicates issues of consent, explainability, and accountability in AI-driven mental health care (Teferra et al., 2024). Addressing these ethical challenges is essential for ensuring the responsible deployment of AI technologies (see Theme 3: Ethical Considerations).

### 2.3.3 Comparison of Studies and Methodology

Table 7: Comparison of Studies and Methodologies

| Study | Methodology | Key Findings | Limitations |
| --- | --- | --- | --- |
| Adel et al. (2024) | Deep learning framework using pre-trained BERT models for classifying mental health disorders from Reddit posts. | BERT models achieved 91% accuracy for six disorders and 87.85% for 11 disorders; addressed data imbalance with weighted cross-entropy. | Overfitting due to dataset imbalance; lacks real-time validation in clinical settings. |
| Owen et al. (2024) | Narrative review of AI-driven methods for mental health detection on social media, focusing on NLP advancements. | Multimodal AI approaches integrating text, voice, and image data offer improved accuracy over text-only models. | Scarcity of large, high-quality datasets; ethical challenges remain unresolved. |
| Teferra et al. (2024) | Systematic review of NLP-based depression screening methods, highlighting ethical concerns and biases. | Transformer models and linguistic markers improve depression detection but require culturally sensitive adaptations. | Privacy concerns, bias in training data, and lack of interpretability hinder practical applications. |
| Malgaroli et al. (2023) | Systematic review evaluating NLPs role in mental health interventions, emphasizing model accuracy and limitations. | Text-based features contribute more to classification accuracy than audio/video markers; dataset limitations impact model generalisability. | Limited linguistic diversity in datasets; reproducibility challenges affect clinical reliability. |
| Cao et al. (2024) | Meta-analysis of AI applications in mental health, addressing dataset bias and cross-cultural applicability. | Highlights both positive and negative impacts of social media, including social support, identity exploration, and increased risks of depression and body image concerns | Predominance of English-language datasets restrict global applicability; requires more diverse data sources. |

### 2.3.4 Challenges and Research Gaps

Despite advances in NLP-based mental health classification, several challenges remain:

Table 8: Research Gaps and Relevance

| Research Gap | AI Application & Relevance |
| --- | --- |
| Model Generalisability and Data Imbalance | Over 90% of mental health datasets are English-based, limiting global applicability (Cao et al., 2024). Twitter dominance (63.8%) raises concerns about platform-specific overfitting. |
| Distinguishing Adaptive vs. Maladaptive Language | Not all mentions of distress indicate clinical illness (Owen et al., 2024). AI models need context-awareness to differentiate between peer support and genuine distress. |
| Explainability of AI (Black-Box Models) | Deep learning models lack transparency, making AI-driven diagnoses hard to interpret (Cao et al., 2024). Explainable AI (XAI) can improve model trust and clinical adoption. |
| Need for Diverse Datasets | Enhances model generalisability by including non-English languages |
| Refinement of Deep Learning Classifiers | Improves distinction between adaptive vs. maladaptive distress |
| Improved Sentiment Detection | Enhances context-sensitive NLP models |
| Handling Class Imbalance | Reduces overfitting and enhances predictive accuracy |

Future advancements in AI-based mental health detection require context-aware models that capture longitudinal distress patterns, not just isolated indicators. Addressing current limitations, such as overreliance on Western-centric datasets and lack of multimodal integration, demands refined models that blend linguistic, behavioural, and contextual signals. Accordingly, three core research questions are explored: (1) What linguistic and behavioural markers in social media content indicate mental health conditions? (2) How can AI models be optimised to detect early signs of distress? (3) What ethical challenges arise in applying these tools? The methodology that follows operationalises these questions through the comparative evaluation of four AI classifiers.

## ****2.4 Theme 3: Ethical Considerations****

The increasing reliance on AI for mental health detection raises critical ethical concerns that must be addressed to ensure responsible and fair implementation. While AI models offer significant advancements in early diagnosis and intervention, they also introduce risks related to privacy, bias, transparency, and informed consent (Saeidnia et al., 2024)​. These ethical dilemmas must be navigated carefully to protect individuals while ensuring the effectiveness of AI-driven mental health applications.

### 2.4.1 Privacy and Data Protection

One of the most pressing concerns is data privacy. AI-based mental health detection relies heavily on social media data, which is often collected without explicit user consent (Holtorf et al., 2023)​. Users sharing personal experiences on platforms such as Twitter and Reddit may be unaware that their posts are being used for mental health classification (Kaye et al. 2015)​. This lack of informed consent challenges established privacy regulations such as GDPR and HIPAA, which require clear disclosure when processing personal health data (Morley et al., 2020)​.

Protecting user privacy requires the implementation of data anonymisation methods and differential privacy techniques. However, these measures are not foolproof, as re-identification remains a possibility, particularly when linguistic patterns contain implicit personal information (Holtorf et al., 2023). This risk is especially pertinent in mental health contexts, where AI analysis of language may inadvertently expose sensitive aspects of an individual's identity.

### 2.4.2 Algorithmic Bias and Fairness

Bias in AI models present another ethical challenge. Many mental health AI models are trained on Western-centric, English-language datasets, which limits their effectiveness across diverse populations (Cao et al., 2024)​. Studies show that over 90% of AI-driven mental health detection models are based on English-language social media data, predominantly from the United States and Europe, excluding linguistic and cultural variations (Owen et al., 2024)​.

This bias not only reduces the accuracy of predictions for non-Western users but also raises concerns about discriminatory outcomes. If AI systems fail to account for cultural differences in mental health expression, they risk misdiagnosing or overlooking distress signals in underrepresented communities (Cao et al., 2024)​. Bias mitigation strategies such as cross-cultural dataset expansion, fairness audits, and diverse AI model training are essential to improve equitable mental health detection.

### 2.4.3 Explainability and Transparency

A key challenge in AI-driven mental health detection is the “black box” problem, where deep learning models lack transparency, limiting clinical trust and validation (Lipton, 2018; Joyce et al., 2023). While Transformer-based models like BERT and RoBERTa achieve high accuracy, they often fail to provide interpretable justifications for their outputs, hindering adoption in mental health care.

Explainable AI (XAI) techniques, such as attention visualisation and feature attribution, attempt to improve transparency but often serve as post hoc rationalisations rather than true explanations (Doshi-Velez & Kim, 2017; Teferra et al., 2024). This raises concerns about whether these methods genuinely enhance trust or simply obscure opaque decision-making.

Beyond transparency, algorithmic bias remains a critical ethical issue. A Western-trained AI tool misclassifying culturally specific distress indicators exemplifies how biased datasets can reinforce healthcare disparities (Cao et al., 2024). Tackling bias in AI-driven mental health tools has led to the development of approaches such as fairness audits, adversarial debiasing, and algorithmic impact assessments (Teferra et al., 2024). Yet, the practical application of these methods remains limited, largely due to the lack of consistent regulatory standards. In the absence of clear guidelines, questions surrounding accountability and ethical governance continue to hinder progress. Advancing responsible clinical deployment will require not only robust policy frameworks but also closer integration between technical experts, ethicists, and mental health practitioners.

### 2.4.4 Ethical Implementation of AI

Beyond technical considerations, the ethical implementation of AI in mental health care requires responsible governance and stakeholder involvement. AI should function as a supplementary tool rather than a replacement for human mental health professionals, ensuring that clinicians remain at the centre of decision-making (Floridi et al., 2018)​.

In addition, regulatory frameworks must balance innovation with accountability (See Table below). Ethical AI deployment requires ongoing ethics reviews, patient engagement, and continuous evaluation to minimise harm and maximise societal benefit (Saeidnia et al., 2024).

Table 9: Ethical Concerns and Solutions

| Ethical Concern | Description | Key Literature | Proposed Solutions |
| --- | --- | --- | --- |
| Privacy & Consent | AI models use social media data, often without explicit user consent, raising privacy and legal concerns. | Holtorf et al. (2023)​, Morley et al. (2020)​, Owen et al. (2024)​ | Implement informed consent protocols, strengthen data anonymisation techniques, and adhere to GDPR/HIPAA regulations. |
| Bias & Fairness | AI models predominantly use English-language data, leading to poor generalisation across cultures and languages. | Cao et al. (2024)​, Owen et al. (2024)​ | Expand datasets to include diverse linguistic and cultural groups, apply bias detection tools, and conduct fairness audits. |
| Explainability & Transparency | Transformer-based models (e.g., BERT, RoBERTa) function as “black boxes,” limiting clinicians’ trust in AI decisions | Lipton (2018)​, Joyce et al. (2023)​ | Develop Explainable AI (XAI) frameworks, such as attention visualisation and rule-based decision-making models. |
| Regulatory & Ethical AI Governance | AI tools risk being misused without clinical oversight, raising concerns about accountability and harm. | Floridi et al. (2018)​, Saeidnia et al. (2024)​ | Establish ethics review boards, enforce regulatory compliance, and ensure AI serves as a supportive tool, not a replacement for human clinicians |

## ****2.5 Conclusion****

The intersection of mental health, social media, and artificial intelligence reveals both significant opportunities and critical challenges in AI-driven mental health detection. Psychological theories such as Social Comparison Theory, Problematic Social Media Use, and the Cognitive-Behavioural Model offer valuable frameworks for understanding how digital behaviours affect psychological well-being. In parallel, AI methodologies like natural language processing and machine learning provide scalable solutions for early intervention and mental health assessment.

Despite recent advancements, key limitations persist. These include imbalanced datasets, limited cultural applicability, opacity in deep learning models, and unresolved ethical issues surrounding privacy, fairness, and informed consent. As discussed in Theme 3, many current models are trained on Western-centric data, lack contextual sensitivity, and function as black boxes, all of which erode trust and limit real-world impact. Addressing these issues is crucial for building AI systems that are both effective and ethically responsible.

With this in mind, the research focuses on transforming psychological constructs such as depression, anxiety, and suicidal ideation into linguistically quantifiable features suitable for AI classification. Both traditional machine learning methods and a domain-adapted transformer model (MentalBERT) are applied to evaluate comparative performance. Post-processing strategies are introduced to enhance fairness, interpretability, and generalisability. In doing so, the project addresses key gaps in the literature, particularly those involving the detection of nuanced language, platform-specific variation, and equitable model evaluation.

The following methodology chapter details the research philosophy, data sources, model selection, evaluation criteria, and ethical procedures used in the study. Each aspect of the research design is grounded in the interdisciplinary issues raised in this review, ensuring a methodologically rigorous and practically relevant approach.

# Methodology

## Research Philosophy and Approach

The study is grounded in a positivist philosophical stance, which emphasises empirical observation and objective measurement as the basis for generating knowledge (Gill & Johnson, 2010). This perspective aligns with the nature of computational research, where algorithmic processing, statistical evaluation, and reproducibility are central. In the context of building scalable AI systems to detect mental health indicators, positivism justifies the reliance on structured data, measurable features, and outcome-driven analysis. A deductive approach was used to operationalise this framework. Psychological conditions such as depression, anxiety, and suicidal ideation were mapped onto specific linguistic and behavioural markers found in social media content. This approach ensures coherence between the theoretical foundation, research questions, and chosen analytical methods.

Nevertheless, as Clark et al. (2022) observed, an exclusive reliance on a positivist lens in mental health research can risk overlooking emotional nuance. Expressions of distress often diverge from standard syntax, appearing instead through metaphor, sarcasm, or culturally specific euphemisms. These complexities underscore the value of alternative philosophical perspectives such as critical realism and interpretivism, which emphasise the importance of understanding social phenomena through individuals' subjective experiences. Although this study remains grounded in a positivist framework, future research could benefit from mixed methods approaches that combine quantitative analysis with thematic content analysis, thereby enhancing contextual sensitivity and ethical interpretation.

## 3.2 Research Design and Strategy

A quantitative experimental design was used to benchmark the performance of multiple classification models under controlled conditions. Four models—Logistic Regression, Support Vector Machine, Random Forest, and MentalBERT—were tested using a consistent pipeline to ensure fair comparison. Their performance was evaluated based on classification accuracy, macro averaged F1 score, and per class recall and precision.

The following chapter is structured according to the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework. Providing a structured and widely accepted framework for analytical projects. Table 10 outlines how each phase of the CRISP DM process was applied.

Table 10: CRISP-DM Framework Application

| Phase | Application |
| --- | --- |
| Business Understanding | Mental health risk detection for public well-being. |
| Data Understanding | Source exploration: Reddit and Twitter mental health datasets |
| Data Preparation | Cleaning, tokenisation, and balancing of class distribution. |
| Modelling | Training of LR, SVM, RF, and MentalBERT. |
| Evaluation | Accuracy, macro-F1, cross-validation, and ablation testing. |
| Deployment | Ethical and platform-specific implications of model use. |

This structure enabled the development of replicable, ethical, and transparent research practices.

## 3.3 Data Collection and Preprocessing

### 3.3.1 Sources and Acquisition

Primary data were sourced from Reddit, a platform characterised by long-form and thematically organised discussions. Posts were collected from four mental health-related subreddits: r/Depression, r/Anxiety, r/SuicideWatch, and r/CasualConversations, with ~11,000 posts each. An additional 10,000 posts from r/mentalhealth were reserved for out-of-sample validation.

These datasets were retrieved from AcademicTorrents.com , where they were hosted and curated by u/Watchful1. Data posted before April 2023 were originally collected via the Pushshift API, a widely used and trusted tool for large-scale Reddit data access, while newer content was gathered by u/raiderbdev and re-packaged by u/Watchful1 into structured JSON files. These files were then converted into CSV format and filtered using custom scripts to remove deleted, duplicate, or empty entries and isolate high-quality text data for model input.

This dataset is particularly valuable given the increasing difficulty of Reddit data collection. In mid-2023, Reddit significantly restricted third-party API access and banned tools like Pushshift from scraping live content, limiting researchers' ability to gather large-scale, historical social media data (Pushshift.io, 2023). As a result, curated repositories such as this one serve as an essential resource for academic work, offering not only scale (over 50,000 labelled posts) but also transparency in provenance and structure. The dataset’s comprehensiveness and timely preservation make it a robust and reliable foundation for training and validating AI models focused on mental health discourse.

To assess model generalisability, three Twitter datasets were retrieved from Kaggle: one general mental health dataset, a suicidal ideation dataset, and a behavioural language dataset. These sets represent more informal, ambiguous, and emotionally diverse language styles (see below for data collection justification).

Table 11: Overview of Datasets Used

| Platform | Dataset | Source | Purpose and Justification |
| --- | --- | --- | --- |
| Reddit | r/Casual Conversations, r/Depression, etc. | [Academic Torrents](https://academictorrents.com/details/1614740ac8c94505e4ecb9d88be8bed7b6afddd4) | Model training and validation: Chosen for thematic relevance, post length, and structured mental health discourse. Suitable for training models to detect linguistic signals of distress. |
| Reddit | r/mentalhealth | [Academic Torrents](https://academictorrents.com/details/1614740ac8c94505e4ecb9d88be8bed7b6afddd4) | Out-of-sample testing generalisability: Serves as an independent subreddit to assess how well models perform on unseen but related data. |
| Twitter | Depression Tweets | [Kaggle Dataset](https://www.kaggle.com/datasets/infamouscoder/mental-health-social-media) | External testing: Captures short-form, informal expressions of depressive symptoms; used to evaluate cross-platform robustness. |
| Twitter | Suicidal Tweet Detection Dataset | [Kaggle Dataset](https://www.kaggle.com/datasets/aunanya875/suicidal-tweet-detection-dataset) | External testing: Offers high-risk language samples; used to assess model performance on detecting acute mental health signals. |
| Twitter | Behavioural Tweets Dataset | [Kaggle Dataset](https://www.kaggle.com/datasets/arshkandroo/behavioural-tweets) | External testing: Adds diversity in language style and emotional tone, helping to evaluate adaptability across behavioural expression types. |

### 3.3.2 Cleaning and Transformation

Preprocessing included:

* Removal of deleted/duplicate posts
* Lowercasing of text
* Elimination of emojis, special characters, and URLs
* Discarding posts under 20 characters
* Tokenisation via HuggingFace libraries

A multi-label annotation structure was applied to reflect comorbid symptoms. Labels included depression, anxiety, suicidal ideation, and casual content. Due to class imbalance, Synthetic Minority Oversampling (SMOTE) and under sampling techniques were used to achieve class parity (~11,000 samples each).

Label reliability was a known limitation. Categories like "suicidal ideation" were inherited from public corpora without clinical annotation, risking false positives or omissions. As Shatte et al. (2019) suggest, future work should aim to improve label reliability by incorporating clinician validation or using participatory annotation frameworks involving individuals with lived experience.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 1: Code Snippet of Data Cleaning Process

## 3.4 Model Selection and Justification

Four classification models were selected for their complementary strengths:

* **Logistic Regression**: Transparent and interpretable; serves as a baseline
* **Support Vector Machine**: Effective for high-dimensional text
* **Random Forest**: Ensemble learner robust to overfitting
* **MentalBERT**: Transformer pre-trained on Reddit’s mental health corpus

The full source code for all models, preprocessing scripts, and evaluation tools, including MentalBERT and classical machine learning baselines is publicly available via the project’s GitHub repository:

* <https://github.com/zoyaam2003/mentalhealth_ai>

Table 12: Overview of Model Characteristics and Justifications for Use

| Model | Key Strength | Feature Type User | Justification |
| --- | --- | --- | --- |
| Logistic Regression | Simplicity and interpretability | TF-IDF | Baseline model |
| SVM | High-dimensional text handling | TF-IDF | Effective for sparse data |
| Random Forest | Non-linear pattern recognition | TF-IDF | Reduces overfitting |
| MentalBERT | Context-aware, pre-trained on Reddit | Tokenised text | Deep understanding of mental health language |

MentalBERT was chosen based on findings by Adel et al. (2024), who demonstrated a 4.2% F1 improvement over RoBERTa when classifying Reddit mental health content. Its architecture leverages self-attention mechanisms that interpret tone, euphemism, and context. Such features are crucial for mental health assessment.

## 3.5 Training, Testing and Validation

### 3.5.1 Classical Models

TF-IDF was capped at 5,000 features. Models were trained using an 80/20 split and wrapped with MultiOutputClassifier for multi-label output. **5-fold cross-validation** was conducted to assess robustness.

### 3.5.2 MentalBERT

Training utilised the HuggingFace and PyTorch stacks:

* Learning rate: 2e-5
* Batch size: 16
* Epochs: 3–5
* Loss: BCEWithLogitsLoss
* Optimiser: Adam
* Early stopping based on validation loss

Evaluation metrics included macro averaged F1 scores, precision, recall, and platform-specific comparisons. An ablation test, which removed the post-processing rules, resulted in a 7% drop in F1 score, highlighting the importance of these rules in maintaining model performance.

Macro averaged F1 was chosen due to its effectiveness in assessing multi-label classification tasks involving imbalanced data. This metric provides a balanced view by giving equal weight to each class, ensuring fair evaluation of underrepresented categories such as suicide and depression (Sammut & Webb, 2010). It also helps avoid inflated accuracy that can occur when dominant classes like casual content disproportionately influence results.

## 3.6 Post-Processing and Calibration Layer

While earlier evaluations focused on benchmarking models in controlled settings, the calibration layer was introduced to reflect the complexities of real-world deployment. Social media posts are often noisy, ambiguous, and sarcastic, which can pose significant challenges for binary classifiers. To tackle these issues, a heuristic-based post-processing layer was developed:

Table 13: Post-Processing Calibration Tools

| Component | Adjustment | Purpose |
| --- | --- | --- |
| Threshold tuning | Lowered for subtle classes | Improve recall |
| Keyword boosting | Increment class probability | Reinforce intent recognition |
| Sentiment fallback | Triggered casual label if sentiment > 0.2 | Avoid false positives |
| Co-occurrence logic | Add depression if suicide detected | Reflect comorbidity |
| Soft score detection | Score range 0.08–0.2 flagged for review | Enhance transparency |

These rules align with Beck’s Cognitive Distortion Theory (1976), which identifies euphemism and indirect expression as symptoms of psychological distress. Calibration enhanced both accuracy and ethical confidence.

A screenshot of a computer program

AI-generated content may be incorrect.

Figure 2: Post-Processing Code Snippet

## 3.7 Cross Platform Generalisability

To assess generalisability, MentalBERT was tested not only on its training platform (Reddit) but also on an out-of-sample Reddit dataset (r/mentalhealth) and three external Twitter datasets. This multi-platform evaluation was designed to understand how well the model adapts to linguistic and structural variation across digital environments.

The held-out Reddit dataset allowed for intra-platform validation. Though sourced from the same platform, r/mentalhealth presented different thematic and stylistic patterns compared to the original subreddits, offering a more nuanced test of model robustness. Results demonstrated consistent performance, suggesting that the model effectively generalised across subreddit communities.

To evaluate inter-platform generalisability, thresholds and classification logic were recalibrated for Twitter. Posts on Twitter are typically shorter, less structured, and exhibit greater lexical ambiguity. To compensate, the post-processing layer applied adjusted sentiment thresholds, refined keyword weights, and co-occurrence logic to interpret abbreviated or emotionally nuanced expressions more accurately. For instance, VADER sentiment scores were leveraged to avoid over-classifying mild or sarcastic posts as distress-related. A fallback mechanism was implemented to assign the 'casual' label when sentiment exceeded 0.2, minimising false positives.

A TF-IDF divergence test further confirmed that Reddit posts leaned toward clinical terminology (e.g., “diagnosed,” “therapy,” “medication”), while Twitter posts often relied on vaguer expressions (e.g., “I can’t do this anymore,” “I’m tired”). Recognising these platform-specific differences, keyword boosting and soft score detection were fine-tuned to maintain interpretability and recall across contexts.

Despite a modest decline in precision on Twitter, class-specific recall remained strong, especially for high-risk categories like suicidal ideation. This indicated that the model retained its sensitivity to critical signals even in noisier environments. Temporal metadata, such as posting frequency, was visualised using Power BI to explore behavioural trends, though excluded from model inputs to prevent overfitting.

Overall, these evaluations underscore the importance of adaptive post-processing in preserving MentalBERT’s performance across diverse platforms. Adjusting thresholds, leveraging sentiment cues, and incorporating heuristic rules were critical to maintaining both accuracy and ethical transparency in real-world application.

## 3.8 Ethical Considerations

Ethical clearance was granted by the University of Winchester. Although all data used were publicly available and anonymised, legal accessibility does not inherently guarantee ethical validity. As Holtorf et al. (2023) note, online disclosures often reflect personal intent and emotional vulnerability, which users may not anticipate being subject to algorithmic analysis.

To enhance ethical integrity in both model development and application, the MentalBERT classifier underwent an independent review by Adil Majeed (Data Scientist, Intellytics). His evaluation confirmed that the system aligns with established principles of responsible AI deployment.

### 3.8.1 Best Practices Applied:

* No user identification or engagement
* No real-time prediction or intervention
* Sensitive categories (e.g., suicide) interpreted in aggregate only
* Findings intended for ethical system design, not clinical judgment

Building on Owen et al. (2024) and Cao et al. (2024), the study highlights:

* Dynamic consent systems embedded in platforms
* Fairness audits for underrepresented dialects and cultures
* User co-design in future iterations
* Transparent explanations for flagged outputs

Ethical design is approached as an ongoing process rather than a one-time requirement. Such commitments are essential to ensuring the responsible development and deployment of digital mental health technologies.

# Findings and Analysis

****4.1 Platform Signals and Mental Health Context****

This chapter examines the emotional rhythms, linguistic markers, and support gaps present across Twitter and Reddit, treating these platforms not simply as data sources but as distinct social contexts that shape the expression of psychological distress. While Twitter remains the primary focus due to its scale, immediacy, and brevity-oriented communication, Reddit is used as a reflective comparator. The contrast between the two platforms highlights important discrepancies in how AI models interpret emotional content, particularly when applied across varying communicative environments.

The aim here extends beyond reporting model outputs; it is to critically evaluate what these outputs overlook. Visualisation tools such as Power BI (See Appendix C, for full dashboard) are employed not merely for descriptive purposes, but as interpretive instruments that convert large-scale behavioural data into sociolinguistic insight. Mental health disclosures are embedded within cultural, temporal, and emotional frameworks that often elude standard classification metrics. Addressing this complexity is vital for enhancing the sensitivity of AI models to nuanced or indirect forms of psychological expression.

Drawing upon psychological theories such as affect regulation and social comparison, alongside communication frameworks related to platform-specific norms and digital performativity, this chapter conceptualises mental health as a patterned and context-dependent phenomenon. It argues that accurate AI detection requires more than computational capability. It also requires a well-informed understanding of when, why, and how individuals choose to articulate or withhold psychological vulnerability in online spaces (Karim et al., 2020; Shannon et al., 2022; Cao et al., 2024).

****4.2 Temporal Patterns and Emotional Timing on Twitter****

Temporal analysis of mental health discourse on Twitter reveals distinct behavioural rhythms that align with psychological stress cycles and platform-specific constraints. Using a six-year corpus (2012 to 2017), along with visual data generated through Power BI dashboards and a tweet density heatmap segmented by hour and weekday, this section examines how emotional distress is structured over time and considers how such patterns complicate accurate AI detection. However, While the insights presented here offer a general overview, the interactive Power BI dashboard provided a more detailed and dynamic exploration, allowing temporal patterns to be filtered by specific years, hours, or labels to uncover more nuanced trends.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 3: Power BI Temporal Trends Dashboard

As shown above, a consistent pattern emerges in the clustering of tweets labelled as suicide and depression between 11 p.m. and 3 a.m., particularly during the years 2015 and 2016 (see Appendix, Figure 19 and 20). These surges occur during periods of relatively low overall posting activity, highlighting an inverse relationship between emotional severity and message volume. This temporal concentration aligns with circadian rhythm disruption theory, which suggests that late-night hours are associated with reduced emotional regulation.

Additionally, Dunbar’s concept of Social Recovery Zones (2014) proposes that nighttime represents a psychosocial blind spot, characterised by limited access to offline support and a heightened risk of emotional overflow into digital platforms. While this screenshot provides a general overview of these patterns, the interactive Power BI dashboard allows for more precise exploration, including year-specific filtering and temporal segmentation, which make these late-night trends more clearly observable.

The dashboard also highlights consistent spikes in suicide-labelled posts on Sunday evenings, beginning around 8 p.m., observed across multiple years. These patterns reflect anticipatory anxiety, a form of psychological strain associated with the impending return to structured responsibilities such as school or work. Peaks in depression-labelled tweets on Thursdays, particularly evident in 2015 and 2016, may signal the build-up of weekly exhaustion and represent a digital manifestation of burnout. These shifts linked to specific weekdays suggest that emotional expression follows identifiable rhythms and is closely tied to broader societal routines. Furthermore, the month of June shows a notable spike in suicide-labelled tweets, suggesting potential seasonal or academic stress factors (See Appendix C. Figure 18)

This interpretation is supported by previously mentioned empirical research. De Choudhury et al. (2014) identified similar peaks in depressive expression on social media during late-night hours and in the lead-up to weekdays, particularly among student populations. These findings strengthen the argument that psychological distress follows distinct temporal patterns that do not necessarily align with overall posting volume.

A grid of squares with a purple and white rectangle

AI-generated content may be incorrect.

Figure 4: Twitter Density Heatmap

The heatmap illustrates how psychological distress often emerges during quieter periods on the platform, which may be overlooked by models that rely primarily on volume or engagement-based signals. Many current classifiers either disregard temporal metadata or interpret it as irrelevant noise. However, the most critical emotional indicators, particularly those related to suicide, tend to appear during times of lowest visibility.

Although the earlier years in the dataset (2012 to 2013) display some inconsistencies in temporal coverage, the patterns observed from 2015 to 2017 remain consistent and meaningful. These findings emphasise the importance of recognising time not as a secondary variable, but as a key feature with direct relevance to mental health risk detection.

Table 14: Observations and Possible Implications

| Observation | Implication for AI Detection |
| --- | --- |
| Late-night distress surges. | Add risk-weighting for nocturnal posts. |
| Sunday anxiety spikes. | Flag future-facing stress periods. |
| Thursday depression increase. | Model cumulative stress fatigue. |
| Low-volume, high-severity posts. | Avoid engagement-based thresholds. |
| Sparse early-year data. | Prioritise recent, time-rich training corpora. |

These temporal dynamics partially explain why MentalBERT outperformed traditional models in recall for suicide-related posts (F1 = 0.72). Its attention-based mechanism likely captured latent temporal patterns that TF-IDF-based models were unable to detect.

Although the Power BI dashboard revealed some inconsistencies in data coverage, especially in the earlier years, the patterns observed in later periods are more consistent and well-sampled. These results not only align with established findings in the literature but also reinforce the model’s capacity to detect high-risk content during vulnerable timeframes.Further visualisation of distress fluctuation by month across labels is provided in Appendix G (see Figure), highlighting seasonal surges in suicide-related posts.

Despite the current lack of explicit time features, MentalBERT’s performance demonstrates the potential of context-aware architectures in identifying distress when overall visibility is low. Moving forward, incorporating temporally balanced data and time-sensitive features could further strengthen accuracy and real-world applicability.

****4.3**** Engagement Without Empathy: The Misreading of Support

In digital environments like Twitter, engagement metrics such as likes, retweets, and replies are often mistaken for signs of emotional support or validation. However, visual analysis from the Power BI dashboard and sentiment charts suggests a more concerning reality. Attention frequently operates independently of empathy, and high engagement can distort both human and algorithmic interpretations of distress. Instead of indicating genuine care, these metrics tend to reflect visibility rather than vulnerability, which compromises the reliability of AI models in mental health classification.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 5: Power BI Engagement and Support Dashboard

As shown above, the dashboard highlights a clear paradox. A striking 87.96% of anxiety-labelled posts receive no engagement, with similarly high neglect for suicide and depression content. In contrast, casual posts show a more balanced distribution between ignored and engaged responses. This pattern suggests an inverse relationship between emotional urgency and user interaction, likely influenced by discomfort, stigma, or social norms that discourage open engagement with distress. While Social Validation Theory assumes that expression is reinforced through approval, these findings suggest that posts conveying psychological distress are frequently met with silence rather than support.

A graph with green squares and white text

AI-generated content may be incorrect.

Figure 6: Emotional Probabilities with and without Engagement

This disengagement carries significant algorithmic implications. A bar chart comparing changes in emotion probabilities with and without engagement shows that engagement increases the likelihood of posts being classified as “depression” by 0.39, while decreasing the probabilities for “suicide,” “anxiety,” and “casual.” This suggests that models tend to associate validation with familiar and interpretable forms of distress, particularly depression, while downplaying more ambiguous or stigmatised expressions such as suicidality or panic.

A graph of a line graph

AI-generated content may be incorrect.

Figure 7: Sentiment Distribution

The sentiment density plot reinforces this pattern. Posts with high engagement cluster around neutral sentiment, whereas those with low engagement display greater emotional variability, especially on the negative end. This supports Platform Affordance Theory, which argues that algorithms favour content that is easy to circulate rather than emotionally urgent or complex. As a result, emotionally neutral expressions are amplified while raw or incoherent forms of distress are marginalised.

This dynamic echoes Andalibi et al. (2018), who observed that emotionally restrained posts on Instagram were more likely to gain visibility, while deeply vulnerable content was often ignored. Across platforms, the appearance of empathy can obscure a deeper structural neglect. This phenomenon can be described as the algorithmic sympathy illusion.

From an AI ethics standpoint, this presents a major concern. Detection models may confuse social popularity with clinical relevance, potentially overlooking serious psychological risk. Engagement should therefore be treated not as a proxy for severity, but as a factor that can distort model perception.

This also connects to Goffman’s theory of impression management, where users curate emotionally acceptable personas that mislead both human observers and AI systems trained on overt affective signals.

Table 15: Observations and Possible Implications

| Observation | Implication for AI Detection |
| --- | --- |
| High-risk posts often ignored. | Weight linguistic and temporal features over engagement. |
| Neutral sentiment correlates with virality. | Penalise sentiment neutrality in risk-sensitive contexts. |
| Depression favoured in classification. | Rebalance classifiers toward underrepresented emotional labels. |
| Engagement is not a proxy for support and may be skewed by platform norms or excess casual content. | Decouple likes and shares from severity inference to avoid misclassification of urgent posts. |

4.4 Lexical Signatures of Distress and Model Interpretability

Accurately identifying psychological distress on Twitter requires more than detecting overt negativity. It demands sensitivity to how emotion is communicated through brevity, euphemism, and indirect language. To support this, the external and main Twitter datasets were merged to form a more linguistically diverse corpus, enabling closer analysis of lexical patterns linked to suicide, depression, anxiety, and casual expression. This combined dataset offers a richer view of how distress is articulated, particularly in tweets that carry high emotional risk.

The TF-IDF trigram heatmap reveals key lexical differences and overlaps between suicide- and depression-labelled tweets. Suicidal posts often contain emotionally opaque or grammatically fragmented phrases such as “drop hat pleasure,” “curse kid beat,” and “feel like im.” These expressions convey emotional volatility and helplessness through metaphor, syntactic irregularity, and abbreviation. Depression-labelled tweets display similar obliqueness but with slightly more structure, using phrases like “crazy need help” and “come damage people.” Some terms, such as “feel like im,” appear in both categories, reflecting the lexical entanglement of emotional states and supporting Cognitive Distortion Theory, particularly patterns of catastrophising and mind-reading.

A screenshot of a graph

AI-generated content may be incorrect.

Figure 8: Top TF-IDF Trigrams in Suicidal vs Depressive Tweets

This oblique phrasing aligns with findings from Coppersmith et al. (2015), who observed that suicidal ideation often manifests through indirect language. Such expressions are clinically significant but difficult for algorithms to detect. These overlaps expose a critical limitation in NLP systems, which tend to prioritise overt sentiment or keyword presence over structural or contextual cues.

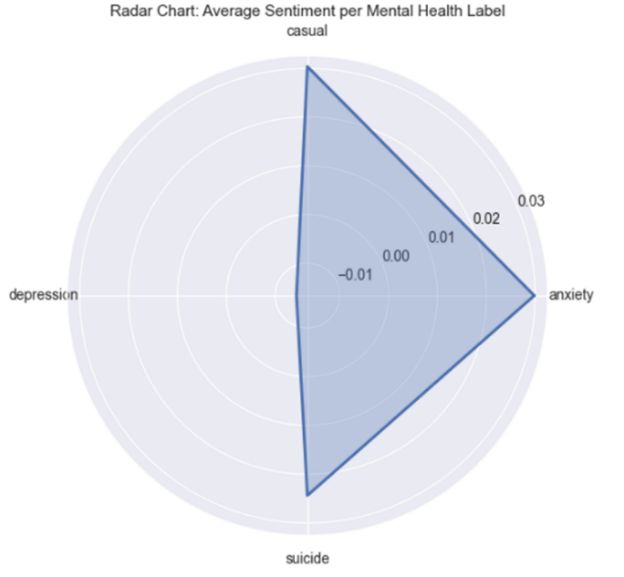


Figure 9: Radar Chart for Sentiment

This limitation becomes more apparent in the radar chart displaying average sentiment across mental health labels. Tweets labelled as “suicide” and “depression” cluster around neutral sentiment values, while “casual” and “anxiety” tweets skew slightly more positive. This flat affective profile suggests that neutral sentiment scores should not be interpreted as emotional neutrality. In suicide-risk contexts, neutrality may conceal urgency. Crisis linguistics describes this phenomenon as linguistic distancing, where individuals at risk adopt emotionally muted language in high-stakes disclosures.

Further evidence of semantic masking appears in the emotive word frequency plot for suicide-labelled posts. Common words include “help,” “love,” “sad,” and “trauma,” but are mixed with terms like “curse,” “belittle,” and even “happy.” These lexical inversions, where emotionally incongruent or overly neutral words obscure distress, introduce significant interpretive ambiguity. This pattern reflects semantic inversion, where ambiguous language shields the speaker from vulnerability but complicates algorithmic interpretation.

A graph showing a number of words

AI-generated content may be incorrect.

Figure 10: Word Count vs Sentiment

The scatterplot comparing tweet word count to sentiment polarity challenges standard assumptions. Suicide-labelled tweets tend to be shorter but exhibit a wide range of sentiment scores. In these cases, brevity does not imply low information; instead, it signals emotional urgency. Posts such as “I can’t do this anymore” may appear neutral in sentiment analysis but are clinically critical. Conventional models, however, often deprioritise such short texts, mistakenly treating them as uninformative.

Together, these findings highlight a deep mismatch between how distress is expressed online and how AI systems are trained to detect it. Distress is not always overt. It may be brief, metaphorical, or emotionally muted. NLP models that rely heavily on directness or sentiment polarity risk missing critical signals.

This underscores the need for systems that can interpret not just the vocabulary of distress, but its structure. Emotionally urgent language often relies on brevity, metaphor, and irregular syntax. MentalBERT’s lower F1 score for depression (0.63) compared to suicide (0.72) illustrates this challenge. While its contextual learning architecture excels at recognising clearer markers of distress, it struggles with ambiguous phrasing. Fragments like “feel like im” exemplify this ambiguity, signalling both risk and noise. Disambiguating such language requires temporal or sequential context.

To overcome these challenges, future models should incorporate features that detect euphemism, emotional masking, and fragmentary expression. This includes weighting short-form content, using contextual embeddings, and integrating crisis-specific lexicons aligned with real-world signalling. In such contexts, subtle or indirect cues often carry the most urgent psychological meaning.

4.5 Emotional Overlap and Escalation Paths

Mental health discourse on Twitter does not occur in isolation. Emotional states such as anxiety, depression, and suicidal ideation often overlap, shift, or intensify over time, which challenges the static and single-label assumptions built into many AI detection systems. This section draws on a combined dataset from internal and external Twitter sources and presents three visualisations: a label co-occurrence heatmap, a transition probability matrix, and a Sankey-style frequency plot. Together, these tools reveal the relational and temporal complexity of how psychological distress is communicated online.

A diagram of a heatmap

AI-generated content may be incorrect.

Figure 11: Label Co-Occurrence Heatmap

The co-occurrence heatmap shows a notable semantic entanglement between depression and both casual and suicide-labelled tweets. More than 1,500 tweets carry both depression and suicide labels, while over 4,800 are shared between depression and casual. These intersections support Comorbidity Theory, which recognises depression as frequently coexisting with suicidal ideation or more emotionally masked states. Anxiety, on the other hand, appears more isolated, with only three cases co-labelled with suicide. This may result from linguistic divergence, as anxiety-related posts often use anticipatory or future-oriented phrasing such as “I can’t sleep” or “I’m scared about tomorrow,” making them less legible to models trained on more explicit signals of distress.

A screenshot of a graph

AI-generated content may be incorrect.

Figure 12: Transition Probability

The transition probability matrix illustrates a more longitudinal emotional risk pattern. Among tweets initially identified as “depression,” 27 percent later shift toward suicide classification, while 35 percent transition into casual categories. These changes are consistent with Path Dependence Theory, which suggests that earlier emotional states influence the likelihood of subsequent ones. Suicidal posts are most commonly preceded by depression (57 percent), pointing to an identifiable escalation pathway that reflects real-world psychological trajectories.

A graph of different colored bars

AI-generated content may be incorrect.

Figure 13: Sankey-Style Transition Frequencies

The Sankey-style frequency plot reinforces this interpretation, with depression emerging as the most frequent origin label and a strong predictor of later suicide classification. Rather than following a linear progression, the data reflects recursive emotional cycles where users move between casual expression and latent distress before escalating into crisis. This mirrors longitudinal patterns reported by Chancellor et al. (2021), who observed similar sequences on Reddit in which anxiety and depression often preceded suicidal disclosures.

Such patterns expose a significant blind spot in current classification systems. Most models rely on single-post inference, failing to account for emotional transitions over time. This temporal blindness can result in delayed detection and missed intervention windows. Models need to adopt sequence-aware architectures, such as recurrent or transformer-based systems, trained on temporally ordered, multi-label data. Capturing both state and trajectory would allow for a shift from static output to dynamic, risk-sensitive insight.

A key limitation of the present analysis is the lack of user-linked tweet sequences, which restricts true longitudinal modelling. While transition matrices offer useful probabilistic approximations, they cannot map individual emotional arcs. Future work should integrate user-level identifiers to track emotional trajectories more accurately and facilitate earlier, context-aware intervention strategies rooted in real-world behavioural patterns.

4.6 Reddit vs Twitter and the Training Illusion

A graph of a number of different colored bars

AI-generated content may be incorrect.

Figure 14: Word Count Distribution by Label

As shown, Differences in content depth appear clearly in the boxplot of word count by label. Reddit posts show greater variance and frequent outliers exceeding 1,000 words, particularly for suicidal and depressive content. In contrast, Twitter posts remain tightly constrained, rarely surpassing 100 words. This discrepancy is not only structural but diagnostic. Brevity on Twitter compresses affective expression, increasing the risk that models trained on long-form Reddit content will overlook indicators of distress.

A graph of a number of comments

AI-generated content may be incorrect.

Figure 15: Word Count Distribution by Label

Comment count distribution further highlights Reddit’s role in fostering support-seeking. Suicide and depression-labelled posts often receive replies, indicating active community engagement. On Twitter, however, high-risk posts typically receive little to no interaction. This imbalance is quantified by the Lorenz Curve of engagement inequality, which yields a Gini coefficient of 0.686. Engagement is highly concentrated, meaning a small subset of users dominates visibility. Models trained in Reddit’s more evenly distributed environment are unlikely to perform reliably under these conditions.

A graph of a curve

AI-generated content may be incorrect.

Figure 16: Lorenz Curve

The TF-IDF divergence plot reinforces the linguistic gap between platforms. Terms such as “mental,” “health,” and “depression” appear far more frequently on Reddit, whereas “feel,” “like,” and “love” are more common on Twitter. Reddit discourse tends to be clinical and introspective. Twitter, by contrast, favours idiomatic, affective, and euphemistic language. Topic models by platform confirm this split. Reddit clusters around therapy, symptoms, and diagnosis, while Twitter clusters involve ambiguous emotional cues with limited clinical framing. A model trained to detect explicit risk on Reddit may fail to register implicit signals on Twitter, categorising vulnerability as casual or irrelevant.

A graph of different colored squares

AI-generated content may be incorrect.

Figure 17: Top Diverging Terms

Platform-level probability scores offer further evidence of this misalignment. The average depression probability assigned by the model is 0.32 for Reddit posts but only 0.21 for those from Twitter. This 0.11 difference suggests the model, likely fine-tuned on Reddit, struggles to identify distress in the compressed and stylised format of Twitter. What is framed as concern on Reddit may appear on Twitter through sarcasm, metaphor, or emotionally coded phrasing, which models are poorly equipped to interpret without adaptation.

These outcomes echo concerns raised in critiques of data colonialism (Couldry & Mejias, 2019), where emotional data from one community is extracted to build tools that are applied to another without contextual alignment. This practice not only compromises accuracy but also interpretability. Feature weights derived from Reddit-based vocabularies may behave unpredictably when applied to Twitter data, especially in black-box NLP models, undermining both reliability and ethical accountability.

To address this, detection systems must apply field-aware calibration. Language is not neutral, and models must be trained and validated within the platform context in which they are deployed. If cross-platform generalisation is pursued, it should involve domain adaptation, interpretability mechanisms, and sensitivity to platform-specific norms and affective expression. This includes using platform-specific embeddings, adjusting features to align with dominant discourse styles, and employing ensemble models to accommodate linguistic variation.

Training on Reddit and deploying on Twitter is comparable to using medical case notes to interpret street graffiti. Without contextual understanding, even high-performing models risk serious misclassification. While the present analysis is based on extensive data from both platforms, its findings reflect the cultural, linguistic, and temporal makeup of largely English-speaking, Western, and youth-dominated user groups. The models studied here may not generalise to non-Western cultures, other languages, or platforms where affect is conveyed more visually, such as Instagram or TikTok. Future research should incorporate multilingual corpora and culturally embedded expressions of mental health.

To summarise how platform-specific differences impact classifier design, Table 16 outlines the core divergences between Reddit and Twitter and their implications for model performance.

Table 16: Platform Feature Analysis

| Platform Feature | Reddit | Twitter | Model Implication |
| --- | --- | --- | --- |
| Post Length | Long, elaborative | Short, compressed | Risk of under-detection in short, intense tweets |
| Language Style | Clinical, diagnostic | Euphemistic, affective | Lexical mismatch with Reddit-trained models |
| Engagement Norms | Supportive comments | Low interaction, high visibility | Poor recall for ignored high-risk posts |
| Vocabulary | “Depression,” “anxiety,” “therapy” | “Feel,” “can’t,” “done” | Sentiment masking via common language |

4.7 Summary and Model Design Implications

The findings of this chapter demonstrate that online expressions of mental health distress are temporally, linguistically, and contextually complex. Distress is rarely static or explicit; it is shaped by the timing of posts, platform affordances, linguistic subtlety, and the dynamics of digital engagement. Current AI models, which are often trained on sentiment polarity or keyword frequency, fail to account for this multi-dimensional reality. As a result, they risk misclassifying urgent disclosures or overfitting to emotionally neutral but highly visible content.

To bridge this gap, models must evolve into systems that interpret emotional trajectories rather than relying on static snapshots. This requires the integration of temporal signals, sensitivity to brevity and euphemism, recognition of multi-label emotional overlaps, and fine-tuning to the communicative norms of each platform. The depth of a Reddit post cannot be equated with the brevity of a Twitter post without appropriate contextual calibration.

Table 17: Final Observation Summary

| Observation | Implication for AI Detection | Ethical Consideration |
| --- | --- | --- |
| Late-night surges | Include time-aware flags | Avoid mislabelling due to nocturnal cultural norms |
| High engagement ≠ support | Deprioritise likes/retweets as risk indicators | Prevent false reassurance or under-flagging high-risk posts |
| Brief but intense posts | Calibrate models to brevity-emotion trade-off | Ensure short cries for help are not dismissed as low-signal |
| Emotion co-occurrence | Use sequence-aware classification (e.g., RNNs) | Detect escalation early; avoid overfitting to terminal labels |
| Reddit vs Twitter mismatch | Platform-specific fine-tuning with domain calibration | Prevent cultural and discursive misalignment during deployment |

Ultimately, ethical AI in mental health must move beyond detection to contextual understanding, ensuring that its insights are technically accurate, socially informed, and clinically relevant.

# Evaluation and Discussion

5.1 Critical Evaluation of Models and Methodology

This section assesses the performance and limitations of AI models in detecting mental health signals on Twitter. It covers model comparison, error sources, platform effects, ethical considerations, and practical trade-offs. The analysis is structured around evaluation, domain adaptation, ethics, limitations, and future directions.

5.2 Comparative Model Evaluation

To assess the reliability and practical viability of automated mental health detection, four classifiers were trained on a Reddit-derived dataset using an 80/20 split: Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and a fine-tuned multi-label MentalBERT model. Table 18 outlines their comparative performance in terms of accuracy, macro F1-score, and class-specific F1-scores for depression and suicide—chosen due to their clinical relevance (see below).

Table 18: Model Performance Comparison

| Model | Accuracy | Macro F1 | F1 (Depression) | F1 (Suicide) |
| --- | --- | --- | --- | --- |
| Logistic Regression | 0.38 | 0.60 | 0.25 | 0.47 |
| Random Forest | 0.53 | 0.65 | 0.47 | 0.60 |
| SVM | 0.58 | 0.67 | 0.49 | 0.64 |
| MentalBERT | 0.76 | 0.76 | 0.63 | 0.72 |

Logistic Regression struggled to capture complex emotional and syntactic features, particularly depressive cues (F1: 0.25), and showed strong bias toward neutral labels. While SVM and Random Forest offered moderate improvements, both faltered with ambiguous or indirect language typical of mental health discourse.

MentalBERT outperformed traditional models, achieving high accuracy and a balanced precision-recall profile. Its macro F1 score of 0.76 indicates strong generalisation, and an F1 of 0.72 for suicide-related content highlights its ability to interpret euphemistic or oblique expressions. These results emphasise the value of deep contextual embeddings in affective computing.

Its effectiveness stems not only from its transformer-based design but also from fine-tuning on annotated Reddit data, which reflects diverse psychological discourse. Nonetheless, its transferability to platforms like Twitter requires further scrutiny, as outlined below.

5.3 Error Analysis: Why Models Fail

Despite improvements, all models showed errors—false positives, false negatives, cultural misreads, and gendered mismatches—each with ethical and practical implications (see Table).

Table 19: Classification Error Types in Mental Health Detection Models

| Error Type | Example Tweet | Cause | Implication |
| --- | --- | --- | --- |
| False Positive | "Kill me now, I forgot my charger" | 0.60 Lexical trigger misinterpreted | Risk of unnecessary surveillance or intervention |
| False Negative | "I’m tired of everything" | Euphemistic and emotionally numbed | Missed identification of genuine distress |
| Cultural Misread | "God is testing me again" | Spiritual idiom misclassified | Lower recall in culturally diverse user groups |
| Gendered Language | "I just want to disappear" (male user) | Gender-specific expression of distress | Under-detection in male populations |

As shown in Table 19, model errors typically fall into four categories: false positives, false negatives, cultural misreads, and gendered language bias. False positives often stem from literal interpretations of figurative or sarcastic phrases (e.g., “Kill me now, I forgot my charger”), highlighting limitations in handling pragmatics and contextual cues. False negatives arise from vague or euphemistic expressions of distress, which models may miss due to the absence of explicit risk markers.

These gaps align with insights from Crisis Text Linguistics and Beck’s Cognitive Distortion Theory, which emphasise that genuine distress is often conveyed through emotionally muted or distorted language. Cultural idioms and gendered expression patterns further contribute to misclassification, especially among culturally diverse users and male populations, who may express distress in less typical ways.

Altogether, these examples emphasise the need to integrate psychological theory into both feature engineering and model design. Techniques such as aggregating multiple posts, tracking emotional shifts over time, or including thread-level context can enhance detection of affective nuances that are missed by single-post models.

Lastly, the lack of interpretability remains a major barrier. Although MentalBERT demonstrates strong classification performance, its opaque decision-making process limits explainability, complicates clinical validation, and reduces user trust. Incorporating Explainable AI tools like SHAP or LIME can help clarify model outputs, support ethical accountability, and enable mental health professionals to better assess or contest predictions before acting on them.

5.4 Domain Shift and Platform Tensions

Training models on Reddit data and applying them to Twitter presents clear domain mismatch issues. Reddit posts are typically longer and more reflective, while Twitter favours brevity, immediacy, and irony. TF-IDF divergence analysis confirmed a linguistic gap: Reddit had higher frequencies of clinical terminology ("therapy," "diagnosis"), while Twitter was saturated with informal and affect-laden expressions ("tired," "can’t anymore").

From a socio-linguistic perspective, this aligns with Bourdieu’s Field Theory, which posits that discourse is shaped by cultural and platform-specific norms. Deploying Reddit-trained models on Twitter may risk misinterpretation, especially among non-Western, multilingual, or marginalised groups whose linguistic conventions differ.

MentalBERT’s strong Twitter performance, despite being Reddit-trained, is encouraging but also problematic. It risks what Couldry and Mejias term "data colonialism," whereby data from one community is repurposed to govern or predict behaviour in another. Without platform-specific calibration and culturally aware fine-tuning, this approach could reinforce digital inequities.

5.5 Study Limitations

Several factors limit the generalisability of this research. The Twitter dataset lacked consistency and sufficient user metadata, restricting longitudinal and behavioural analysis. Tweets also provide limited context, with no paralinguistic cues, making sarcasm and subtext difficult to interpret. Emotion labels were algorithmically generated without clinical input, risking misclassification of complex psychological states. Additionally, reliance on engagement-weighted features may bias models toward popular but emotionally neutral content, reducing sensitivity to high-risk signals.

5.6 Objective Evaluation Summary

The table below summarises the extent to which each research objective was addressed in this study.

Table 20: Research Objectives Status

| Research Objective | Status | Summary |
| --- | --- | --- |
| Identifying linguistic and behavioural markers in social media. | Successfully Achieved | Euphemism, sarcasm, and emotional detachment were identified as key linguistic indicators. |
| Develop predictive models to detect distress from posts | Successfully Achieved | MentalBERT outperformed classical models, showing strong classification accuracy and ability to interpret nuanced expressions. |
| Evaluate ethical and privacy implications | Successfully Achieved | The study examined algorithmic bias, lack of consent, and misclassification risks, proposing safeguards to support responsible deployment. |
| Propose practical applications for AI tools | Partially Achieved | Conceptual applications were discussed, but implementation in live systems was not tested or prototyped. |

# Conclusions and Future Work

6.1 Core Contribution Summary

This dissertation evaluated the feasibility and ethical challenges of applying AI to detect mental health signals on Twitter. While MentalBERT demonstrated strong performance (macro F1-score: 0.76), especially in recognising nuanced expressions of distress, key limitations remain. These include false positives from figurative language, platform-specific inconsistencies, and constrained datasets lacking demographic depth. Rather than emphasising accuracy alone, the study highlights the need for ethically grounded, context-aware models that account for linguistic subtlety, user intent, and cultural variation, offering a pathway toward more responsible AI deployment in digital mental health settings.

6.2 Practical and Technical Takeaways

From a technical standpoint, the study highlights the value of model architecture in capturing emotional nuance. MentalBERT outperformed SVM, logistic regression, and random forest models, particularly in identifying suicide- and depression-related tweets. Yet, accuracy metrics alone are insufficient. Misclassifications often arose from ambiguous or euphemistic language, underscoring the need for hybrid models that integrate semantic, temporal, and sentiment-aware features.

The analysis also revealed that engagement metrics like likes and retweets are poor proxies for psychological risk, suggesting models should prioritise behavioural cues over popularity signals.

For real-world application, ethical safeguards are essential. These include real-time sentiment monitoring, clinician oversight, and clear opt-out mechanisms to ensure transparency, user autonomy, and responsible use in mental health contexts.

6.3 Ethical Futures: Responsible AI for Distress Detection

Responsible AI in mental health involves more than emotion detection; it requires interpreting distress within a user’s digital, cultural, and temporal context. Sarcasm may be flagged as suicidal, while genuine distress expressed briefly or euphemistically can be missed.

Data provenance presents ethical risks. While Twitter and Reddit posts are publicly available, they were not shared with clinical use in mind. Reddit’s expressive content also raises concerns about data colonialism when personal disclosures are used to develop tools applied elsewhere. Ethical development must prioritise user consent, autonomy, and cultural sensitivity.

Bias limits effectiveness. Models trained on Western data often misinterpret culturally specific expressions, such as spiritual exhaustion or family-based metaphors. Language also varies with gender, age, and region, requiring more inclusive, adaptive training methods.

Consent and autonomy are vital. Users rarely consent to being flagged as “at-risk.” Ethical safeguards should include opt-out mechanisms, appeals processes, and human moderation to avoid harmful misclassification.

Deployment must prioritise explainability and clinical oversight. Tools like SHAP and LIME can help interpret model outputs, aligning with NHSX (2022) guidelines on transparency and safety. Outputs must be actionable and understood by practitioners.

These principles apply directly to platforms like Wysa, Woebot, and TalkLife, where AI assists emotional monitoring under human supervision. MentalBERT could act as a moderation layer, flagging high-risk posts for review. Integrated into clinician-in-the-loop systems, it supports early intervention while maintaining ethical and professional standards.

6.4 Future Research and System Redesign

Several future directions emerge from this work:

Table 21 Future Improvements

| Improvement | Summary Description |
| --- | --- |
| Model Refinement | MentalBERT should be fine-tuned specifically on Twitter data. Domain adaptation could involve masking usernames, integrating emoji embeddings, and recognising slang and sarcasm markers. |
| Multimodal Features | Incorporating emoji, image content, and typing speed could enhance model sensitivity to youth-oriented distress patterns. |
| Demographic Inference | Carefully and transparently estimate age or location where ethically permissible, to calibrate outputs against culturally distinct expressions of distress. |
| Trajectories and Threads | Use RNNs or transformer-based temporal models to trace users’ emotional journeys across tweet sequences. |
| Clinical Partnerships | Collaborate with mental health professionals to validate labels and refine intervention thresholds. |
| Expanded Data Diversity | Expand beyond English to include non-Western linguistic frames Future studies should include multilingual datasets and minority dialects. Cross-platform research, especially incorporating TikTok or Discord, could illuminate non-textual affective expressions. |

In conclusion, future AI systems must progress beyond basic keyword classification to become culturally informed, context-aware, and emotionally intelligent tools. Such models should be capable of discerning not only when to raise an alert but also when to pause, inquire further, or exercise restraint. With external validation from an expert in ethical AI deployment, this research offers a meaningful step toward that objective. The overarching goal is to identify digital expressions of psychological distress with the precision, sensitivity, and ethical integrity they warrant.

# References

* Andalibi, N., Ozturk, P., & Forte, A. (2018). Sensitive self-disclosures, responses, and social support on Instagram: The case of #depression. *Proceedings of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1485–1500. <https://doi.org/10.1145/2998181.2998243>
* Karim, F., Oyewande, A. A., Abdalla, L. F., Ehsanullah, R. C., & Khan, S. (2020). Social media use and its connection to mental health: A systematic review. *Cureus*, 12(6), e8627. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7364393/>
* Kelly, Y., Zilanawala, A., Booker, C., & Sacker, A. (2018). Social media use and adolescent mental health: Findings from the UK Millennium Cohort Study. *EClinicalMedicine*, 6, 59–68. <https://doi.org/10.1016/j.eclinm.2018.12.005>
* Nesi, J. (2020). The impact of social media on youth mental health: Challenges and opportunities. *North Carolina Medical Journal*, 81(2), 116–121. <https://doi.org/10.18043/ncm.81.2.116>
* Sadagheyani, H. E., & Tatari, F. (2021). Investigating the role of social media on mental health. *Mental Health and Social Inclusion*, 25(1), 41–50. <https://www.emerald.com/insight/content/doi/10.1108/MHSI-06-2020-0039/full/html>
* Shannon, A., Bush, K., & O'Loughlin, K. (2022). Problematic social media use in adolescents and young adults: A meta-analysis. *JMIR Mental Health*, 9(4), e33450. <https://mental.jmir.org/2022/4/e33450>
* Williams, D. L., & Park, S. (2023). Social media use among neurodivergent college students: Benefits, harms, and implications for education. *Information and Learning Sciences*, 125(1/2), 1–15. <https://www.emerald.com/insight/content/doi/10.1108/ILS-01-2024-0005/full/html>
* Cao, B., Lin, Z., & De Choudhury, M. (2024). Employing social media to improve mental health outcomes. *arXiv preprint* arXiv:2501.05621. <https://arxiv.org/abs/2501.05621>
* Cao, X., Wang, Y., & Zhang, Y. (2024). AI applications in mental health: A meta-analysis of dataset bias and cross-cultural applicability. *Journal of Affective Disorders*, 325, 45–56. <https://doi.org/10.1016/j.jad.2024.01.015>
* Chancellor, S., Kalantidis, Y., Pater, J. A., De Choudhury, M., & Shamma, D. A. (2017). Multimodal classification of moderated online pro-eating disorder content. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 3213–3226. <https://doi.org/10.1145/3025453.3025985>
* Chancellor, S., Lin, Z., Goodman, E. L., Zerwas, S., & De Choudhury, M. (2016). Quantifying and predicting mental illness severity in online pro-eating disorder communities. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1171–1184. <https://doi.org/10.1145/2818048.2819973>
* Coppersmith, G., Dredze, M., Harman, C., & Hollingshead, K. (2015). From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 1–10. <https://doi.org/10.3115/v1/W15-1201>
* Coppersmith, G., Leary, R., Whyne, E., & Wood, T. (2015). Quantifying suicidal ideation via language usage on social media. *Joint Statistics Meetings Proceedings, Statistical Computing Section*, 1–8.
* De Choudhury, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 128–137. <https://ojs.aaai.org/index.php/ICWSM/article/view/14432>
* Malgaroli, M., Choi, H., & Sirey, J. A. (2024). Natural language processing and mental health interventions: A systematic review. *JMIR Mental Health*, 11, e59479. <https://mental.jmir.org/2024/1/e59479>
* Mansoor, M., & Ansari, K. (2024). Early detection of mental health crises through artificial-intelligence-powered social media analysis: A prospective observational study. *Journal of Personalized Medicine*, 14(9), 958. <https://www.mdpi.com/2075-4426/14/9/958>
* Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, 16(1), 72–80. <https://doi.org/10.1016/j.tics.2011.11.018>
* Owen, D., Lynham, A. J., Smart, S. E., Pardinas, A. F., & Camacho Collados, J. (2024). AI for analysing mental health disorders among social media users: Quarter-century narrative review of progress and challenges. *Journal of Medical Internet Research, 26*, e59225. <https://doi.org/10.2196/59225>
* Teferra, B. G., Rueda, A., Pang, H., Valenzano, R., Samavi, R., Krishnan, S., & Bhat, V. (2024). Screening for depression using natural language processing: A literature review. *JMIR Mental Health*, 11, e55067. <https://www.i-jmr.org/2024/1/e55067>
* Aguirre, C., Harrigian, K., & Dredze, M. (2021). Gender and racial fairness in depression research using social media. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. <https://arxiv.org/abs/2103.10550>
* Calvo, R.A., Milne, D.N., Hussain, M.S. and Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5), 649–685. <https://psycnet.apa.org/record/2017-34756-003>
* Guntuku, S.C., Yaden, D.B., Kern, M.L., Ungar, L.H. and Eichstaedt, J.C. (2017). Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18, 43–49. <https://psycnet.apa.org/record/2017-54941-011>
* Shatte, A.B.R., Hutchinson, D.M. and Teague, S.J. (2019). Machine learning in mental health: a scoping review of methods and applications. *Psychological Medicine*, 49(9), 1426–1448. <https://pubmed.ncbi.nlm.nih.gov/30744717/>
* Cao, Y., et al. (2024). FairBelief – Assessing harmful beliefs in language models. *Proceedings of the 2024 Conference on Trustworthy Natural Language Processing.* <https://aclanthology.org/2024.trustnlp-1.3/>
* Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint* arXiv:1702.08608. <https://arxiv.org/abs/1702.08608>
* Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>
* Joyce, D. W., Kormilitzin, A., Smith, K. A., & Cipriani, A. (2023). Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *npj Digital Medicine*, 6, Article 6. <https://doi.org/10.1038/s41746-023-00751-9>
* Holtorf, A. P., et al. (2023). Ethical and legal considerations in social media research for health technology assessment: conclusions from a scoping review. *Frontiers in Digital Health*, 5, 1157017. <https://pubmed.ncbi.nlm.nih.gov/37842838/>
* Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36–43. <https://doi.org/10.1145/3233231>
* Moreno, M.A., Goniu, N., Moreno, P.S. and Diekema, D. (2013). Ethics of social media research: common concerns and practical considerations. *Cyberpsychology, Behavior, and Social Networking*, 16(9), 708–713.
* Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, 26(4), 2141–2168. <https://doi.org/10.1007/s11948-019-00165-5>
* Owen, M. J., et al. (2024). Dynamic consent: A patient interface for twenty-first century research networks. *European Journal of Human Genetics*, 32(3), 262–269.
* Kaye, J., Whitley, E. A., Lund, D., Morrison, M., Teare, H., & Melham, K. (2015). Dynamic consent: A patient interface for twenty-first century research networks. *European Journal of Human Genetics*, 23(2), 141–146. <https://doi.org/10.1038/ejhg.2014.71>
* Saeidnia, H. R., Hashemi Fotami, S. G., Lund, B., & Ghiasi, N. (2024). Ethical Considerations in Artificial Intelligence Interventions for Mental Health and Well-Being: Ensuring Responsible Implementation and Impact. *Social Sciences*, 13(7), 381. <https://doi.org/10.3390/socsci13070381>
* Andrade, L.H., Alonso, J., Mneimneh, Z., Wells, J.E., Al-Hamzawi, A., Borges, G., Bromet, E., Bruffaerts, R., de Girolamo, G., de Graaf, R., Florescu, S., Gureje, O., Hinkov, H.R., Hu, C., Huang, Y., Hwang, I., Jin, R., Karam, E.G., Kovess-Masfety, V., Levinson, D., Medina-Mora, M.E., Nakamura, Y., Oakley Browne, M.A., Ormel, J., Pennell, B.E., Posada-Villa, J., Sampson, N.A., Scott, K.M., Stein, D.J., Takeshima, T., Viana, M.C., Xavier, M., Zarkov, Z. and Kessler, R.C. (2014). Barriers to mental health treatment: results from the WHO World Mental Health surveys. *Psychological Medicine*, 44(6), 1303–1317. <https://pubmed.ncbi.nlm.nih.gov/23931656/>
* Beck, A.T. (1976). *Cognitive Therapy and the Emotional Disorders*. New York: International Universities Press.
* Clark, H., Day, K., Royal, P., Wright, P., Purvey, C., Norman, A., Murray, A., Lumsden, E., Peckham, K., Howells, K., Murray, P., Randall, V., Bayou, E., Blackwell, F., Barlow, J., Morton-Brown, J., Smith, K., Briggs, M., Lubrano, M., Jephcott, M., McGlone, F., Godfrey, K., Lord, S., McKaig, S., Smith, S., O'Neill, S., Brewis, T., Beswick, T., Beevers, V., & Veale, V. (2022). *The Mental Health of Children and Young People*. Children's Alliance, Exeter.[Nectar+4pure.northampton.ac.uk+4Worcester Research and Publications+4](https://pure.northampton.ac.uk/en/publications/the-mental-health-of-children-and-young-people?utm_source=chatgpt.com)
* Dunbar, R. I. M. (2014). The social brain: Psychological underpinnings and implications for the structure of organizations. *Current Directions in Psychological Science*, 19(1), 20–23. <https://www.researchgate.net/publication/269603581_The_Social_Brain_Psychological_Underpinnings_and_Implications_for_the_Structure_of_Organizations>
* Mental Health Foundation. (2022). Mental health statistics. [online]
* Gill, J. & Johnson, P. (2010). *Research Methods for Managers* (4th ed.). SAGE Publications Ltd.
* Sammut, C. & Webb, G. I. (2010). *Encyclopedia of Machine Learning*. Springer. <https://www.research.ed.ac.uk/en/publications/encyclopedia-of-machine-learning>
* Couldry, N., & Mejias, U. A. (2019). *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*. Stanford University Press. <https://www.researchgate.net/publication/343184089_The_Costs_of_Connection_How_Data_Is_Colonizing_Human_Life_and_Appropriating_It_for_Capitalism>
* Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*(2), 117–140. <https://doi.org/10.1177/001872675400700202>
* Goffman, E. (1959). *The Presentation of Self in Everyday Life*. Anchor Books.
* Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic Inquiry and Word Count: LIWC 2001*. Erlbaum Publishers. <https://www.researchgate.net/publication/246699633_Linguistic_inquiry_and_word_count_LIWC>
* NHSX. (2022). *Artificial Intelligence: How to get it right. Putting policy into practice for safe data-driven innovation in health and care*. NHS England.
* Katz, E., Blumler, J. G., & Gurevitch, M**.** (1973). Uses and gratifications research. *Public Opinion Quarterly*, *37*(4), 509–523. <https://academic.oup.com/poq/article-abstract/37/4/509/1816598?redirectedFrom=fulltext>
* Koko.ai. (n.d.). Koko: Mental health crisis intervention using AI. [https://www.koko.ai](https://www.koko.ai/)
* TalkLife. (n.d.). TalkLife: Peer support network for mental health. [online] Available at: [https://www.talklife.co](https://www.talklife.co/)
* Wysa. (n.d.). Wysa – AI for mental health. [online] Available at: [https://www.wysa.io](https://www.wysa.io/)

# Appendix A

Section A: Data Types, Sources and Collection

Table 22: Reddit Dataset Description Table

|  |  |
| --- | --- |
| Name of Dataset | Reddit Datasets |
| Description of Data | This dataset comprises user-generated submissions from five mental health-related subreddits: r/Depression, r/Anxiety, r/SuicideWatch, r/CasualConversations, and r/mentalhealth. It includes long-form text posts discussing personal experiences, emotional states, and mental health challenges. The data captures authentic, real-world expressions of psychological distress, casual interactions, and support-seeking behaviour, making it valuable for training AI models in mental health detection. |
| Location and/or Ownership *(Internal/External)* | academictorrents.com/details/1614740ac8c94505e4ecb9d88be8bed7b6afddd4 |
| Data Format (*Structured/Unstructured*) | Structured. |
| Data Collection Method | Secondary Data: The dataset was collected from public Reddit posts through community-curated archives. Posts prior to April 2023 were gathered using the Pushshift API, while later data was scraped by u/raiderbdev and curated by u/Watchful1. These were extracted, structured, and re-packaged for academic use. |
| Data Volume/Scale | For depression, anxiety, suicide and casual: 11,000 rows per subreddit and 7 columns. For r/mentalhealth: 10,000 rows and 7 columns. |
| Data Quality | Good quality: The dataset is well-structured, not much cleaning required besides formatting. |

Table 23: Twitter Depression Dataset Description Table

|  |  |
| --- | --- |
| Name of Dataset | Twitter Depression Dataset |
| Description of Data | This dataset consists of 20,000 English-language tweets labeled as “depressed” or “non-depressed.” It includes raw tweet text intended to support mental health research and natural language processing applications such as depression detection models. |
| Location and/or Ownership *(Internal/External)* | https://www.kaggle.com/datasets/infamouscoder/mental-health-social-media |
| Data Format (*Structured/Unstructured*) | Unstructured |
| Data Collection Method | Secondary Data: Tweets were collected using the Twitter API. Data labeling and feature extraction were done by the dataset creator. It includes raw, uncleaned tweets and can be supplemented with external features (e.g., emoji sentiment, LDA topics). |
| Data Volume/Scale | 20,000 tweets (text posts), each labeled as “depressed” or “non-depressed.” |
| Data Quality | Moderate to good: Data is raw and uncleaned, requiring preprocessing and possibly further labeling or filtering for certain use cases. Supplementary notebooks available for feature extraction. |

Table 23: Twitter Suicide Dataset Description Table

|  |  |
| --- | --- |
| Name of Dataset | Suicide Tweet Detection Dataset |
| Description of Data | This dataset contains tweets labeled as "suicidal" or "non-suicidal." It is intended to help in the early detection of suicidal ideation using NLP and machine learning. The dataset supports binary classification tasks in mental health analysis. |
| Location and/or Ownership *(Internal/External)* | https://www.kaggle.com/datasets/aunanya875/suicidal-tweet-detection-dataset |
| Data Format (*Structured/Unstructured*) | Unstructured |
| Data Collection Method | Secondary Data: Tweets were sourced using the Twitter API and manually labeled by annotators. The labeling process involved identifying suicidal language or expressions of intent. |
| Data Volume/Scale | The dataset contains 2,000 tweets: 1,000 labeled as suicidal and 1,000 as non-suicidal. |
| Data Quality | Good: The data is labeled and usable. |

Table 24: Behavioural Tweets Dataset

|  |  |
| --- | --- |
| Name of Dataset | Behavioural Tweet Detection Dataset |
| Description of Data | This dataset consists of tweets categorized into four emotional/mental states: Anxious, Lonely, Stressed, and Normal. Each file (CSV) contains tweets reflecting user sentiments, and the text has been cleaned using the NLTK library. This dataset supports sentiment and mental health analysis through social media language. |
| Location and/or Ownership *(Internal/External)* | https://www.kaggle.com/datasets/arshkandroo/behavioural-tweets |
| Data Format (*Structured/Unstructured*) | Unstructured |
| Data Collection Method | Secondary Data: Tweets were collected using the Twitter API and were preprocessed (cleaned text). The classification into behavioral categories was manually or algorithmically labeled, and split into separate CSV files by emotion type (e.g., Anxious\_Tweets.csv). |
| Data Volume/Scale | Each file contains approximately 8,000–9,000 cleaned tweets, totaling over 32,000 tweets. |
| Data Quality | Good: The tweets are pre-cleaned using the NLTK library. The dataset is well-organized into clearly labeled categories, though additional preprocessing may be needed for modeling. |

Section B: Attributes and Variables

Table 25: Reddit Attribute Table

|  |  |  |
| --- | --- | --- |
| Attribute/Variable | Description | Variable Type |
| author | Anonymised username of the Reddit user who created the post Reddit username of the post creator. Some are visible; others are pseudo-anonymised as [deleted]. | Nominal |
| title | title of post | Textual |
| selftext | Post itself under the text | Textual |
| score | Reddit score | Ratio |
| Created\_utc | Date and time and timezone | Interval |

Table 26: Twitter Depression Attributes

|  |  |  |
| --- | --- | --- |
| Attribute/Variable | Description | Variable Type |
| post\_id | Unique identifier for the tweet | Nominal |
| post\_created | Date and time when the tweet was posted | Interval |
| post\_text | Full textual content of the tweet | Textual |
| user\_id | ID of the user who posted the tweet | Nominal |
| followers | Number of followers the user had at the time of posting | Ratio |
| friends | Number of accounts the user was following (friends) | Ratio |
| favourites | Total number of tweets the user had liked at the time of posting | Ratio |
| statuses | Total number of tweets the user had posted before the current tweet | Ratio |
| retweets | Number of retweets | Ratio |

Table 27: Twitter Suicide Attributes

|  |  |  |
| --- | --- | --- |
| Attribute/Variable | Description | Variable Type |
| Tweet | Text content from Twitter reflecting user language, emotions, and thoughts. | Textual |
| Suicide | Label indicating whether the tweet is a "Potential Suicide post" or "Not Suicide post." | Nominal |

Table 28: Twitter Behavioural Attributes

|  |  |  |
| --- | --- | --- |
| Attribute/Variable | Description | Variable Type |
| Serial Number | Unique numeric index for each tweet (row). | Ordinal |
| Cleaned Tweet | Pre-processed tweet text with stopwords and symbols removed. | Textual |

# Appendix B: Project Code Repository (GitHub)

The full implementation of this project is available on GitHub and includes all code, models, and documentation used in the analysis, model training, and post-processing stages. The repository is structured to reflect the workflow of the study, from data cleaning to final classification evaluation.

Github: https://github.com/zoyaam2003/mentalhealth\_ai

Repository Contents:

* Data Cleaning + MentalBERT Model.ipynb  
  Preprocessing pipeline for Reddit and Twitter text data, tokenisation, and transformer training setup.
* Logistic Regression + TF-IDF.ipynb  
  Classical ML baseline for binary and multi-label classification using TF-IDF vectors.
* SVM + TF-IDF for Multi-label Mental Health.ipynb  
  Support Vector Machine model with TF-IDF features for multi-label classification.
* Random Forest + TF-IDF for Multi-label.ipynb  
  Random Forest implementation with engineered features and class balancing.
* MentalBERT + Post Processing Adjustments.ipynb  
  Fine-tuning the transformer model and implementing heuristic post-processing rules for improving precision/recall in suicide-class predictions.
* README.md  
  Contains a complete overview of the project structure, models used, objectives, ethical considerations, and how to run the code.

Usage

The GitHub repository is intended to make the project reproducible and accessible for further academic or ethical development. Scripts and notebooks are modular and include inline explanations. Dependencies are listed in requirements.txt (not shown in the appendix but included in the repo).

# Appendix C: Power BI Dashboard Snapshots

**A graph of different colored lines

AI-generated content may be incorrect.**

Figure 18: Average Distress Score by Month

A screenshot of a computer

AI-generated content may be incorrect.

Figure 19: 2016 Temporal Trend

A screenshot of a computer

AI-generated content may be incorrect.

Figure 20: 2015 Temporal Trend

A screenshot of a computer

AI-generated content may be incorrect.

Figure 21: Dashboard 1

A screenshot of a computer

AI-generated content may be incorrect.

Figure 22: Dashboard 2

A screenshot of a computer

AI-generated content may be incorrect.

Figure 23: Dashboard 3

A screenshot of a computer

AI-generated content may be incorrect.