universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Leveraging Natural Language Processing (NLP) for Mental Health: A Comparative Study of Sentiment Analysis Models

verfasst von | submitted by

Berna Franko BA

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 977
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Business Analytics
Betreut von | Supervisor:	ao. Univ.-Prof. Mag. Mag. Dr. Werner Winiwarter

Acknowledgements

I would like to express my profound gratitude to my supervisor, Univ.-Prof. Mag. Mag. Dr. Werner Winiwarter, whose guidance and support have been invaluable throughout the development of this thesis. His willingness to make time for my questions and his insightful advice have been crucial to my academic growth and the successful completion of this research.

I am also grateful to the faculty and staff at the University of Vienna, particularly within the Masters in Business Analytics program, for providing a supportive and enriching environment that has facilitated my learning and research endeavors.

Special thanks to my colleagues and fellow students for their collaboration and the stimulating discussions that have greatly contributed to my understanding of the subject matter.

In addition, I would like to acknowledge the authors and researchers whose work has informed and inspired this thesis. Their contributions to the fields of Natural Language Processing and mental health have laid the groundwork for my study.

Finally, I extend my heartfelt appreciation to my family and friends for their unwavering support and encouragement throughout this journey. Their belief in me has been a constant source of motivation.

Thank you!

Abstract

Natural Language Processing (NLP) offers promising avenues for mental health assessment by analyzing linguistic features in social media texts.

This study compares sentiment analysis models, including traditional machine learning algorithms (Logistic Regression, SVM, Random Forest) and advanced deep learning models (LSTM, BERT), to determine their efficacy in classifying mental health conditions such as depression, anxiety, and stress.

Utilizing a comprehensive dataset from platforms like Reddit and Twitter, the research assesses model performance based on accuracy, precision, and recall, while addressing ethical concerns associated with privacy.

The findings highlight the potential of sentiment analysis for early detection and continuous monitoring of mental health, advocating for its integration into personalized interventions and public health strategies. Challenges include data diversity and ethical considerations, necessitating improvements in model robustness and application protocols.

This study underscores the importance of leveraging NLP for timely mental health interventions, contributing to enhanced diagnostic accuracy and support systems.

Keywords: *NLP, Sentiment Analysis, Mental Health, Depression, Anxiety, Stress, Machine Learning, Deep Learning, Social Media, Accuracy, Precision, Recall, Early Detection*

Kurzfassung

Das ist eine deutsche Kurzfassung meiner in Englisch verfassten Masterarbeit.

Die Verarbeitung natürlicher Sprache (NLP) bietet vielversprechende Ansätze für die Bewertung der psychischen Gesundheit durch die Analyse linguistischer Merkmale in sozialen Medien.

Diese Studie vergleicht Modelle zur Sentiment-Analyse, einschließlich traditioneller maschineller Lernalgorithmen (logistische Regression, SVM, Random Forest) und fortschrittlicher Deep-Learning-Modelle (LSTM, BERT), um ihre Wirksamkeit bei der Klassifizierung psychischer Gesundheitszustände wie Depression, Angst und Stress zu bestimmen.

Unter Verwendung eines umfassenden Datensatzes von Plattformen wie Reddit und Twitter bewertet diese Arbeit die Modellleistung anhand von Genauigkeit, Präzision und Recall, wobei ethische Aspekte des Datenschutzes berücksichtigt werden.

Die Ergebnisse unterstreichen das Potenzial der Sentiment-Analyse für die Früherkennung und kontinuierliche Überwachung psychischer Gesundheit und sprechen für die Integration in personalisierte Interventionen und öffentliche Gesundheitsstrategien. Herausforderungen umfassen die Vielfalt der Daten und ethische Überlegungen, die Verbesserungen der Modellrobustheit und der Anwendungsprotokolle erfordern.

Diese Studie hebt die Bedeutung der Nutzung von NLP für zeitnahe Interventionen im Bereich psychischer Gesundheit hervor und trägt zu verbesserter diagnostischer Genauigkeit und zu unterstützenden Versorgungssystemen bei.

Schlüsselwörter: *NLP; Sentiment-Analyse; psychische Gesundheit; Depression; Angststörung; Stress; Maschinelles Lernen; Deep Learning; Soziale Medien; Genauigkeit; Präzision; Recall; Früherkennung.*

Contents

1 Introduction

The majority of people suffer from mental health disorders like depression, anxiety, and stress disorders, which are among the most urgent public health concerns in the world and cause enormous personal and societal burdens. Detecting in advance and monitoring mental health issues is still challenging despite improvements in clinical diagnosis and treatment, continuously depending on self-reports and professional interviews that can be laborious, subjective, and narrowly focused.

Nowadays, much of the textual data highlights users' ideas, feelings, and mental states due to the growth of social media sites like Reddit and Twitter. This digitalization presents a special opportunity to make measurements, using non-intrusive Natural Language Processing (NLP) tools for mental health screening. In particular, sentiment analysis with NLP has demonstrated potential in recognizing the power of language regarding to the mental health concerns, allowing for early detections and continuous monitoring. This technique outputs views, emotions, and attitudes from text.

Sentiment analysis models have became better with the developments in deep learning and machine learning. Even though they have shown success with traditional algorithms like Random Forest, Support Vector Machines (SVM), and Logistic Regression, there are still challenges in observing linguistic subtleties. On the other hand, deep learning models such as Bidirectional Encoder Representations from Transformers (BERT) and Long Short-Term Memory (LSTM) networks have demonstrated better outcomes in comprehending context and understand the linguistic signals, which makes them ideal for mental health classification tasks.

Along with these advancements, there is still a significant gap in the methodical comparison of these various models' outcome in the field of sentiment analysis for mental health. For the advancement of automated mental health assessment tools, it is significant to observe which models work better/best and what elements, like data quality, preprocessing methods, and model architecture, reflect their accuracy.

This thesis aims to fill this gap by conducting a comparative analysis of various machine learning models applied to social media data to classify mental health related sentiments. The study evaluates and compares the overall performance of traditional algorithms and deep learning approaches, with the aim of analyzing the most effective methods to support early detection, ongoing monitoring, and timely intervention. Ultimately, this research aims to contribute to the development of accessible, objective, and scalable tools that can aid mental health professionals and raise awareness of individuals in order to manage mental health more proactively.

1.1 Problem Statement

Over the past decade, social media platforms have been extensively used to compile datasets for various mental health issues. Even though there have been numerous advancements in mental health diagnostics and treatment, there is still a significant gap in early detection and continuous monitoring of mental health conditions. Traditional methods mostly rely on self-reported surveys and clinical interviews, which can be time-consuming and subject to bias.

Hengle et al. (2024) have pointed out that thanks to the developments of web technologies, social media platforms such as Reddit and Twitter have become the pri-

mary source for mental health support and information exchange. These platforms provide a secure anonymous environment that facilitates in-depth discussion about mental health. As a result, there is a focus on "digital psychiatry," which delves into mental health topics and language usage on these platforms in order to enhance the discovery, understanding, and detection of mental health issues. However, despite significant efforts, many concerns related to these resources persist (Hengle et al., 2024). The integration of sentiment analysis and mental health care promises to bridge this gap by providing real-time, scalable, and objective assessments of mental health statuses from everyday textual communications (B. Liu, 2015).

1.2 Methodology

This study aims to compare the performance of various machine learning models including traditional algorithms (Logistic Regression, SVM, Random Forest) and advanced deep learning models (LSTM, BERT) in accurately classifying mental health related sentiments from social media texts.

The key research question is: *Which machine learning model (Logistic Regression, SVM, Random Forest, LSTM, BERT, etc.) provides the most accurate classification of mental health related sentiments, and what are the key factors influencing their performance?*

Specifically, the research seeks to evaluate whether NLP techniques combined with these models can accurately determine early signs of mental health conditions, and to determine key factors such as preprocessing, feature engineering, and model architecture that affect their performance. Ultimately, the goal is to show an outcome how NLP based sentiment analysis can present more accessible, objective, and mental health assessments, early detection and ongoing monitoring. Methodologically, the study involves collecting and preprocessing diverse datasets from platforms like Reddit and Twitter using Python libraries, and evaluating model performance based on metrics like accuracy, precision, and recall.

We refined the language (English and translation to German) and produced exploratory model outputs using AI tools (ChatGPT, OpenAI GPT-5.1).

1.3 Literature Review

The significance of this study lies in its potential to enhance the accuracy and efficiency of mental health diagnostics through the application of sentiment analysis. By comparing the performance of various machine learning (ML) models, this research aims to identify the most effective approaches for classifying mental health statuses based on textual data. The findings of this study could inform the development of real-time mental health monitoring tools, contribute to the design of supportive mental health chatbots, and provide valuable insights for academic research on mental health patterns.

Natural Language Processing (NLP) has gained considerable attention for its potential to analyze social media text data in identifying linguistic markers associated with mental health issues (Tang, 2024). Mental health conditions such as anxiety, stress, depression, self-harm, and suicidal ideation are widespread global concerns that significantly impact public health (X. Wang et al., 2024). Early detection through

timely interventions can mitigate the progression of these conditions, thereby improve individual outcomes and reduce long-term risks.

According to the World Health Organization (WHO & ILO, 2022), mental health is an increasingly critical aspect of overall well-being, affecting millions worldwide. The rise of digital communication platforms offers new opportunities to monitor mental health via textual data analysis. Sentiment analysis, a subfield of NLP, enables the extraction of opinions, emotions, and attitudes from text, facilitating the classification of mental health related sentiments (Calvo et al., 2017). Calvo et al. (2017) highlight that mental health applications rely on interdisciplinary collaboration between researchers and practitioners from fields such as computational linguistics, human-computer interaction, and mental health services.

Sentiment analysis, also referred to as opinion mining, involves determining and categorizing opinions and emotions expressed in text (Ravi & Ravi, 2015). Basically, it helps leveraging language patterns in social media posts to detect symptoms of depression, anxiety, and other disorders (Guntuku et al., 2019).

We can say that the newest improvements in ML and NLP have significantly improved sentiment analysis performance. Machine learning algorithms such as SVMs and deep learning models, which capture linguistic nuances more effectively, have largely replaced traditional dictionary-based approaches that rely on pre-established word lists (Pang & Lee, 2008). These developments enable more accurate, scalable, and real-time analyses of large online datasets, assisting individuals who might be at risk. Nonetheless, deploying NLP in mental health contexts usually raises ethical and privacy concerns that require careful consideration (Guntuku et al., 2019).

Beyond individual detection, sentiment analysis can track changes over time, offering early warnings. Integrating textual data with other sources can provide a comprehensive view of mental health status, facilitating early and personalized interventions. At the population level, sentiment analysis can reveal broader mental health trends, informing policy and resource allocation. However, challenges such as data diversity, preprocessing inconsistencies, and evaluation metrics especially with imbalanced datasets must be addressed to improve reliability (Alahmadi et al., 2025).

Linguistic challenges, biases from preprocssing and hyperparameter tuning might effect model robustness. Ongoing research emphasizes the need for continuous refinement of models to enhance accuracy and fairness (Alahmadi et al., 2025).

Finally, online support communities and forums serve as vital resources for individuals seeking help. (Calvo et al., 2017) note that online peer support groups have become one of the most accessible and reliable platforms for people to share their mental health concerns and seek for some help, especially with the younger populations.

1.4 Research Objectives

1. Comparing classical ML (LR, SVM, RF) and deep models (CNN, LSTM, BERT) for mental health text classification.

2. Quantifying the impact of representations (TF–IDF, Word2Vec, sentence-transformer embeddings) and imbalance handling in minority/risk classes.

3. Evaluating risk-aware operation (testing, class specific thresholds) for safety critical labels (e.g., *Suicidal*).

1.5 Contributions

1. A unified, reproducible comparison across representations and models on a large-scale mental health sources.

2. Suggesting transparent, lightweight benchmarks for transformer performance.

3. Providing a risk-aware evaluation method (macro/weighted F1, PR curves, calibration) for high-impact classes.

1.6 Scope, Ethics, and Reproducibility

This thesis focuses on analyzing publicly available, identified texts; models support screening/monitoring and do not provide diagnosis. We follow privacy and crisis safe practices, report subgroup metrics where feasible, and release seeds, package versions, and configuration details for reproducibility.

1.7 Organization of the Thesis

This section provides a guide for the reader to show the logic of the work and what each chapter presents.

2. **Social Media:** Defining social media as data generating, technologically support communication systems; reviewing changes in infrastructure (mobility, datafication, AI mediation), interaction styles (masspersonal), and psychological self-effects. It emphasizes the need for text analytics and clarifies how texts from platforms are an insightful indicator for mental states.

3. **Mental Health:** Demonstrating biopsychosocial and ecological models, clinical/medical assessment (including "supportive" monitoring), literacy/stigma, structural and digital effects, and key conditions. It concludes by discussing the role of language in mental health and its connection to NLP.

4. **Natural Language Processing (NLP):** Formalizing the task (label space, problem setup, ethics/scope), detailing preprocessing (tokenization, normalization, identification) and text representations (TF–IDF, Word2Vec, sentence/-transformers), and defining evaluation criteria (accuracy, macro/weighted F1, PR, calibration). This chapter focuses on the analytical toolkit utilized in further experiments.

5. **Exploratory Data Analysis (EDA):** Describing the corpus, class imbalance, and notable linguistic features; providing descriptive figures (distributions, correlations, word clouds). EDA justifies modeling choices (e.g., class weighting, thresholds) and informs hyperparameter ranges.

6. **Methods:** Identifying models (LR/SVM, trees/ensembles, CNN/LSTM, BERT), training procedures, and risk-aware evaluation (per-class PR, cali-

bration, class specific thresholds for high-risk labels such as Suicidal). Outputs include a clear protocol that supports comparability between models.

7. **Large Language Models (LLMs):** Describing LLMs relation to supervised models; covering safety, privacy, fairness, and governance; outlining prompting/RAG and uncertainty handling. This frames responsible use and future deployment pathways.

8. **Findings:** Reporting comparative results with tables/figures, interpreting error patterns (especially minority classes), and debating compromises between accuracy, readability, and cost.

9. **Conclusion**: Summarizing contributions, limitations, and actionable future work (e.g., imbalance mitigation, domain adaptation, governance). It links back to the objectives and underlines where the approach can be extended in practice.

Building on the research objectives, the next chapter examines *social media* as the primary context and data source for this study. We discuss platforms and their infrastructural changes (mobility, data sharing, AI mediation), describe interaction styles and psychological self-effects that shape language use, and outline ethical limits for the use of public text. This provides the theoritical and practical basis for addressing online language as empirical signals of mental states, which the subsequent NLP chapters operationalize.

2 Social Media

Social media is widely understood as digital communication platforms that empower users to create, share, and exchange content while taking part in networked interactions with others. However, as Carr and Hayes (2015) emphasize, defining social media is not only just listing popular tools such as Facebook, Instagram or Twitter.

The foundation of social media lies not in specific applications, but in the **interactive, participatory, and networked** nature of communication. These sociotechnical qualities remain relevant as technologies evolve, extending beyond individual platforms, tools, or software.

2.1 Developments in Social Media

The continuous development of both the technical infrastructure and social dynamics of social media will profoundly impact communication research and applications. As Carr and Hayes (2015) suggest, these new developments will reshape not only the nature of online interaction but also the theoretical frameworks and methodologies academics use to study them.

2.1.1 Transforming Infrastructures: How Data and Mobility Affect Social Media

The core elements of social media and the internet are rapidly evolving. With the use of mobile technologies such as wearable devices, tablets, and smartphones on the rise, the primary means of reaching the internet has moved from traditional desktop platforms to portable, application-based interfaces. Thanks to the *Internet of Things*, where connected objects such as watches, cars, and even home appliances blur the line between online and offline activities, this mobility is allowing social media to become embedded in daily life.

Meanwhile, the computational architecture of these platforms is becoming increasingly *data-driven*. Also, the emergence of Web 3.0, commonly referred as the *Semantic Web*, offers highly personalized user experiences using ML, large-scale data processing, and adaptive algorithms. With the aim of predicting preferences and maximize engagement, many platforms currently use advanced recommendation systems that analyze users' networks, interactions, and activities. As these systems evolve, social media conversations may increasingly involve artificial agents or algorithmic equivalents; consumers may not always be able to distinguish automated programs from human participants. Current technologies like Siri, Google Now, and conversational AI systems such as Cleverbot illustrate this emerging trend of *AI-mediated communication* (Carr & Hayes, 2015).

Moreover, these developments extend beyond communication to large-scale predictive analytics. As Chauhan et al. (2021) states, social media data are used as predictive frameworks capable of utilizing sentiment analysis and ML to forecast real-world events such as election outcomes. These platforms process large volumes of user generated content to extract social opinions and behavioral patterns, thereby transforming ordinary communication into measurable and useful data streams. This development reflects the increasing incorporation of deep learning architectures, such as recurrent and convolutional neural networks, into the analytical core of social media platforms, enabling more context-sensitive and effective data interpretation.

Consequently, distinguishing between human communication and computational inference is becoming increasingly difficult, altering the way meaning, emotion, and influence spread in digital environments (Chauhan et al., 2021).

2.1.2 Changes in Interactions: Constant, Massive, and Masspersonal Communication

The nature of social interaction on digital platforms is changing significantly, combining aspects of *mass communication* and *interpersonal exchange*. As noted by Carr and Hayes (2015), this evolving "masspersonal" environment renders interpersonal communication to reach wider audiences with unprecedented access and enables messages initially intended for broad audiences to be received and interpreted in an extremely personal way.

Thus, social media encourages interactions that function both individually and collectively. Posts, stories, and tweets aimed at a broad audience can still appear personalized and encourage the establishment of a close relationship between users and content creators. The emergence of algorithmically controlled or group managed social media profiles that mimic authentic, unique voices contributes to the blurring of communication boundaries. As a result, users may perceive genuine intimacy even when engaging with automated systems or coordinated branding teams, an effect that echoes the "illusion of intimacy" described in parasocial interaction theory (Carr & Hayes, 2015).

These developments raise critical *ethical concerns*. As algorithmic and automated agents increasingly simulate human interaction, they have the capacity to evoke emotional responses, influence consumer behavior, and even shape personal decision-making. This form of mediated persuasion extends beyond conventional targeted advertising to encompass social environments in which algorithms generate seemingly interpersonal messages tailored to user data and preferences. Such interactions may inadvertently foster dependence, manipulation, or emotional attachment toward nonhuman entities, highlighting the need for clearer ethical frameworks and digital literacy in online engagement.

These changing modes of interaction increasingly shape not only how people communicate but also how they perceive themselves and others online.

2.2 Psychological and Self-Effects of Social Media Use

Social media has a strong impact on users' psychological mental states and self-perceptions, in addition to structural and behavioral changes. Recent studies introduced the concept of *self-effects*, which is based on the cognitive, emotional, and attitudinal consequences that message creators experience from their own online expressions (Valkenburg, 2017). Self-effect theory, unlike previous communication theories that focused on the effects of media messages on recipients, acknowledges that producing and disseminating content can have an equal effect on the sender.

According to Valkenburg (2017), self-effects emerge through four primary mechanisms: *self-persuasion*, where expressing an opinion strengthens one's own belief; *self-concept change*, where self-presentations shape identity; *expressive writing*, where online disclosure supports emotional well-being; and *political deliberation*, where engagement in public discourse reinforces cognitive engagement and civic behavior. The opportunities offered by social media are scalability, asynchronous

communication, and manageable hints. It allows users to improve how they present themselves, anticipate criticism, and absorb other people's reactions, which in turn strengthens these self-effects in the online environment.

In other words, in addition to social media communication channels, they function as interactive psychological feedback areas and enable users to continuously co-construct their own self-concepts through expression, reflection, and perception.

2.3 Effects of Social Media on People and Social Organizations

While the development of social media has brought about unprecedented opportunities for communication, participation, and access to information, it has also raised significant concerns regarding its effects on both individuals and society. Zeitel-Bank and Tat (2014) argue that social media represent a double-edged phenomenon: they enhance global communication, learning, and democratization, yet they simultaneously pose serious challenges for psychological well-being, social behavior, and data security.

Societal Level

Social media have transformed traditional one-way communication models into interactive, participatory systems. Information flows are no longer controlled by centralized authorities, as individuals can directly share opinions and mobilize others globally. This shift fosters democratic participation and transparency but also facilitates misinformation, privacy risks, and cybercrime. As Zeitel-Bank and Tat (2014) note, the rise of user-generated content and the growing influence of digital opinion leaders blur boundaries between journalism, marketing, and everyday discourse, reshaping how societies form public opinion.

Individual Level

Pervasive social networking affects emotional health, identity construction, and relationships. Drawing on neuroscience perspectives, Zeitel-Bank and Tat (2014) suggest that continuous engagement in virtual environments may influence attention, empathy, and emotion regulation. Through neuroplasticity, repeated online interactions can encourage behaviors such as shorter attention spans, intensified social comparison, and dependence on digital feedback loops. Reduced accountability and anonymity are also associated with cyberbullying, deception, and diminished self-control.

These individual-level effects are particularly visible in the rise of influencer culture, where exposure to idealized imagery shapes users' self-perceptions and emotional well-being.

This pattern is further reinforced by evidence that *risky ICT use* and *moral disengagement* are strong predictors of cyberbullying behaviors (L. Chen et al., 2017). Together, these findings highlight how social media affordances can both facilitate connection and magnify harmful behavioral tendencies.

Emotional dynamics are central online: rapid, affect-laden exchanges can amplify both supportive and harmful outcomes. Given social media's permanence in everyday life, Zeitel-Bank and Tat (2014) call for responsible and reflective use to ensure

benefits outweigh risks.

2.3.1 Influencer Culture and Appearance Anxiety

Recent studies have discussed the psychological effects of influencer culture, particularly in terms of social media users' concerns about their self-perception and appearance. According to Deng and Jiang (2022), exposure to idealized images shared by both human influencers (HIs) and virtual influencers (VIs) can significantly increase users' appearance anxiety. In an experimental study involving young women aged 18 to 35, participants reported feeling more concerned about their appearance when viewing photos of influencers, whether computer-generated or real. Remarkably, women respondents who examined virtual influencers reported feeling less concerned about their appearance than those who viewed real influencers. This suggests that the perceived artificiality of VIs may directly reduce the tendency to compare oneself (Deng & Jiang, 2022).

The study highlights that social media amplify mechanisms of *upward social comparison*, where users evaluate themselves against idealized beauty standards frequently presented online. These comparisons, often subconscious, reinforce unrealistic aesthetic ideals and contribute to negative emotional outcomes such as self-objectification, shame, and low self-esteem. The findings show that user generated and algorithmically curated visual content fosters environments where, based on appearance, validation dominates social interaction, which can increase mental health risks such as anxiety and depression. Consequently, Deng and Jiang (2022) emphasize the importance of fostering digital literacy and critical awareness to mitigate the psychological risks associated with influencer content.

As evidenced by the psychological and social effects of online behavior, digital participation and mental health are closely linked. Understanding the impact of social media on well-being is crucial, as social media often reinforces mechanisms such as comparison, seeking approval, and emotional contagion. Therefore, the next section examines the concept of mental health itself, the main factors affecting mental health, and how the digital environment influences these characteristics (Deng & Jiang, 2022).

2.4 NLP for Social Media Data (Bridge to Chapter 7)

Natural language processing (NLP) can identify early indicators of mental health risk on social media streams. As Coppersmith et al. (2018) mentioned, these studies demonstrate that deep learning classifiers applied to posts from the previous six months, particularly for suicide tendencies, can distinguish users who attempted suicide from a matched control group with promising ROC performance months before the crisis. This supports large-scale screening and referral processes. Furthermore, the studies emphasize ethical requirements (consent, privacy, participant and non-participant approaches). In this thesis, we treat social media texts as analyzable signals and implement NLP techniques in subsequent sections (preprocessing, representations, supervised models). We address governance issues in our security section.

2.5 Summary

In conclusion, the evolution of social media brings about complex changes in the areas of interaction, infrastructure, and psychology. As digital platforms become more mobile, algorithmic, and interactive systems, communication is becoming more individualized and widespread. Consequently, examining social media requires a holistic perspective that considers interpersonal, social, and technological factors influencing how people interact, communicate, and construct their identities in online environments.

Therefore, in the following chapter, we will examine how these psychological and communication systems interact with broader concepts of mental health in digital societies.

3 Mental Health

Having awareness of mental health has become increasingly important in public and scientific discourse. Mental health is now recognized as a cornerstone of overall well-being rather than a separate or secondary domain. It involves not only the absence of mental illness but also the presence of psychological balance, social connectedness, and emotional resilience (Bhugra et al., 2012).

As Bhugra et al. (2012) emphasize, individuals with good mental health can better maintain their emotional balance under pressure and, most importantly, build lasting relationships and adapt to change. The National Collaborating Centre for Mental Health (2011) further define mental health within a clinical and societal framework, describing it as a continuum that encompasses conditions such as depression, anxiety, and post-traumatic stress disorder. Complementing these perspectives, the concept of *mental health literacy "knowledge and beliefs about mental disorders that help to identify, treat, or prevent them"* links understanding with concrete action (Jorm, 2012). Surveys show persistent gaps in public recognition of disorders, knowledge of effective help, and first-aid skills; however community, school, web-based, and Mental Health First Aid interventions can improve these outcomes over time (Jorm, 2012). When these perspectives come together, it becomes clear that mental health is both an individual and a societal resource, influenced by biological, psychological and social variables, and that improving it requires integrated strategies that encompass both public health systems and individual care.

3.1 Dimensions and Determinants of Mental Health

Mental health arises from the continuous interaction of biological, psychological, and social factors, reflecting what (Engel, 1977) conceptualized as the *biopsychosocial model* of health. This framework posits that well-being cannot be understood solely in biological terms but must also account for cognitive, emotional, and environmental influences. On a biological level, genetic predispositions, neurochemical imbalances, and chronic physical illnesses can significantly shape an individual's vulnerability or resilience to mental disorders. Psychological factors, such as coping strategies, self-efficacy, and the ability to regulate emotions, also play a critical role in maintaining mental stability. (Ryff, 1989) emphasized that psychological well-being extends beyond the absence of distress to include self-acceptance, purpose in life, and personal growth. Social determinants, including socioeconomic status (SES), education, and access to supportive networks, further influence mental health outcomes. (Marmot, 2005) argued that inequality, marginalization, and social exclusion are among the strongest predictors of poor mental health, highlighting the need for policies that address structural disparities. Complementing this view, (Antonovsky, 1996) introduced the concept of *salutogenesis*, emphasizing a sense of coherence and meaning as protective resources that promote psychological resilience despite adversity. Together, these perspectives underline that mental health is a dynamic state shaped by the complex interplay between biological mechanisms, individual psychological processes, and the broader social environment.

3.1.1 Explanation of the Biopsychosocial Model

The biopsychosocial model as shown in Figure 1, first conceptualized by (Engel, 1977), illustrates how mental health is shaped by the continuous interaction of **bi-**

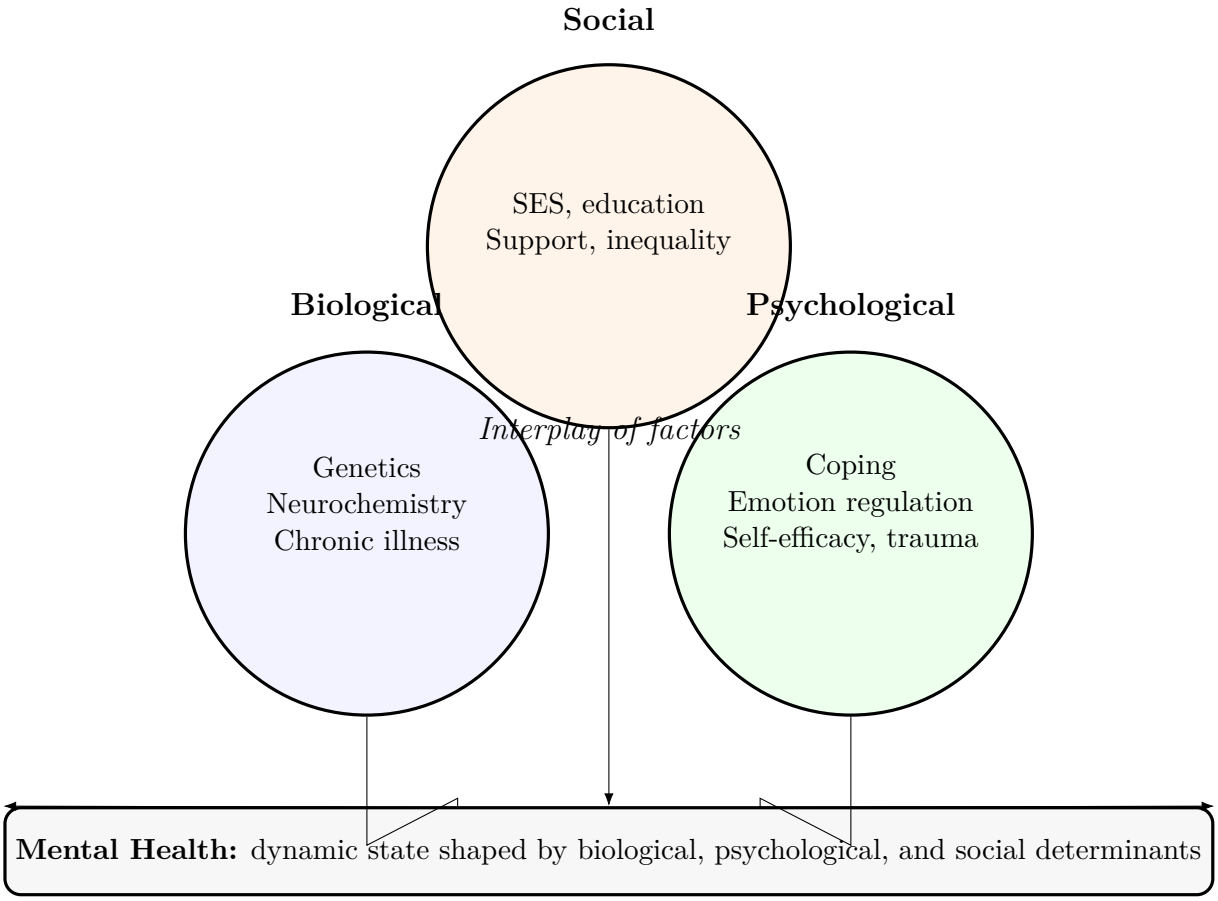

Figure 1: Biopsychosocial model with separated domains and contributions to mental health

ological, **psychological**, and **social** determinants. Rather than viewing mental health as a purely medical or psychological issue, the model emphasizes a dynamic, multidimensional system in which biological predispositions, individual coping mechanisms, and sociocultural contexts jointly influence well-being and vulnerability to mental disorders.

From the *biological* perspective, genetic inheritance, neurochemical balance, and chronic health conditions contribute to an individual's baseline susceptibility to mental health difficulties. The *psychological* domain encompasses cognitive and emotional processes such as resilience, self-efficacy, coping strategies, and trauma responses, all of which affect how people interpret and manage stress. The *social* dimension includes factors like socioeconomic status (SES), educational access, stigma, and the quality of community or family support networks, which can either buffer or exacerbate psychological strain (Marmot, 2005).

Reflecting the balance and interaction of biological, psychological, and social influences, mental health is a *dynamic state* that lies at the intersection of these three areas. This integrative view supports the notion of *salutogenesis* proposed by (Antonovsky, 1996), emphasizing the maintenance of well-being rather than merely the treatment of illness. Similarly, (Ryff, 1989) links personal growth, purpose, and autonomy to mental health outcomes. Together, these perspectives underline that sustainable mental health interventions must address not only individual symptoms but also the broader ecological and social contexts in which individuals live.

Ecological Perspective

Building on biopsychosocial accounts, an ecological view situates determinants across nested systems—micro (family, peers), meso (links between settings), exo (institutions such as media, workplaces), and macro (policy, norms) and stresses that cross-level interactions and cumulative risks shape outcomes. Reviews of mental health research using Bronfenbrenner's framework show that studies which explicitly model interactions within and between systems yield the most actionable guidance (i.e., what works for whom, for which outcome), whereas merely listing factors at several levels tends to produce broad, non-specific advice (Eriksson et al., 2018).

3.2 Clinical Assessment and Diagnosis

Clinicians diagnose mental health problems through a staged, clinician-led process that integrates multiple sources of evidence. First, therapists obtain a structured history and evaluate symptoms against recognized diagnostic criteria, then complement this with standardized self-reports and ecological momentary assessment that capture fluctuations in daily life. Where appropriate, they add objective signals from digital tools (e.g., heart rate/variability, sleep, activity, speech or smartphone use) to monitor severity and change over time. These data streams support outcome monitoring and "supportive monitoring", allowing the clinician to adjust the intervention when predefined targets are not met (for example, modifying psychotherapy tasks or changing medication). Importantly, algorithmic scores and app outputs function only as screening and monitoring aids; final diagnosis and individualized treatment planning require professional judgment that considers context, comorbidity, somatic differentials, and risk (including suicidal cases). In short, E-Health expands the assessment and tracking, but does not replace the therapist's diagnostic

role (Lüttke et al., 2018).

3.3 Mental Health Literacy

Mental health literacy (MHL) connects knowledge with action; it comprises the following (Jorm, 2012):

(i) Knowledge of ways to avoid mental disorders;

(ii) Detection of developing disorders;

(iii) Knowledge of therapies that are proven and options for obtaining assistance;

(iv) Understanding of practical self-help for milder problems; and

(v) First-aid skills to assist in crisis.

Evidence across countries shows common shortfalls in correctly recognizing disorders (e.g., depression vs. "stress"), delays in help-seeking (often years between onset and treatment), and skepticism toward some recommended treatments; these gaps are most consequential in adolescence and early adulthood when many disorders first emerge (Jorm, 2012). At the same time, the public frequently endorses and uses self-help strategies; population data suggest "overlapping waves of action," with readily available self-help peaking in mild distress and professional care increasing with severity (Jorm, 2012).

MHL can be improved, for example, rigorous evaluations report benefits from entire community campaigns (e.g., depression literacy, attitude shifts, reduced suicidal acts), based on school and university education, Mental Health First Aid training (better knowledge, confidence, and helping behaviour), and high-quality information websites that can even reduce symptoms in some users (Jorm, 2012). Culturally adapted approaches are essential to address diverse explanatory models and language barriers. As a policy lever, raising MHL supports prevention, earlier intervention, greater adherence to care based on evidence, and broader public support for adequately resourced services (Jorm, 2012).

3.3.1 Social Distance: Causes, Consequences, and Coping Strategies

Social distance is people's willingness to avoid contact with someone who has a mental disorder, which is a core, measurable facet of stigma with distinct predictors and levers (Jorm & Oh, 2009). Key points:

(i) *Measurement and validity:* Social distance is a reliable, largely one-dimensional measure, although it correlates with reported real-world contact.

(ii) *Who desires more distance:* Higher in older than younger adults; declines across adolescence; only weakly linked to lower education; large cross-national differences exist.

(iii) *Experiences that shift distance:* Sensational media (such as violent situations) may increase distance; human contact (and one's own lived experience) generally decreases distance; but clinicians may not exhibit less distance despite significant interaction.

(iv) *Which presentations elicit more distance:* There is a typical hierarchy: drug dependence disorders > schizophrenia > depression/anxiety; the same symptoms elicit more distance when portrayed in men than women.

(v) *Labels and help-seeking:* Diagnostic labels can increase distance (effects vary by label and familiarity). Among the general public, knowing someone sought professional help can sometimes increase distance; among more knowledgeable groups (family members, professionals), *not* seeking appropriate help can increase distance.

(vi) *Causal beliefs:* Endorsement of claims of "brain disease" or "weakness of character" has been associated with increased distance; there is no discernible correlation with "biochemical imbalance"; the results of genetic and psychological explanations are inconsistent; public opinions are often multi-causal.

(vii) *What works:* Planned interventions (e.g., contact-based education, Mental Health First Aid) can reduce social distance with effects that persist for months; future work should track behaviour change, not only vignette responses.

Contact-based education (e.g., Mental Health First Aid) not only raises knowledge but can reduce desired social distance in population studies (Jorm & Oh, 2009).

3.4 Environmental and Structural Determinants

Mental health is influenced by both personal risk as well as external circumstances: Evidence gathered by a quasi-experimental panel relates shifts in Emergency Department (ED) visits, self-reported days with bad mental health, and suicides to temporary weather shocks and local system capacity (Mullins & White, 2019). Key external determinants include:

- *Ambient temperature:* Nonlinear associations show more suicides and worse mental health on hotter (and some colder) days, net of location-by-month and year effects (Mullins & White, 2019).
- *Insurance coverage rules:* Strong mental health parity laws (covering mental health on the same terms as physical health) are associated with lower suicide rates, suggesting policy can buffer environmental risk (Mullins & White, 2019).
- *Provider availability*: Residence in mental health professional shortage areas signals limited access and may exacerbate temperature in mental health sensitivity (Mullins & White, 2019).
- *Treatment infrastructure:* Greater density of substance-abuse treatment centers is linked to fewer suicides and drug deaths, indicating system capacity matters for crisis prevention (Mullins & White, 2019).
- *Local resources:* County income moderates vulnerability, plausibly via ability to afford cooling/heating and to access timely care (Mullins & White, 2019).

These environmental and structural levers complement biopsychosocial determinants: Modifying exposure (e.g., heat mitigation) and strengthening systems (coverage, providers, treatment capacity) can reduce population risk even when individual-level vulnerabilities persist (Mullins & White, 2019). These external conditions set

the stage for technology and mediated exposures at work and online, which we examine next.

Ecological analyses recommend testing cross-level effects (e.g., how school/work policies or media exposure [exo/macro] modify family or peer processes [micro/meso]) rather than treating each level in isolation (Eriksson et al., 2018).

3.4.1 Family Stress and Adolescent Mental Health During COVID-19

Longitudinal U.S. research implies a family-stress pathway that starts with pandemic stress and ends with parents' symptoms, parenting, and perceived teenage mental health (Morgan et al., 2025). Elevated COVID-19 stress has been correlated with higher levels of anxiety among parents and depressive symptoms, which may have contributed to more poor parenting. Numerous hazards were associated with different practices:

(i) The connection with subsequently perceived depression among adolescents was mediated by inconsistent discipline;

(ii) The association with subsequently experienced adolescent anxiety was mediated by psychological control; and

(iii) Increased child-parent relationship quality did not overcome the effects of psychological control, but it did reduce the relationship between inconsistent discipline and later perceived teenager depression/anxiety.

Practically, pair mental health supports for parents with skills that promote consistent, non-coercive discipline and making parent-adolescent relationship quality much stronger during large-scale stressors (Morgan et al., 2025).

3.4.2 Population Groups and Stressors During COVID-19

Distancing and lockdowns altered routine in ways that made particular populations more vulnerable (Chugh, 2020):

(i) *Older adults:* Increased risk of anxiety and depression due to increased fear and depressed mood combined with the salience of medical vulnerability, and the sudden cessation of regular social interactions (walks, parks, prayers, visits);

(ii) *Children and adolescents:* Increased time spent indoors and significant increases in online learning and entertainment; reduced face-to-face interaction outside the home; changes in sleep or appetite, persistent irritability or agitation, or noticeable quietness or withdrawal are considered warning signs; when these symptoms persist and interfere with daily functioning, urgent psychological treatment is recommended; and

(iii) *Adults of working age:* Factors such as the rapid transition to working from home or being fired, financial uncertainty, and separation from family/friends, have led to isolation, insecurity and rumination.

These stressors are considered ecological (macro-politics, external-level institutions, micro-routines) and interact with individual vulnerabilities, reinforcing the need for multi-level supports (Chugh, 2020).

3.4.3 Resilience-oriented Self-care During Prolonged Disruption

Beyond clinical pathways, low-cost practices can buffer distress and support adherence to health protective behaviors (Chugh, 2020):

(i) Establishing a simple daily routine to restore predictability and perceived control;

(ii) Searching accurate, concise information about the illness to reduce uncertainty;

(iii) Aiming to have regular physical activity (e.g., 30–40 minutes on most days);

(iv) Maintaining sleeping hygiene and balancing nutrition to support immunity and mood;

(v) Minimizing or cutting alcohol, nicotine, and excess caffeine that can lead to rise in anxiety and sleep;

(vi) Setting boundaries on pandemic news exposure to prevent continual distressing content;

(vii) Practicing brief relaxation (breathing, mindfulness, or yoga) to down-regulate arousal;

(viii) Maintaining social interpersonal connections via phone/video with loved ones and acquaintances;

(ix) Leveraging one's own experiences that provide purpose as an emotional framework, for instance values or faith; and

(x) Organizing deliberate rest and preventing depletion.

Reframing the disruption (for instance, as a temporary pause that enables reflection and prioritization) can further aid coping.

3.5 Mental Health, Nutrition, and Elite Sport

The interaction between mental health and nutritional status has become a major focus in sports psychiatry and exercise science. As Halioua et al. (2025) noted elite sport imposes exceptional physical and psychological demands on athletes. While performance optimization has long emphasized physical training and nutrition, it is now clear that dietary practices directly influence both physical and mental well-being.

Athletes' eating behaviours span a continuum from based on the performance choices to disordered eating and clinical eating disorders, which remain among the most prevalent psychiatric conditions in competitive sport (Sundgot-Borgen & Torstveit, 2004). The prevalence is particularly high in weight-sensitive disciplines where a lean physique or low body mass is perceived as advantageous (Byrne & McLean, 2002).

The concept of the *Female Athlete Triad* originally linked energy availability, menstrual function, and bone health in female athletes (Nattiv et al., 2007). This framework evolved into the *Relative Energy Deficiency in Sport (REDs)* model, introduced by the International Olympic Committee (IOC) to capture the broader

physiological and psychological effects of low energy availability (LEA) across genders (Mountjoy et al., 2018, 2014, 2023). REDs is defined as a syndrome of impaired physical and/or psychological functioning caused by insufficient energy intake relative to expenditure.

Recent IOC statements highlight that mental health problems can happen prior to, simultaneously with, or result from REDs (Pensgaard et al., 2023). Psychological consequences include disordered eating, mood and anxiety disorders, sleep disturbances, and exercise addiction. Risk factors may be *intentional* such as deliberate pursuit of low weight—or *unintentional*, for instance through inadequate nutritional knowledge, time constraints, or increased training loads without sufficient caloric compensation.

Despite growing recognition, the intersection of nutrition and mental health in elite sport remains methodologically underexplored. Many studies rely on self-reported energy intake and psychological measures, introducing bias and limiting causal interpretation (Halioua et al., 2024). As Halioua et al. (2025) emphasize, advancing this field requires longitudinal designs and close interdisciplinary collaboration between sports psychiatrists and sports nutritionists.

Clinically, integrated treatment pathways established in general psychiatry and eating disorder care are still emerging in elite sport contexts. Evidence-based, multidisciplinary models combining psychiatric, psychological, and nutritional expertise are needed to guide diagnosis, treatment, and safe choices about returning to play (Schulz et al., 2025). Early educational interventions such as the FUEL program for female endurance athletes show promise in improving awareness and promoting healthier habits (Solstad et al., 2025). Together, these developments underline that maintaining athletes' mental health requires coordinated attention to both psychological and nutritional domains.

Beyond elite contexts, physical activity more broadly plays a crucial role in mental health promotion and recovery. Eather et al. (2023) argue that participation in sport and regular physical activity enhances emotional well-being, cognitive function, and resilience while reducing the risk of anxiety and depression. Their research shows that both team and individual sports can serve as protective factors against mental illness, with team sports offering additional benefits through social connection and belonging. These findings underscore the therapeutic potential of sport not only for high-performance athletes but also for the general population recovering from mental health challenges.

The mechanisms connect sport participation and mental health reflect a biopsychosocial interaction: physiological benefits (e.g., endorphin release, improved sleep and stress regulation), psychological effects (e.g., mastery, self-esteem, and goal setting), and social dimensions (e.g., inclusion and shared identity) converge to support well-being. These mechanisms are parallel to those observed in elite sport, where structured activity, motivation, and social support also mediate resilience and mental performance (Eather et al., 2023; Halioua et al., 2025).

Integrating sports psychiatry and nutrition not only supports athlete well-being but enhances sustainable performance. Understanding and preventing REDs and hence, its mental health sequelae should be central to athlete management, policy, and education. Moreover, recognizing the broader mental health benefits of sport participation highlights the need for inclusive, community-based interventions that

leverage physical activity as both prevention and treatment in mental health care.

3.6 From Determinants to Mental Health Statuses

Consistent with a WHO style continuum, mental health spans from flourishing to distress to diagnosable disorder. Movement along this continuum is dynamic and bidirectional with physical health: mental disorders elevate risks for communicable and non-communicable diseases and injuries, while these conditions, in turn, increase the likelihood and persistence of mental disorders. Because comorbidity affects detection, adherence, and outcomes, integrating mental health assessment and based on evidence care into general health services is essential (Prince et al., 2007).

No health without mental health

Beyond their direct contribution to global disability, common mental disorders shape and are shaped by other health conditions. Mental illness increases risk for communicable and non-communicable diseases and injuries; conversely, many physical conditions raise the risk of mental disorder. Such comorbidity delays help-seeking, complicates diagnosis, reduces adherence, and worsens prognosis. The implication is programmatic: mental health should be integrated across primary and specialty care (e.g., HIV, TB, maternal/child health, chronic disease management), not siloed as a separate domain (Prince et al., 2007).

An ecological lens helps move from broad determinants to targeted action by:

(i) examining interactions across micro–macro systems;

(ii) considering cumulative risk; and

(iii) tailoring interventions by subgroup and outcome (e.g., anxiety vs. depression).

Evidence indicates this yields more specific, policy relevant recommendations than factor listings alone (Eriksson et al., 2018).

3.7 Global Policy Framing: Transforming Mental Health Systems

A recent appraisal of the WHO World Mental Health Report argues that meaningful progress requires system transformation on three interdependent fronts:

(i) Revaluing mental health as intrinsic to human, social, and economic well-being

(ii) Shaping environments that promote mental well-being and prevent conditions across the life course

(iii) Modernizing services toward community and people centred, and rights respecting care (Freeman, 2022)

The report defines mental health as a state that enables people to cope with daily stresses, notice abilities, learn and work well, and contribute to their communities—more than the mere absence of disorder.

Operationally, this framing separates what the health sector should lead (e.g., evidence-based, community-delivered services; integration with primary care;

recovery-oriented practice) from actions other sectors must own (poverty alleviation, violence prevention, education, housing, justice). Service networks are expected to include both integrated physical–mental health delivery points and specialist mental health services, with a decisive shift away from long-stay psychiatric hospitals and coercive practices toward community supports grounded in human-rights and lived-experience leadership. For countries, the pathway is not a single blueprint but a set of principle in multi-sector engagement, prevention and promotion, community care, and accountability that can be adapted to context (Freeman, 2022).

3.8 Working Conditions, Digitalization, and Social Media

Rapid digitalization has transformed workplace organization, job demands, and social interaction, and this in turn affects mental health through both risks and protective factors. Information and communication technology *ICT-supported* flexible working can, on the one hand, increase autonomy and reduce commuting stress; on the other hand, it can accelerate work, blur the boundaries work and non-work life, and increase expectations of availability, all of which raise the risk of stress and exhaustion (Eurofound and the International Labour Office, 2021; Messenger, 2019; Siegrist & Wahrendorf, 2016). The meta-analytic study underlines the relationship between psychosocial work characteristics and mental health outcomes by linking the imbalance between rewards and the pressure and effort required at work to depression and anxiety. (Siegrist & Li, 2017; Theorell et al., 2015).

Psychosocial exposures are further altered by digital platforms. Algorithmic management and performance monitoring may reduce perceived control and fairness, but the pressure of constant connection and notifications may cause sleep disorders due to increased cognitive tension (Meijerink et al., 2021; Tarafdar et al., 2019; Wood et al., 2019). We can say that stress factors such as technological overload, technological invasion, and technological complexity are associated with lower well-being and increased burnout, and are referred to as "technostress" (La Torre et al., 2019; Tarafdar et al., 2007). Availability of mental health resources, information sharing, and social support can be facilitated through digital tools, well-defined boundaries, and employee participation (Brough et al., 2020; Carolan et al., 2017).

Social media, both inside and outside the workplace, adds an additional layer. Evidence suggests small average negative associations between heavy social media use and mental health in youth, with substantial heterogeneity across individuals and contexts (Orben, 2020; Orben & Przybylski, 2019). Among adults, social media can provide professional networking and peer support but may also foster to upward social comparison, cyberbullying, and distraction, which are linked to depressive and anxiety symptoms (Best et al., 2014; Verduyn et al., 2017). The mental health impact of digitalization is therefore contingent on design features (e.g., autonomy, transparency, boundary control), organizational policies (e.g., right to disconnect), and individual differences (e.g., coping styles), aligning with biopsychosocial accounts of stress and resilience (Antonovsky, 1996; Karasek Jr, 1979).

Beyond appearance-related pressures, social media environments also facilitate harmful interpersonal dynamics such as cyberbullying, which has become one of the most pervasive online risks. A large-scale meta-analysis by L. Chen et al. (2017) synthesized findings from 81 empirical studies comprising nearly 100,000 participants and identified key predictors of both cyberbullying perpetration and victimization.

The study found that *risky ICT use*, *moral disengagement*, *depression*, and *social norms* were the strongest predictors of online aggression, while *risky ICT use* and *traditional bullying victimization* most strongly predicted victimization.

Policy and practice implications include integrating psychosocial risk assessment into digital transformation projects; co-designing ICT use norms (e.g. availability windows, notification hygiene, and recovery periods); ensuring algorithmic accountability and worker voice; and offering targeted digital literacy and mental health supports. Such measures can shift digitalization from a source of chronic strain to a resource that enhances well-being and performance (OECD, 2020; WHO & ILO, 2022).

These design and policy levers reflect a broader public health lesson: embedding mental health into mainstream prevention and chronic-disease programs yields greater population benefit than siloed services (Prince et al., 2007).

Consistent with ecological theory, platform/organizational rules (exo/macro) can amplify or buffer individual stress processes (micro), so policy levers should be evaluated as cross-level interventions (Eriksson et al., 2018).

During COVID-19, these dynamics were most visible among youth on mainstream platforms.

3.8.1 Social Media and Youth Mental Health During COVID-19 (Survey Findings)

A national survey of Australian youth (16–25; n=371; June–Oct 2020) observed high distress and near-universal platform use (Bailey et al., 2022). Key patterns were:

(i) Psychological distress was substantial (over two in five scored in the severe or extremely severe range on depression or anxiety, DASS-21[1]);

(ii) 96% used social media daily and about two-thirds increased use during lockdowns;

(iii) heavier daily use (>7 hours) correlated with higher depression, anxiety, and stress scores;

(iv) platforms were used for crisis support: about one-third sought help for their own suicidal thoughts/self-harm and about one-half supported someone else;

(v) support commonly occurred on YouTube, Instagram, and Facebook; confidence to give/receive help was modest, and some helpers felt worse afterwards; and

(vi) risk was concentrated among gender-diverse youth and those with prior mental-health diagnoses.

These data indicate that social platforms are both exposure channels and reachable support points for at-risk youth during widespread disruption.

[1]The Depression Anxiety Stress Scales-21 (DASS-21) is a 21-item self-report instrument that measures psychological distress, focusing on depression, anxiety, and stress in adults and adolescents aged 17+.

Implications for Platforms and Practice:

(i) Targeted outreach to higher-risk groups;

(ii) Concise, crisis-safe resources for help-seekers and help-givers (what to say, when to escalate, after-care);

(iii) Safety nudges and crisis links embedded in the apps where support happens;

(iv) Privacy-respecting ways to treat very heavy daily use as a possible distress signal; and

(v) Guidance to reduce helper burden (Bailey et al., 2022).

Taken together, these findings underscore that social media platforms are not only venues for information exchange but also environments where psychological distress, social support, and identity formation unfold in real time. Beyond mainstream networks, similar psychosocial mechanisms operate in other digital contexts such as dating applications and online subcultures. These platforms, while designed for connection, can intensify feelings of exclusion, rejection, or inadequacy among certain user groups. Understanding these dynamics is particularly relevant in examining the intersection of digital romance, incel identity[2], and mental health.

3.8.2 The Intersection of Digital Romance, Incel Subculture, and Mental Health

Quantitative evidence links adverse dating-app experiences to poorer mental and relational health among self-identified incels versus other men. In an online study (incels $n = 38$; non-incel men $n = 107$; $M_{\text{age}} \approx 23.6$), incels showed consistently worse profiles: higher depressive symptoms (CESD-R[3]: 58.6 vs. 40.7), greater rejection sensitivity (15.1 vs. 10.4), stronger fear of being single (25.7 vs. 16.8), higher dating anxiety (67.2 vs. 49.4), lower global self-esteem (1.94 vs. 4.38), lower secure attachment (6.71 vs. 14.43), higher anxious attachment (16.6 vs. 12.5), and more relationship-contingent self-esteem (43.9 vs. 40.5). Perceived popularity on apps was markedly lower (1.21 vs. 2.67). Despite more liberal strategies (wider age/radius; "swipe right" 66.9% vs. 46.4%), incels reported far fewer matches (4.5% vs. 31.3%), more non-responses to first messages (75% vs. 38%), and fewer in-person outcomes (ever dated: 33% vs. 61.7%; relationship: 0% vs. 29.2%; sex: 12.8% vs. 58.1%). In multivariable models, *lower self-esteem* and *lower secure attachment* uniquely predicted incel status, net of other factors (Sparks et al., 2024).

Implications for Mental Health Statuses and Care:

- *Status shifts:* Repeated app rejection happens with elevated depression/anxiety and fragile self/relational schemas (relationship-contingent self-esteem, insecure attachment).
- *Assessment:* Screen for rejection sensitivity, fear of being single, dating anxiety,

[2]The term *incel* refers to individuals, typically young men, who describe themselves as involuntarily celibate and unable to form sexual or romantic relationships. It is often associated with online communities that express resentment toward women and sexually active men.

[3]The CESD-R (Center for Epidemiologic Studies Depression Scale) is a 20 items self-report questionnaire that assesses depressive symptoms and is especially useful for tracking symptom changes over time.

global/contingent self-esteem, and attachment security when app-based rejection is salient.

- *Targets:* CBT[4] for rejection/threat appraisals and graded social-approach work alongside attachment-informed interventions to build security; coach prosocial, effective online approaches.
- *Context:* Contextualize outcomes within platform ecology (competition, choice overload), avoiding pathologizing labels while addressing safety and stigma.

3.8.3 Cyberbullying, Health, and Protective Family Factors

Bullying and cyberbullying are intentional, goal-oriented aggressive acts intended to harm another person and therefore belong to violent social conduct. They are best understood within social psychology: Aggression is shaped not only by individual traits but also by situational forces such as authority/power, group norms and identity, and the human need to belong. When belonging or self-esteem is threatened—through rejection or exclusion or aggressive responses become more likely; online settings can amplify these dynamics by increasing reach and persistence. This framing complements operational definitions by locating cyberbullying within broader processes of power, belonging, and context (Navarro et al., 2016).

Social-Psychology Mechanisms

Cyberbullying is shaped by group and contextual processes (Navarro et al., 2016):

(i) Group norms and bystander dynamics (status, identity, contagion) steer participation;

(ii) Links with aggressive/violent conduct; perceived power/authority lowers inhibition;

(iii) Reduced affective and/or cognitive empathy is associated with perpetration;

(iv) Moral disengagement (anonymity, distance, audience) enables harm without guilt;

(v) Moral judgments and value climates (perceived injustice, weak rule internalization) increase risk;

(vi) Family context matters: Parental warmth/communication and active mediation reduce involvement, whereas conflictual ties and low guidance increase it; and

(vii) Gender patterns are mixed; simple male–female assumptions are not supported consistently.

Common Forms

Operationally, cyberbullying manifests through recurring online behaviors, including (Hinduja & Patchin, 2008; Willard, 2006):

(i) Insults, humiliation, ridicule, and denigration;

[4]Cognitive Behavioral Therapy (CBT) is a structured, goal-oriented psychotherapy that focuses on changing unhelpful thoughts and behaviors to improve emotional well-being and manage issues like anxiety and depression.

(ii) Defamation/reputation damage, spreading embarrassing content;

(iii) Exclusion (being removed from group chats/games; being ignored);

(iv) Impersonation and harm via fake accounts;

(v) Threats, harassment, and persistent tracking (cyberstalking); and

(vi) Non-consensual disclosure of personal information (doxxing) or images.

These features can exacerbate mental health impacts especially depression, anxiety, and PTSD and disrupt school and family functioning (Kowalski & Limber, 2013).

Effects of cyberbullying

Evidence across countries shows that cyberbullying threatens adolescent health and well-being and can, in some respects, be more harmful and longer-lasting than traditional bullying because content persists, spreads rapidly, and often reaches a large unseen audience (Navarro et al., 2016, pp. 8–10). Reported impacts include:

(i) *Physical:* Headaches, stomachaches, sleep disturbance, fatigue, backache, loss of appetite, and other somatic complaints;

(ii) *Psychological/emotional*: Fear (even terror), anxiety, sadness, anger, depressive symptoms, and increased suicidal ideation (with some studies noting higher risk than in traditional bullying);

(iii) *School-related:* Decreased in motivation and concentration, absenteeism, and academic performance problems; and

(iv) *Psychosocial:* Isolation and loneliness, ostracism/exclusion, and threatened belonging or self-esteem.

Severity varies with situational factors such as anonymity, the size/visibility of the audience (bystanders), and whether the victim knows the perpetrator, underscoring the need for multi-level prevention and response (Navarro et al., 2016).

Health links and Protective Family Factors

Evidence from a large high school sample in China links cyberbullying victimization to broad health and damages in mental health, between parent and child connection moderating some effects (Zhu et al., 2021). In a stratified sample (n=3,232; ages 15–17) from Xi'an, lifetime victimization was reported by 22.2% and past-year victimization by 6.3%. The following model shows:

(i) poorer overall health (SF-12[5]) among victims (lifetime and past year);

(ii) higher depressive symptoms (BDI-II[6]) and PTSD symptoms (UCLA PTSD-RI[7]) for both lifetime and past year victims;

[5]The SF-12 is a short self-report instrument that measures the effect of health on everyday functioning and is commonly used to assess quality of life.

[6]The Beck Depression Inventory-II (BDI-II) is a 21-item questionnaire designed to assess the severity of depression and is commonly used in clinical practice to monitor symptoms and treatment progress.

[7]The UCLA Child/Adolescent PTSD Reaction Index for DSM-5 (PTSD-RI) is a revised self-report measure that evaluates post traumatic stress symptoms in school aged children and adolescents.

(iii) higher odds of problem drinking, cigarette smoking, and gambling among victims (odds ratios roughly 1.3–2.2, depending on outcome and timing); and

(iv) significant buffering by stronger parent and child attachment for depressive and PTSD symptoms (but not for general health or addictive behaviors).

These findings reinforce that online relational aggression has offline consequences across multiple domains and that supportive family bonds can attenuate internalizing symptoms after victimization.

Implications for Prevention and Response

The evidence summarized above makes clear that cyberbullying is not merely a social or disciplinary issue but a significant public-health concern that intersects with family systems, education, and digital governance. Because harms emerge from the interaction of individual vulnerabilities, peer dynamics, and technological affordances, effective prevention must operate across multiple levels. In particular, interventions that combine emotional skills training, parental engagement, and based in school policy enforcement show the strongest promise for sustainable impact. At the same time, digital platforms and AI-driven moderation tools play an increasingly central role in detection, reporting, and early intervention, highlighting the need to align technical innovation with psychosocial understanding.

(i) Embed parent components (monitoring, open communication, supportive problem-solving) into anti-cyberbullying programs;

(ii) Equip schools with policies and education that target online relational harms and promote supportive peer climates;

(iii) Integrate screening for cyberbullying in adolescent mental health assessments (depression/PTSD, sleep, somatic complaints, and substance use);

(iv) Coordinate with platforms on reporting tools and rapid guidance for victims and families;

(v) School actions: Clear policies, teacher engagement, bystander training, and restorative responses (Navarro et al., 2016);

(vi) Family actions: Strengthen parent–child relations and active mediation/monitoring of online activity (Navarro et al., 2016); and

(vii) Community/platform actions: Media literacy, easy reporting, and norm-setting campaigns (Navarro et al., 2016).

In ecological terms, family processes (micro) can buffer damages originating in digital environments (exo), suggesting cross-level intervention points. These same cross-level, dynamics—linking individual experience, social context, and technological mediation are increasingly relevant as artificial intelligence tools begin to shape how mental health risks are detected, assessed, and managed online. Therefore, the next section evaluates emerging evidence and ethical considerations surrounding AI-based mental health tools.

3.9 Evaluating AI Mental Health Tools

Tools that use generative AI and large language models for mental health are appearing at pace, however, their quality, safety, and fairness are uneven. To make evaluations comparable, Golden and Aboujaoude (2024) introduce FAITA–Mental Health, a concise rubric (0–24 points) tailored to AI-enabled interventions. The FAITA (Framework for the Assessment of Intelligent Technologies in Mental Health) rubric provides a standardized, based on evidence approach to asses the safety, efficacy, and ethical quality of AI-based mental health tools (Golden & Aboujaoude, 2024).

It scores six areas:

(i) *Credibility* — is the aim clear, grounded in evidence, and does the tool keep users engaged? (0–6);

(ii) *User experience* — does it adapt to the individual, feel natural to interact with, and offer simple ways to give feedback or get help? (0–6);

(iii) *User agency* — who controls personal data, and does the design build self-management rather than dependence? (0–4);

(iv) *Equity and inclusivity* — is the content culturally responsive and actively checked for bias and fairness? (0–4);

(v) *Transparency* — are ownership, funding, development process, and governance explained in plain language? (0–2); and

(vi) *Crisis management* — can the tool recognize emergencies and connect users to appropriate local resources with documented follow-through? (0–2).

In practice, FAITA can serve as a purchasing and research checklist alongside model update logs, tests for hallucinations and safety guardrails, and privacy/security reviews. As the authors note, the framework still needs formal validation (e.g., inter-rater reliability and links to clinical or safety outcomes), so it should complement rather than replacing risk management, bias audits, and human oversight (Golden & Aboujaoude, 2024).

Given that young people tend to seek and provide support on mainstream platforms, evaluation rubrics for AI tools should incorporate crisis-safe communication standards (e.g., based in assistance prompts, helper after care) alongside usability and equity checks (Bailey et al., 2022; Golden & Aboujaoude, 2024).

Together, these frameworks and evaluation tools demonstrate how mental health support is increasingly shaped by digital infrastructures and artificial intelligence. However, understanding and evaluating technological interventions ultimately requires grounding in the core conditions they are intended to address. Therefore, the following subsection provides a concise overview of common mental health disorders which defines their features, lived experiences, and implications for care in order to contextualize how AI and digital systems interact with real-world mental health needs.

3.10 Overview of Common Mental Health Conditions

Common mental health conditions span mood, anxiety, neuro-developmental, stress, and personality disorders. The following is a concise overview for a clarification; detailed diagnostic criteria are provided in the *DSM-5-TR*[8] and *ICD-11* (American Psychiatric Association, 2022; *International Classification of Diseases 11th Revision (ICD-11)*, 2019).

Depression

Depression is characterized by a persistent low mood and/or anhedonia[9], accompanied by cognitive, somatic, and functional impairment. Recurrent episodes are common (American Psychiatric Association, 2022). Globally, depressive disorders rank among the leading contributors to the years of living with disability (GBD 2019 Diseases and Injuries Collaborators, 2020).

Individuals experiencing depression often attribute it to a combination of stressful life events and biological factors. While both psychotherapy and medication are viewed as helpful, psychological care is generally preferred due to concerns about medication side effects or dependency. Despite a high perceived need for care (estimated at 50–80%), help-seeking is frequently delayed by stigma, financial barriers, and doubts about treatment efficacy. Vulnerable populations such as minorities, youth, and older adults face heightened risks of unmet needs. Exploring beliefs across identity, cause, timeline, consequences, and control, and applying shared decision-making, can enhance treatment engagement and adherence (Prins et al., 2008).

Depression predicts disability progression and higher all-cause mortality, increases risk of coronary heart disease and stroke, and commonly follows myocardial infarction and stroke where it worsens outcomes. Treatments reliably improve depressive symptoms, though their impact on long-term cardiovascular health remains mixed (Prince et al., 2007).

Pharmacological Treatment Antidepressant medications, particularly selective serotonin reuptake inhibitors (SSRIs), are widely used as first-line pharmacological treatments for depressive disorders. At a population level, an ecological analysis of 29 European countries found that increases in antidepressant utilization (defined daily doses per 1,000 inhabitants per day) were associated with reductions in standardized suicide mortality rates, even after adjustment for economic indicators, alcohol consumption, unemployment, and divorce rates (Gusmão et al., 2013). The authors caution that these data do not establish causality, but they support the view that appropriate antidepressant use, embedded within broader primary-care and public mental health strategies, can contribute to suicide prevention rather than increasing the risk of suicide. Beyond clinical outcomes, how depression is perceived socially strongly influences recovery, help-seeking, and reintegration.

Public attitudes and Stigma Although depression is now discussed more openly, everyday social attitudes have shifted only gradually. Long-term survey data show limited progress: while people express slightly greater understanding, feelings of

[8] The DSM-5-TR represents the latest update of the Diagnostic and Statistical Manual of Mental Disorders and provides uniform guidelines for diagnosing and categorizing mental health disorders.

[9] Anhedonia is the diminished capacity to experience pleasure or interest in previously enjoyable activities, often described as emotional numbness.

discomfort and a tendency to keep distance remain widespread. This indicates that simply providing information is not enough to change behaviour. More enduring improvements depend on opportunities for direct contact with those who have lived experience, supportive social policies, and access to effective clinical care (Angermeyer & Matschinger, 2004). Targeted web-based modules for staff and students have also reduced social distance and personal/perceived stigma while improving knowledge, offering a scalable complement to contact-based approaches (Finkelstein & Lapshin, 2007). Although depression generally elicits less desired social distance than schizophrenia or substance-use disorders, meaningful distance remains, reinforcing the need for sustained anti-stigma work alongside clinical care (Jorm & Oh, 2009). Compared with schizophrenia, depression more often elicits pro-social responses (pity, desire to help) and less perceived dangerousness/unpredictability. In this survey, labeling the vignette as a "mental illness" tended to reduce anger for depression (but increased fear for schizophrenia), and psychosocial-stress explanations aligned with more supportive reactions (Angermeyer & Matschinger, 2003).

Depression vs. Schizophrenia A large population survey found common ground—both conditions were widely recognized as mental illnesses; acute stress was frequently endorsed as a cause; a poor "natural course" but favourable treatment prognosis was expected. Key differences mattered for stigma: people with schizophrenia were far more often viewed as dangerous and unpredictable and evoked more fear, whereas people with depression evoked more pity and helping intentions. Labeling as "mental illness" affected emotions in opposite ways (increasing fear for schizophrenia, reducing anger for depression). Emphasizing psychosocial stress was linked to more supportive reactions, while "brain disease" explanations increased fear chiefly for schizophrenia. These patterns argue for tailored anti-stigma work: reduce blame for both; address violence stereotypes explicitly for schizophrenia; and use contact- plus education-based strategies for depression (Angermeyer & Matschinger, 2003).

Schizophrenia

A primary psychotic disorder defined by persistent delusions and/or hallucinations with disorganized thinking/behavior, alongside negative symptoms (e.g., avolition, anhedonia, social withdrawal) and frequently measurable cognitive impairment. Onset is typically in late adolescence or early adulthood, with a modest male predominance; social and occupational decline is common but outcomes are heterogeneous rather than uniformly poor. Symptom structure is often described in positive, negative, and disorganization domains.Established biological findings include lateral ventricular enlargement and small (≈2%) whole-brain volume reductions; neurochemical evidence implicates dopamine dysregulation and glutamatergic (NMDA receptor[10]) disturbance. Risk is largely polygenic with contributions from birth and early life factors. First-line treatment remains dopamine D_2–receptor–blocking antipsychotics (with clozapine for treatment-resistant illness), while cognitive-behavioural therapy adds small symptom benefits; psychosocial and recovery-oriented supports are crucial for functioning. Current debates include the extent to which psychosis lies on a continuum, the roles of cannabis and childhood adversity, and whether progressive brain change occurs after onset (Jauhar et al., 2022).

[10]NMDA receptors are glutamate-activated ion channels that support synaptic plasticity, learning, memory, and other brain functions. Dysfunction is associated with neurological and psychiatric conditions such as Alzheimer's disease, epilepsy, and schizophrenia.

Epidemiology Contrary to older dogma, frequency measures vary markedly across place and population. Median incidence is about 15.2 per 100,000 with a fivefold inter-site spread; men are affected more often than women (rate ratio ≈1.4). Lifetime morbid risk clusters around 7.2 per 1,000 (closer to 0.7% than the colloquial "1%"). Migrant groups show higher incidence and prevalence, and incidence tends to be higher in urban versus mixed urban–rural settings. People with schizophrenia have a substantially elevated mortality (all-cause standardized mortality ratio ≈2.6), and this mortality gap has widened in recent decades. Higher latitude is linked to higher prevalence (and higher male incidence), and developed economies report higher prevalence than less-developed settings. These gradients point to multiple social and environmental modifiers and argue for a targeted early intervention in urban and migrant communities and for systematic physical-health management to close the mortality gap (McGrath et al., 2008).

Bipolar Disorders

Defined by recurrent episodes of mania/hypomania and depression with marked functional impairment and elevated suicide risk if untreated; care centres on mood stabilisation (e.g., lithium and other mood stabilisers/atypical antipsychotics), psychoeducation, and relapse-prevention therapies (American Psychiatric Association, 2022; Grande et al., 2016). Beyond symptom control, structured psychosocial treatments improve course and functioning, including group psychoeducation, interpersonal and social rhythm therapy, and focused on family therapy (Colom et al., 2003; Frank et al., 2005; Miklowitz et al., 2003). Recovery is also shaped by resilience factors: in a longitudinal study of adults with bipolar disorder, higher self-management, care, and confidence correlated with better well-being, quality of life, and lower functional impairment; critically, baseline self-confidence predicted later gains in personal recovery, and improvements in self-confidence mediated benefits from interpersonal support and self-care (Echezarraga et al., 2018). Practically, treatment plans should pair pharmacotherapy and relapse-prevention skills (sleep–wake regularity, early-warning monitoring, coping plans) with targeted strengthening of self-confidence, self-management, and self-care to support sustained recovery.

Anxiety Disorders

Include generalized anxiety, panic disorder, social anxiety, and specific phobias; characterized by excessive fear/anxiety and avoidance (American Psychiatric Association, 2022). Evidence-based treatments include CBT and SSRIs/SNRIs (Cristea et al., 2015; *Generalised Anxiety Disorder and Panic Disorder in Adults: Management*, 2011).

Patient perspective People with anxiety disorders typically rate psychological treatments especially skills-based cognitive and exposure-based approaches—as their best option; medication is often seen as adjunctive. About half to four-fifths perceive a need for care, yet many face stigma and practical/cost barriers or attribute symptoms to purely physical causes, which delays help-seeking. Mapping illness beliefs (what it is, why it happens, how long it lasts, effects, and how it can be controlled) and aligning plans accordingly can increase uptake and persistence with evidence-based care (Prins et al., 2008).

Anxiety disorders generally attract less social distance than psychotic or substance use disorders, but avoidance is still common and improves with informed contact (Jorm & Oh, 2009).

Stress Related Disorders (including PTSD)

Stress is the psychobiological response to perceived demands or threats. Briefly, controllable stress can be adaptive; chronic, uncontrollable stress dysregulates autonomic/ HPA systems[11] and contributes to allostatic load (cumulative wear-and-tear) linked to anxiety, depression, sleep and cardiometabolic risk (McEwen, 2004). PTSD is a trauma-related disorder that follows exposure to a qualifying event (actual or threatened death, serious injury, or sexual violence) via direct experience, witnessing, learning about it occurring to a close other, or repeated/extreme exposure to aversive details. In order to meet diagnostic criteria, symptoms must occur in four clusters: intrusion, avoidance, changes in mood or cognition, and increased arousal or reactivity. They must persist for more than one month and cause significant distress or impairment. Dissociation or delayed-onset may also apply. Acute stress disorder shares a similar symptom picture but spans 3 days to 1 month; adjustment disorder applies to non–Criterion-A[12] stressors with disproportionate distress/impairment (American Psychiatric Association, 2022). Effective care prioritises early identification and trauma-focused psychotherapies (e.g., prolonged exposure[13], cognitive processing therapy[14], TF-CBT[15], EMDR[16], written exposure therapy[17]), with SSRIs/SNRIs[18] as adjuncts where indicated (Bisson et al., 2018); Guidelines for the Management of Conditions Specifically Related to Stress, 2018). Delivery via secure telehealth is generally non-inferior to in-person for PE and CPT, with promising support for WET and TF-CBT and practical guidance on safety and alliance (Bucur et al., 2024). Beyond discrete traumas, chronic stressors in work and daily life matter: high job strain and effort–reward imbalance are associated with elevated depressive and anxiety symptoms (Theorell et al., 2015).

The Social Ecology of PTSD PTSD risk and recovery are strongly shaped by social context (Charuvastra & Cloitre, 2008):

(i) Interpersonal (human-intent) traumas carry higher conditional PTSD risk than accidents or natural disasters;

(ii) Perceived, functional social support before and after trauma is among the strongest correlates of lower PTSD risk/severity; network size alone is less predictive;

[11] The hypothalamic–pituitary–adrenal (HPA) axis functions as a central stress response pathway, controlling cortisol secretion and supporting the body's physiological stability.

[12] Non–Criterion-A stressors are significant life events that do not meet the DSM definition of trauma (e.g., divorce, job loss) but can still cause marked distress or impairment.

[13] Prolonged Exposure (PE) therapy is a trauma-focused treatment that reduces PTSD symptoms by safely revisiting and processing traumatic memories.

[14] Cognitive Processing Therapy (CPT) is a structured form of cognitive-behavioral intervention designed to help individuals reinterpret and modify harmful trauma-related beliefs.

[15] Trauma-Focused Cognitive Behavioral Therapy (TF-CBT) is a structured psychotherapy for children and adolescents that integrates cognitive and behavioral techniques to address trauma symptoms.

[16] Eye Movement Desensitization and Reprocessing (EMDR) is a therapy that uses guided eye movements to help individuals process and reduce distress from traumatic memories.

[17] Written Exposure Therapy (WET) is a brief trauma-focused intervention where individuals process distressing experiences through structured writing sessions.

[18] Selective Serotonin Reuptake Inhibitors (SSRIs) and Serotonin–Norepinephrine Reuptake Inhibitors (SNRIs) are classes of antidepressant medications commonly prescribed to reduce symptoms of depression and anxiety.

(iii) Negative social reactions (blame, disbelief, conflict) robustly predict worse and more persistent PTSD; positive support helps most when matched to needs and from trusted sources;

(iv) In youth, caregiver proximity and functioning are pivotal; parental PTSD/avoidance predicts child symptoms, and early maltreatment elevates later PTSD risk;

(v) Social support reduces avoidant coping; low support links to withdrawal that maintains symptoms;

(vi) supportive bonds can downshift threat responses via social-neurobiological pathways (e.g., affiliation systems).

Conceptual Refinements Social context effects are not only protective but also bidirectional: supportive environments can reduce symptoms, while symptoms can erode support over time. Beyond perceived support, network diversity and neighborhood cohesion matter; negative reactions to disclosure are uniquely harmful; and before and during trauma contexts shape later risk. Conservation of resources theory offers a useful frame for resource loss cascades and intervention targets (Vogt et al., 2017).

Design and measurement notes Use longitudinal/cross-lag designs; include pre/during/post-trauma social measures; complement self-report with observed interactions and social-network mapping; test moderators (e.g., gender, community context) to tailor interventions (Vogt et al., 2017).

Cumulative Trauma and Coping Evidence in community samples shows dose–response relations and coping links:

(i) cumulative interpersonal traumas (e.g., child abuse plus adult rape) markedly increase odds of probable PTSD, while intact social support can buffer symptom severity in dual-trauma cases (Schumm et al., 2006);

(ii) approach-oriented coping is associated with better functioning, and early changes in coping track subsequent PTSD severity (Gutner et al., 2006; Tiet et al., 2006).

Assessment and Intervention Implications

(i) Documenting trauma type (human-intent vs. accidental) and mapping helpful vs. harmful responses across the network;

(ii) Addressing negative reactions directly (psychoeducation, bystander coaching) and arranging matched, practical support;

(iii) Prioritizing therapeutic alliance and early skills (emotion regulation, interpersonal effectiveness) before/alongside exposure for complex/interpersonal trauma;

(iv) Involving caregivers for youth to restore safety and reduce avoidance; for adults, mobilizing peer/community ties;

(v) Including graded social-approach homework to counter avoidance and strengthen perceived support.

Telehealth Delivery (Evidence and Practice)

(i) Delivering PE and CPT via video conference as non-inferior to in-person; WET and TF-CBT showing promising telehealth outcomes (Bruce et al., 2025);

(ii) Setting safety protocols (patient location each session, local emergency contacts) and arranging a technical fallback (phone) before starting;

(iii) Using measurement-based care and ensuring clear audio/video to support alliance; conducting brief check-ins to maintain engagement between sessions;

(iv) Providing structured instrumental support (e.g., a trained peer accompanying via secure video) when exposure homework is difficult to complete alone, to improve completion and retention (Bruce et al., 2025).

In today's digital spaces, repeated rejection on dating apps which can push people toward more depression, rejection sensitivity, and dating anxiety while wearing down self-esteem and attachment security, sharpening where assessment and prevention should focus online.

Taken together, a biopsychosocial and ecological lens—pairing evidence-based trauma therapies with optimisation of social supports and flexible delivery (including telehealth)—provides the organising principle for this section and links directly to the next parts on determinants, interventions, and evaluation.

Borderline Personality Disorder (BPD)

Borderline personality disorder is a severe condition marked by persistent instability in emotion, self-image, and close relationships, together with impulsivity and recurrent self-harm or suicidality. In the general population it affects around 1–2% (lifetime up to 5.9%), and comorbidity with mood and anxiety disorders, PTSD, and substance use is the rule rather than the exception (American Psychiatric Association, 2022; Grant et al., 2008). Care is primarily psychological and structured: dialectical behavior therapy shows the strongest evidence for reducing self-harm and overall severity, with additional benefits reported for mentalization-based, schema-focused, and transference-focused treatments (Linehan, 2015; Størbø et al., 2020). From a public-health perspective, BPD remains poorly recognized and lay people rarely label it correctly, often confuse it with depression or schizophrenia, and are less inclined to recommend professional help-underscoring the need for targeted literacy initiatives and clear referral pathways (Furnham et al., 2015).

Prospective Suicide Risk in BPD: In the 10-year Collaborative Longitudinal Study of Personality Disorders (CLPS; $N = 701$), BPD (excluding the self-injurious behavior criterion) was the strongest baseline factor associated with prospectively observed suicide attempts (OR=4.18, 95% CI: 2.68–6.52 [19]) after adjusting for demographics and comorbidities. When BPD criteria were modeled simultaneously, *identity disturbance* (OR=2.21, 95% CI: 1.37–3.56), *chronic feelings of emptiness* (OR=1.63, 95% CI: 1.03–2.57), and *frantic efforts to avoid abandonment* (OR=1.93, 95% CI: 1.17–3.16) remained independent predictors over follow-up—features that may be overlooked in risk assessments (Yen et al., 2021). Childhood sexual abuse was

[19] OR stands for Odds Ratio, a measure of the strength of association between a predictor and an outcome. A 95% Confidence Interval (CI) indicates the range within which the true value is expected to fall with 95% certainty.

also independently associated with attempts (OR=2.12, 95% CI: 1.33–3.36), whereas obsessive–compulsive personality disorder was linked to lower odds (OR=0.47, 95% CI: 0.30–0.73) (Yen et al., 2021).

A BPD diagnosis is most valuable when it guides access to effective, compassionate treatment rather than serving as a reductive label. Used well, it provides a shared language for collaborative formulation, helps patients and clinicians source reliable information, and supports research into what works; used poorly, it can invite pessimism or obscure the person's individuality (Bateman & Krawitz, 2013). Contemporary classification emphasizes impairments in self and interpersonal functioning with graded severity, alongside trait qualifiers; within such dimensional systems, a borderline pattern qualifier/type remains available to capture the characteristic instability in affect, self, relationships, and control of impulses (see also American Psychiatric Association, 2022). For practical diagnosis, five or more of the familiar features should be present across contexts (e.g., frantic efforts to avoid abandonment, unstable and intense relationships, identity disturbance, impulsivity in risky domains, recurrent self-harm/suicidality, affective lability, chronic emptiness, intense anger, and stress-related paranoia/dissociation) (American Psychiatric Association, 2022). Regardless of framework, clinicians are encouraged to disclose and discuss the diagnosis collaboratively and to link it to structured, evidence based care (Bateman & Krawitz, 2013; Linehan, 2015; Størbø et al., 2020).

Suicidal Thoughts and Behaviors; Self-Harm

Suicide risk intersects with mood, psychotic, and personality disorders; means restriction, brief contact interventions, and evidence-based therapies reduce risk (*Preventing Suicide: A Global Imperative*, 2014; Zalsman et al., 2016). Non-suicidal self-injury (NSSI) often reflects emotion regulation difficulties and requires tailored assessment (Kiekens et al., 2020). Beyond clinical factors, short-run environmental stressors also matter: county–month panel evidence (1960–2016) shows suicide rates vary systematically with ambient temperature across nonlinear bins after accounting for location-by-month and year effects, indicating sensitivity to weather shocks (Mullins & White, 2019). The authors further examine policy and access modifiers including strong mental health parity laws, mental health professional shortage designations, substance-abuse treatment capacity, and county income to assess whether supportive insurance, greater provider availability, and treatment infrastructure attenuate temperature-related suicide risks (Mullins & White, 2019).

Within BPD, risk formulation should not focus solely on affective instability and impulsivity: identity disturbance, chronic emptiness, and abandonment fears independently predict suicide attempts over long-term follow-up, alongside elevations with PTSD, alcohol/substance use disorders, and childhood sexual abuse (Yen et al., 2021). Clinically, these domains warrant routine assessment and targeted intervention within structured therapies (e.g., DBT [20] with modules addressing identity/interpersonal instability, MBT [21], schema-focused approaches).

[20] Dialectical Behavior Therapy (DBT) is an evidence-based psychotherapy that combines acceptance and change strategies to reduce self-harm, improve emotion regulation, and strengthen interpersonal functioning.

[21] Mentalization-Based Therapy (MBT) is a psychotherapy that helps individuals improve their ability to understand and interpret their own and others' thoughts, feelings, and intentions, supporting emotional regulation and relationship stability.

Attention-Deficit/Hyperactivity Disorder (ADHD)

A neurodevelopmental disorder marked by a persistent pattern of inattention and/or hyperactivity–impulsivity that is inconsistent with developmental level, begins in childhood, and causes clinically significant impairment across settings; symptoms frequently persist into adulthood and show high heritability (American Psychiatric Association, 2022; Faraone et al., 2021). ADHD is associated with educational, occupational, and social difficulties and elevated risk of comorbid conditions (e.g., oppositionality, anxiety, depression) and injury; effective care is typically multimodal (psychoeducation, behavioral/parent training, school accommodations, cognitive–behavioral strategies, and pharmacotherapy with stimulants or evidence-based nonstimulants) (American Psychiatric Association, 2022; Faraone et al., 2021; Pliszka, 2007).

Symptoms typically emerge in the early school years, with substantial persistence through adolescence and into adulthood for many (American Psychiatric Association, 2022; Faraone et al., 2021). In caregivers, younger maternal age among mothers with ADHD tendencies has been linked to poorer mental health, underscoring the value of early identification and tailored support for younger mothers (Marumoto et al., 2025). A state of well-being in which individuals realize their abilities, can cope with normal stresses, work productively, and contribute to their community; it reflects not merely the absence of mental disorder but the presence of psychological functioning and resilience shaped by biological, psychological, and social factors (e.g., stressors, supports) (World Health Organization, 2022).

Living with ADHD and Mental Health Functioning improves with structured supports across the lifespan:

(i) Predictable routines, organizational aids, and environmental scaffolding;

(ii) Behavioral/parent-management training and school accommodations for children;

(iii) CBT/executive-skills coaching for adolescents and adults;

(iv) Evidence-based pharmacotherapy when indicated;

(v) Sleep, physical activity, and substance-use risk reduction;

(vi) Screening and support for caregiver mental health and parenting stress (e.g., difficulties understanding the child's needs, inadequate understanding from others), with priority attention to younger mothers with ADHD tendencies.

(American Psychiatric Association, 2022; Marumoto et al., 2025; Pliszka, 2007; World Health Organization, 2022).

Personality Disorders

Enduring patterns of inner experience and behavior deviating from cultural expectations, inflexible and pervasive, with distress or impairment; dimensional models in ICD-11 emphasize severity and trait domains (*International Classification of Diseases 11th Revision (ICD-11)*, 2019). Structured psychotherapies have the strongest evidence base (Livesley et al., 2018).

Across conditions, comorbidity is common, social determinants shape onset and

course, and access to timely, evidence-based care is critical. Prevention and intervention strategies should integrate clinical, psychological, and social components, consistent with the biopsychosocial model (Antonovsky, 1996; Engel, 1977; Marmot, 2005).

3.10.1 Gender Bias in Mental Health

By gender bias we mean a systematic distortion in assessment, diagnosis, risk estimation, or treatment that is attributable to the patient's gender rather than to clinically relevant need or evidence. Bias can arise from stereotypes about symptom expression, from instruments normed on one gender, from clinical heuristics used under uncertainty, and from structural factors that differentially shape access and labeling (Eriksen & Kress, 2008; Ruiz & Verbrugge, 1997). In practice, even when formal criteria exist, judgments may still be swayed by gendered social norms—as shown in a randomized trial where identical paraphilic vignettes[22] were pathologized less often when the subject was female (with the reverse for masochism when full criteria were met) (Fuss et al., 2018).

Operationally, bias can manifest at multiple points:

(i) Measurement bias (screeners cut-points, items, or examples reflecting one gender's presentation). Across identical vignettes, male targets elicit greater desired social distance than female targets, demonstrating how gendered expectations shape public responses (Jorm & Oh, 2009);

(ii) Diagnostic heuristics (e.g., externalizing assumed "male", internalizing "female");

(iii) Risk and dangerousness appraisals (higher perceived risk assigned to men for the same description);

(iv) Treatment allocation (differences in referral, medication, or psychotherapy offers at equal severity); and

(v) Documentation and legal labels (language and codes that carry stigma or legal consequences).

Bias mitigation should include structured diagnostic checklists for borderline presentations, routine audits of case reviews for gender-related effects, and an explicit distinction between atypical but consensual behavior and diagnosable disorder. These practices are consistent with equity-oriented standards for tool evaluation (Golden & Aboujaoude, 2024).

An example from randomized vignette evidence with mental health professionals (n=546) shows that diagnostic judgments about atypical sexuality can be swayed by patient gender, independent of criteria (Fuss et al., 2018). Key findings:

(i) Female subjects were less often judged "mentally disordered" for exhibitionistic, frotteuristic[23], sexual-sadistic, and pedophilic cases, especially when criteria were ambiguous;

[22]Paraphilic disorders involve atypical sexual interests that may cause distress, impairment, or risk of harm, such as voyeurism, exhibitionism, or fetishistic behaviors.

[23]Frotteuristic disorder involves recurrent urges or behaviors of touching or rubbing against a non-consenting person, typically in crowded public settings, and is considered a paraphilic disorder when it causes distress or risk of harm.

(ii) For sexual masochism, the pattern reversed: when full DSM-5 criteria were met, women were more often pathologized;

(iii) Stigma profiles differed by gender (men perceived as more dangerous, greater social distance); sexual orientation had no effect; and

(iv) Pedophilia was seen as more "biological" than other paraphilias, but all paraphilias were more stigmatized (yet less pathologized) than psychosis.

Implications:

(i) Reinforcing the DSM-5 distinction between paraphilia and paraphilic disorder (distress/harm/consent);

(ii) Using structured checklists and peer review when criteria are borderline;

(iii) Monitoring/auditing gender effects in case conferences; and

(iv) Including bias-awareness training and transparent documentation of decision thresholds.

This chapter has framed mental health as a dynamic, biopsychosocial, and ecological phenomenon, shaped by biological susceptibilities, psychological processes, social structures, and wider environmental conditions. We reviewed how mental health literacy links knowledge to action, how structural levers (coverage, provider capacity, policy) and digitalization influence risk and resilience, and how condition-specific evidence (e.g., BPD's suicide risk profile; ADHD's life course and caregiver context) refines assessment and care. Cross-cutting themes included stigma and social distance, family and community level buffers, and equity concerns such as gender bias. As a result, effective responses require pairing evidence-based clinical care with context-sensitive prevention, system design, and fair measurement.

Mental health is also fundamentally communicated and negotiated through language: symptoms are narrated, support is offered or withheld, stigma is expressed, and policies are codified in text. Increasingly, this language is digital, appearing in clinical notes, peer-support forums, social media posts, help-seeking messages, and policy documents, creating a large-scale, time-stamped record of psychological states and social contexts. Natural Language Processing (NLP) provides the toolkit to transform this unstructured text into actionable signals that complement traditional measures. Applications include tracking population level literacy and stigma, detecting risk and triaging support, evaluating care quality and access, operationalizing biopsychosocial constructs in context, and embedding safety, equity, and governance into AI-enabled tools.

In sum, this chapter emphasized that mental health must be understood as both a clinical and social phenomenon, shaped by structures, language, and lived experience. The next chapter introduces NLP methods that ranging from preprocessing to embeddings and transformer models and showing how linguistic signals can be harnessed to strengthen screening, monitoring, and service improvement.

4 Natural Language Processing (NLP)

Natural language processing (NLP), a branch of artificial intelligence, automatically analyzes digital text to draw conclusions about human language and emotions. These findings can lead to the development of personalized advertising or other types of interventions, among other things. While NLP techniques are widely used in sectors such as marketing to analyze emails and social media posts, they also offer significant potential in the fields of medicine and mental health. Natural language processing (NLP) can support mental health and well-being in these situations by providing automated assessments and treatments based on the analysis of textual communication (Calvo et al., 2017). NLP provides a collection of computational techniques that allow machines to analyze, make sense of, and produce human language in both written and spoken form.

These techniques include (Calvo et al., 2017):

1. *Tokenization*: The process of dividing raw text into smaller units called tokens (words, sub-words, or sentences). It converts unstructured language into a structured format for subsequent analysis and serves as a key preprocessing step for tasks such as parsing and feature extraction. Common approaches include word-level tokenization, sub-word methods (e.g., morphemes, character n-grams), and sentence segmentation.

2. *Sentiment Analysis:* Examining text to identify the feelings or attitude conveyed, helping to assess the writer's mood or perspective.

3. *Named Entity Recognition (NER):* The identification and classification of key elements within the text, such as names of people, organizations, or locations.

4. *Part-of-Speech Tagging:* The assignment of parts of speech (e.g., nouns, verbs, adjectives) to each word in a sentence to understand its grammatical structure.

5. *Syntax and Semantic Analysis:* The understanding of the grammatical structure and meaning of sentences.

6. *Topic Modeling:* The identification of the main topics or themes within a collection of texts.

4.1 Application of NLP in Mental Health

As a result of the COVID-19 pandemic, interest in the use of NLP techniques in mental health research has increased. Social media networks contain a wealth of textual information on people's emotional states, mental health issues, and behavioral patterns (Bucur et al., 2024). Bucur et al. (2024) comprehensively analyzed NLP techniques for modeling depression on social media, highlighting progress from early feature engineering and traditional machine learning classifiers to sophisticated techniques using neural networks and word embeddings. Their survey also addresses ethical considerations surrounding data collection and processing, highlighting the importance of responsible research practices.

Research in this field has increased significantly, particularly since 2020, thanks to

workshops and collaborative activities focused on mental health modeling. Specialized dictionaries have been developed to aid in the identification of emotions and disorders. For example, Mohammad (2024) developed a dictionary containing over 44,000 English words related to anxiety, thereby enhancing the ability of NLP models to accurately analyze large text corpora containing anxiety-related content. These resources help to create more complex and reliable algorithms (Mohammad, 2024).

Traditional classifiers such as Logistic Regression, SVM, Naive Bayes, and Random Forests have frequently been used in early studies based on linguistic cues such as negative words, first-person pronouns, and emotional expressions (Bucur et al., 2024). Nowadays, deep learning models, particularly those using contextual word embeddings, have proven to be more effective in identifying small linguistic indicators associated with mental health conditions.

NLP is associated with mental health in a number of important ways (Bucur et al., 2024):

- *Automated Screening and Diagnosis:* Analyzing clinical responses and social media materials to detect early signs of depression, anxiety, and similar illnesses, hence facilitating rapid therapeutic intervention.
- *Monitoring and Support:* The continuous analysis of internet communications has made it possible to monitor mental health status in real time. This allows for early detection of deterioration and the timely provision of assistance.
- *Personalized Interventions:* Taking into account every individual's unique language patterns and feelings, NLP-powered chatbots and artificially intelligent agents may offer personalized coping mechanisms, resources, and therapeutic signs.
- *Research and Public Health Insights:* Policy and resource allocation are determined based on trends and patterns in the occurrence of mental health issues among different population groups, as revealed by large-scale text analysis.

As a result, thanks to scalable, automated, and sophisticated text data analysis, NLP offers powerful tools to improve mental health screening and intervention. The goal of ongoing developments is to increase model accuracy, resolve ethical issues, and expand applications in public health and clinical contexts.

4.2 Mental Health and Language

Language offers important insights into mental health. Consistent with previous research, individuals with mental health issues such as bipolar disorder, anxiety disorders, depressive disorders, and borderline personality disorder (BPD) exhibit distinctive language patterns (Amartey, 2023). For example, people with depression tend to use second-person or plural pronouns (e.g., "you," 'we') less frequently and first-person singular pronouns (e.g., "I," "me") more frequently, and they use symbolic language and objective expressions less frequently (Pennebaker & Chung, 2011). In texts related to mental health issues, expressions of loneliness and negative effects are becoming increasingly common.

An important work by Hofmann et al. (2012) investigates mental distortions that express excessive or irrational thought patterns associated with psychopathology and

emotional distress. These typically manifest as negative automatic thoughts that influence judgments and actions. For example, overgeneralization (for instance, "always," "never") and hyperbole (anticipating the worst possible outcome) are common in anxiety and lead to increased physiological arousal and persistent worry. Identifying these linguistic tendencies can aid in assessment and intervention (Hofmann et al., 2012; Pennebaker & Chung, 2011).

Disorganized or illogical speech can sometimes be linked to more serious mental illnesses such as schizophrenia. In such illnesses, mental processes can become disjointed and difficult to understand (Meyer & Marks, 2018). When these impairments are combined with delusions, hallucinations, and executive function disorders, they can significantly disrupt daily functioning and quality of life.

By identifying and correcting cognitive distortions, it becomes possible to create more efficient, personalized treatment plans (Hofmann et al., 2012). However, linguistic indicators should be carefully evaluated because they are correlational rather than diagnostic, vary between individuals and situations, and can be influenced by situational, cultural, and platform norms.

4.3 Sentiment Analysis

Also known as opinion mining, sentiment analysis is performed to extract evaluative or emotional attitudes from text at the document, sentence, or aspect level (Jurafsky & Martin, 2023; B. Liu, 2015; Pang & Lee, 2008). This thesis aims to provide a comprehensive categorization of mental health classifications (Normal, Depression, Suicidal Tendencies, Anxiety, Stress, Bipolar Disorder, Personality Disorder). Models map a text $\mathbf{x}$ to a label $y \in \mathcal{Y}$ by estimating $p_\theta(y \mid \mathbf{x})$ (e.g., softmax for LR/BERT) and predicting $\hat{y} = \arg\max_y p_\theta(y \mid \mathbf{x})$.

We evaluate with accuracy and weighted/macro F1 to reflect overall correctness and balance under class imbalance. For risk-sensitive labels (e.g., *Suicidal*), precision–recall curves and class-specific thresholds are reported; probability calibration may be applied to align scores with empirical frequencies. Featurizations considered include sparse lexical TF–IDF, averaged Word2Vec, sentence level embeddings (Sentence-Transformers), and contextual encoders (BERT). Classifiers span linear margins (Logistic Regression, SVM), trees/ensembles (Decision Tree, Random Forest), and neural models (CNN, LSTM, BERT).

This chapter contains a brief description of the NLP toolkit implemented in the thesis. This toolkit encompasses responsible data processing and preprocessing (tokenization, normalization, redaction of personal information), text representations (dictionaries, TF-IDF, Word2Vec, sentence-transformer embeddings), model families (linear margins, trees/ensembles, CNN/LSTM, BERT), and risk-aware evaluation (macro/weighted F1, PR curves, calibration). We also consider why language is a useful model for mental states and how various approaches strike a balance between cost, accuracy, and reviewability.

The next chapter, *Exploratory Data Analysis (EDA)*, prepares these methods for deployment by:

1. profiling the corpus (source mix, text lengths, vocabulary) and quantifying class imbalance;

2. visualizing linguistic features and correlations to motivate representation choices and guard against multicollinearity;

3. identifying risks (e.g., rare classes, outliers) that inform class weighting, thresholding, and calibration.

These diagnostics ground the modeling decisions that follow and ensure a fair and reproducible comparison between approaches.

5 Exploratory Data Analysis

In this thesis, the aim is to apply various machine learning and deep learning models to analyze sentiment and diagnose mental health conditions based on textual data from social media. The Python code explores different techniques, including traditional machine learning models, neural networks, and based on transformer models in order to determine the most effective approach for this task.

5.1 Data Source and Overview

This research utilizes a comprehensive dataset curated by Suchintika Sarkar on Kaggle (https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health, (Sarkar, 2024)), which amalgamates data from multiple sources to create a robust resource for sentiment analysis and chat-bot development.

The dataset integrates information from the following Kaggle datasets:

- 3K Conversations Dataset for Chatbot
- Depression Reddit Cleaned
- Human Stress Prediction
- Predicting Anxiety in Mental Health Data
- Mental Health Dataset Bipolar
- Reddit Mental Health Data
- Students Anxiety and Depression Dataset
- Suicidal Mental Health Dataset
- Suicidal Tweet Detection Dataset

Each record is a short statement annotated with one of seven categories: *Normal, Depression, Suicidal, Anxiety, Stress, Bipolar, or Personality Disorder* because the texts originate from diverse platforms (e.g., Reddit posts, tweets, and chatbot conversations), the corpus captures varied linguistic styles and contexts, providing a realistic testbed for sentiment and mental health classification.

In this chapter, we delve into the exploratory data analysis (EDA) of our mental health conversation dataset. Exploratory data analysis (EDA) must be used to discover underlying patterns and distributions in our mental health discussion dataset. To ensure the complexity and subtleties of the data are handled appropriately, these insights are crucial for guiding future modeling efforts.

5.1.1 Data Preparation

After reviewing the CSV file and removing unnecessary columns, it is crucial to load the dataset and perform the necessary preparatory steps. Ultimately, this process will facilitate the management of missing values and the conversion of lowercase text. For a better understanding of the dataset, the database becomes much more prepared to visualize the distribution of mental health conditions.

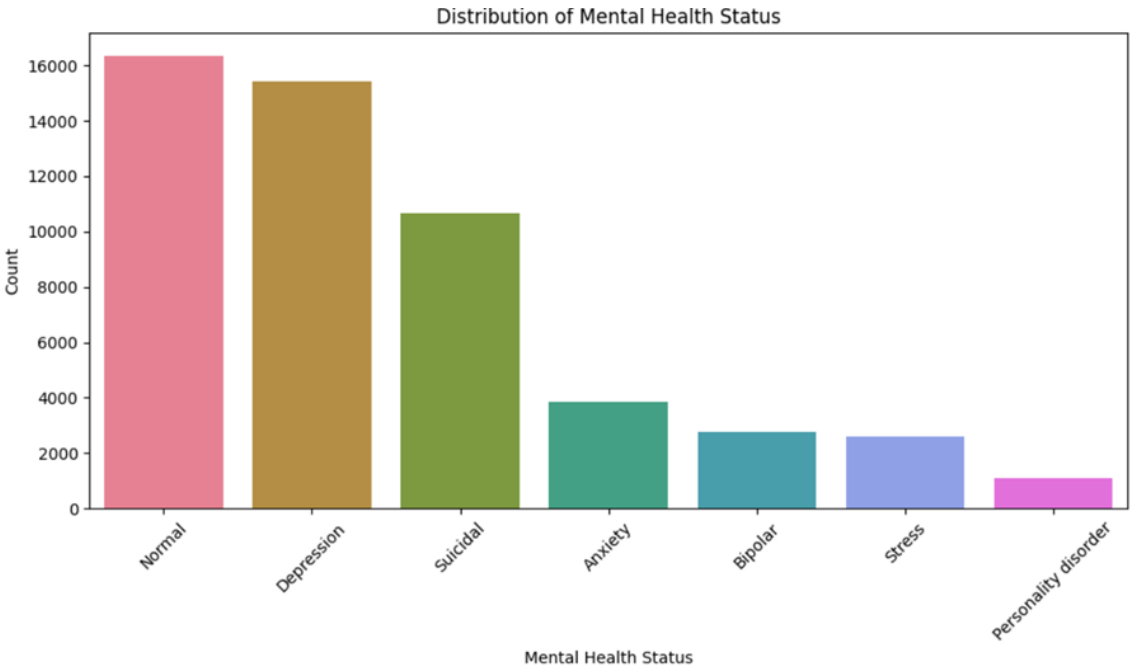

Figure 2: Distribution of Mental Health Status (Bar Chart)

5.2 Visualizing Data Distributions

We first observe at the marginal distribution of annotated mental health classes for the purpose of determining class imbalance and inform subsequent modeling decisions

As seen in Figure 2, the corpus is clearly skewed: *Normal*, *Depression*, and *Suicidal* constitute the majority of examples, whereas *Personality Disorder* is rare, with *Bipolar* and *Stress* also comparatively underrepresented. Such imbalance can bias learning toward the majority classes and inflate overall accuracy.

Several measures were subsequently employed in the study as a result of these observations, including layered train-test splits, macro and weighted F1 reporting with accuracy, and the use of techniques aware of imbalance during model building (such as class weighting, resampling, or threshold calibration). By setting expectations about which classes are most likely to be confused, label prior also informs error analysis and feature engineering.

As implemented in Listing 1, we visualize the class counts; Figure 3 summarizes the relative frequency of labels: *Normal* ($\approx 31\%$), *Depression* ($\approx 29\%$), *Suicidal* ($\approx 20\%$), *Anxiety* ($\approx 7\%$), *Bipolar* ($\approx 5\%$), *Stress* ($\approx 5\%$), and *Personality Disorder* ($\approx 2\%$). Distribution is clearly skewed toward *Normal*, *Depression*, and *Suicidal* with *Personality Disorder*, *Stress*, and *Bipolar* underrepresented. This imbalance increases skewed decision boundaries and inflated overall accuracy. An attempt will be made to reduce the dominance of the majority class by using stratified splits in future trials, publishing macro and weighted F1 scores along with accuracy, and employing adjustments that might account for imbalance (such as class weighting or resampling).

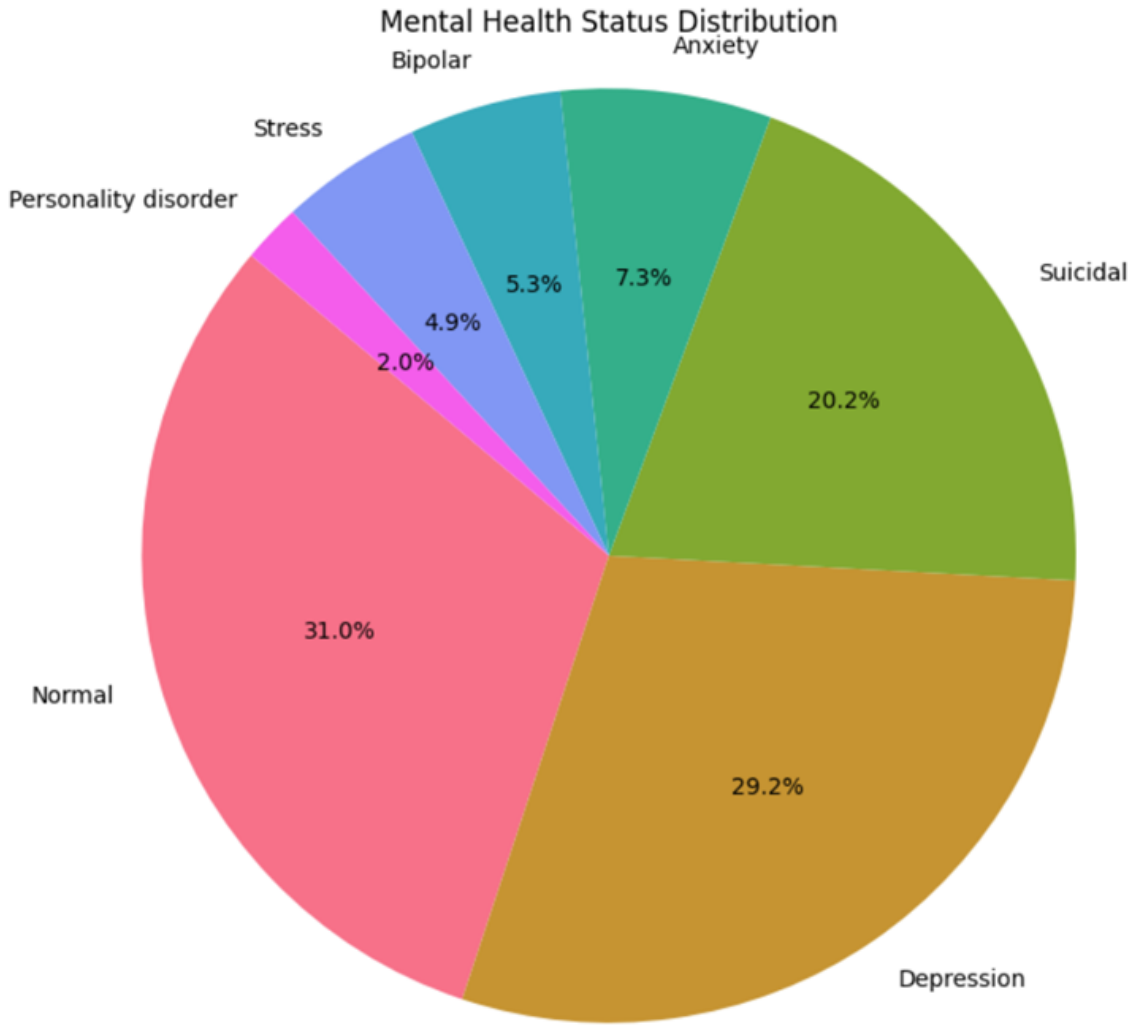

Figure 3: Mental Health Status Distribution (Pie Chart)

Listing 1: Distribution of Mental Health Statuses

```
# Visualizing the distribution of mental health statuses
plt.figure(figsize=(10, 6))
sns.countplot(data=data, x="status",
              order=data["status"].value_counts().index)
plt.title("Distribution of Mental Health Status")
plt.xlabel("Mental Health Status")
plt.ylabel("Count")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

5.2.1 Word Cloud Analysis

In this section, we generate a text cloud from preprocessed sentences to provide a qualitative perspective on word meaning. To reduce noise, we converted the text to lowercase, tokenized and lemmatized it, and removed punctuation marks and stop words. The Listing 2 shows how the visualization is produced and Figure 4 highlights high frequency unigrams (e.g., “feel,” “want,” “know,” “life”), offering a descriptive overview of dominant themes in the corpus. Word clouds are used here solely as a contextual aid for exploration, as they encode frequency rather than class specificity or polarity; subsequent quantitative analyses identify the terms that are truly distinctive among the tags.

Listing 2: Generating a Word Cloud

```
# Generating a word cloud
all_text = ' '.join(data['statement'].tolist())
wordcloud = WordCloud(width=800, height=400, background_color='
    white').generate(all_text)
plt.figure(figsize=(10, 5))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.show()
```

Figure 5 presents per class word clouds that qualitatively summarize salient vocabulary within each mental health label. Complementary descriptive analyses indicate a strong association between statement length and token count and moderate associations with vocabulary size, consistent with known regularities of written text. Correlations between surface features and labels are generally modest but informative for feature selection and transformation (e.g., normalization, log scaling) and for diagnosing potential multicollinearity. Given the pronounced class imbalance, the subsequent modeling pipeline employs stratified splits, reports macro and weighted F1 alongside accuracy, and considers imbalance aware procedures such as class weighting or resampling to ensure robust estimation and evaluation. The remainder of this chapter details the methodological framework: data sources, preprocessing steps, feature construction, model families, and evaluation protocol. By applying a unified, reproducible pipeline across all models, the study provides a rigorous basis for comparative analysis of mental health text classification.

5.2.2 Visualizing Linguistic Features by Mental Health Status

We explore the distribution of various linguistic features across different mental health statuses. These visualizations provide insights into how language use varies

Figure 4: WordCloud Analysis

Figure 5: Word Cloud on each Mental Health Status

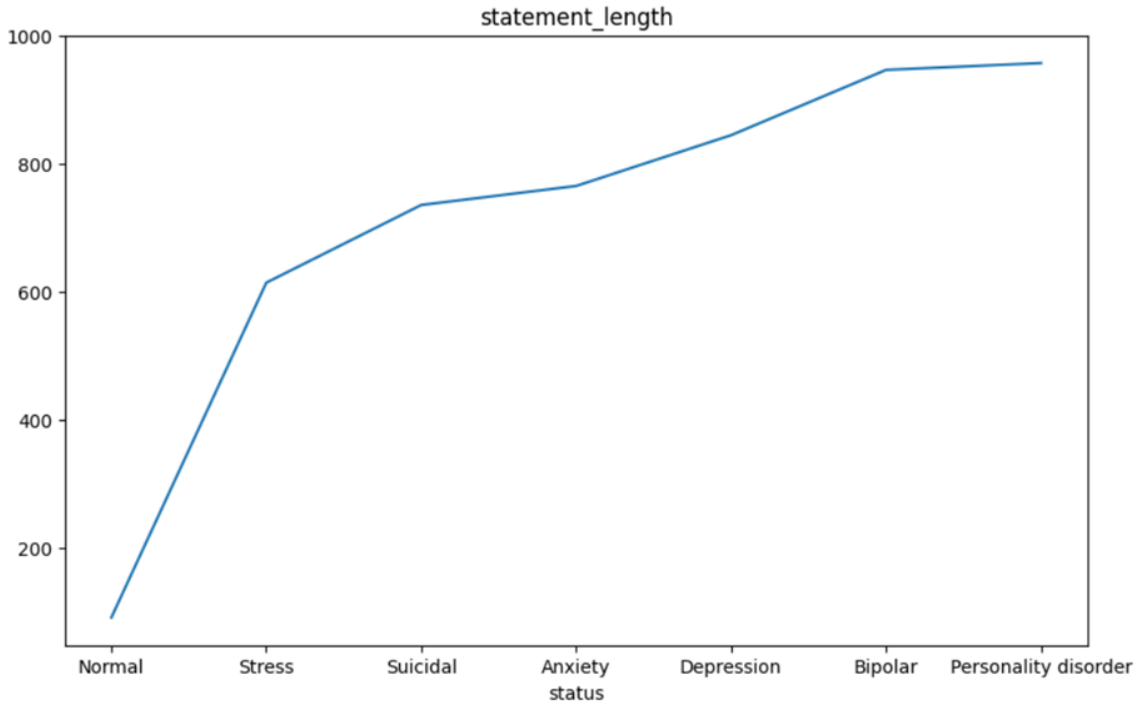

Figure 6: Average Statement Length by Mental Health Status

among individuals with different mental health conditions, which is crucial for feature engineering and model development.

The line chart (Figure 6) illustrates the average statement length across different mental health statuses. It is notable that individuals with personality disorders and bipolar disorder tend to make longer comments. These comments may reflect more complex narratives or complex emotions.

Listing 3: Word Count Distribution by Mental Health Status

```
# Creating a violin plot for word count distribution by status
plt.figure(figsize=(14, 7))
sns.violinplot(x="status", y="word_count", data=df)
plt.title("Word Count Distribution by Mental Health Status")
plt.xlabel("Status")
plt.ylabel("Word Count")
plt.xticks(rotation=45)
plt.tight_layout()
```

The violin plot presented in Figure 7 illustrates the distribution of word counts across various mental health statuses, providing a comprehensive view of linguistic behavior among individuals. This visualization integrates the advantages of both box plots and density plots, revealing not only the central tendency but also the distribution shape of word counts for each status. Notably, individuals categorized as “Suicidal” and “Bipolar” exhibit significantly higher word counts, suggesting a tendency to express their thoughts and feelings in a more elaborate manner, potentially indicative of the complexity of their emotional states. In contrast, “Anxiety” and “Normal” statuses are characterized by lower word counts, reflecting more concise expressions that may denote different cognitive processing styles or emotional regulation strategies. The presence of outliers, particularly within the “Suicidal” category, highlights instances of exceptionally long statements, which could provide valuable qualitative insights into the experiences of individuals facing severe mental

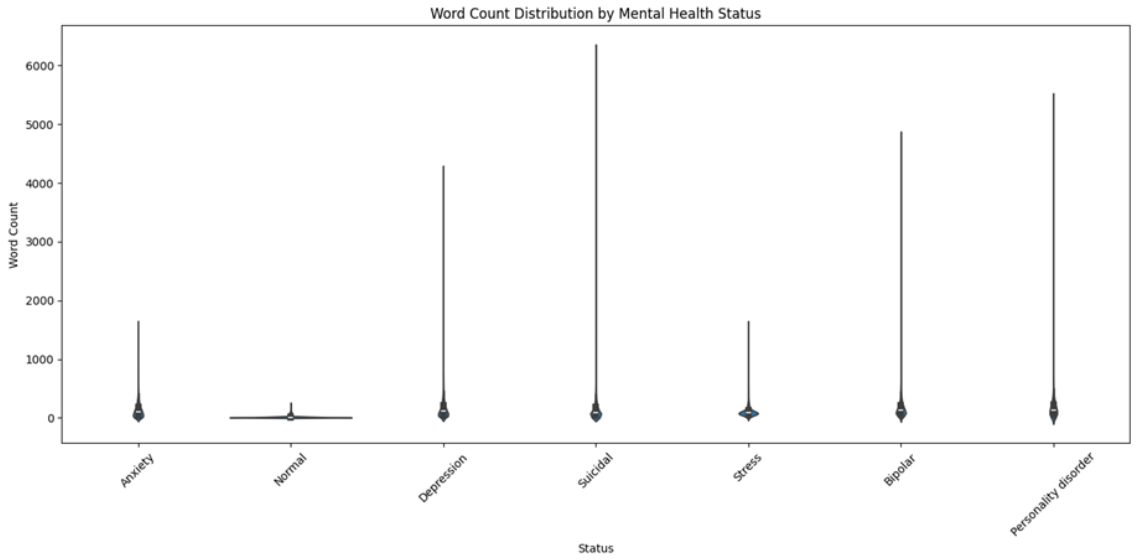

Figure 7: Violin plot for word count

health challenges. Additionally, the narrower distributions for "Stress" and "Personality Disorder" suggest more uniformity in expression, indicating common patterns in emotional or cognitive states among these individuals. These insights are crucial for feature engineering in predictive modeling, as longer statements may serve as indicators of more severe mental health conditions, while shorter statements could signify less complexity in emotional expression. By incorporating word count as a feature in our models, we aim to enhance the predictive accuracy and reliability of mental health status classification, thereby underscoring the importance of linguistic features in understanding mental health and developing effective interventions as also seen in Listing 3.

As shown in Figure 8, the average number of words per statement, revealing a similar trend as statement length. The increase in word count for bipolar and personality disorders suggests a richer linguistic output.

The average text length plot (Figure 9) indicates subtle variations across mental health statuses, with stress-related statements exhibiting slightly longer average word lengths. This could imply the use of more complex vocabulary in expressing stress.

Vocabulary size (Figure 10) across statuses highlights linguistic richness, particularly for bipolar and personality disorders. A broader vocabulary may be indicative of diverse expression patterns or complex thought processes.

Hence, the visualizations mentioned above emphasize how important language features are in understanding mental health issues. This information helps create reliable and understandable models by informing selection and transformation of features for predictive modeling.

The distribution of text lengths across sentiment categories offers insight into linguistic patterns associated with distinct psychological states. A box plot (Figure 11) summarizes both variability and central tendency for Anxiety, Normal, Depression, Suicidal Status, Stress, Bipolar, and Personality Disorder. The median (horizontal line) indicates the typical text length for each category. Although Suicidal Status and Stress have similar medians, Suicidal Status shows a higher frequency of out-

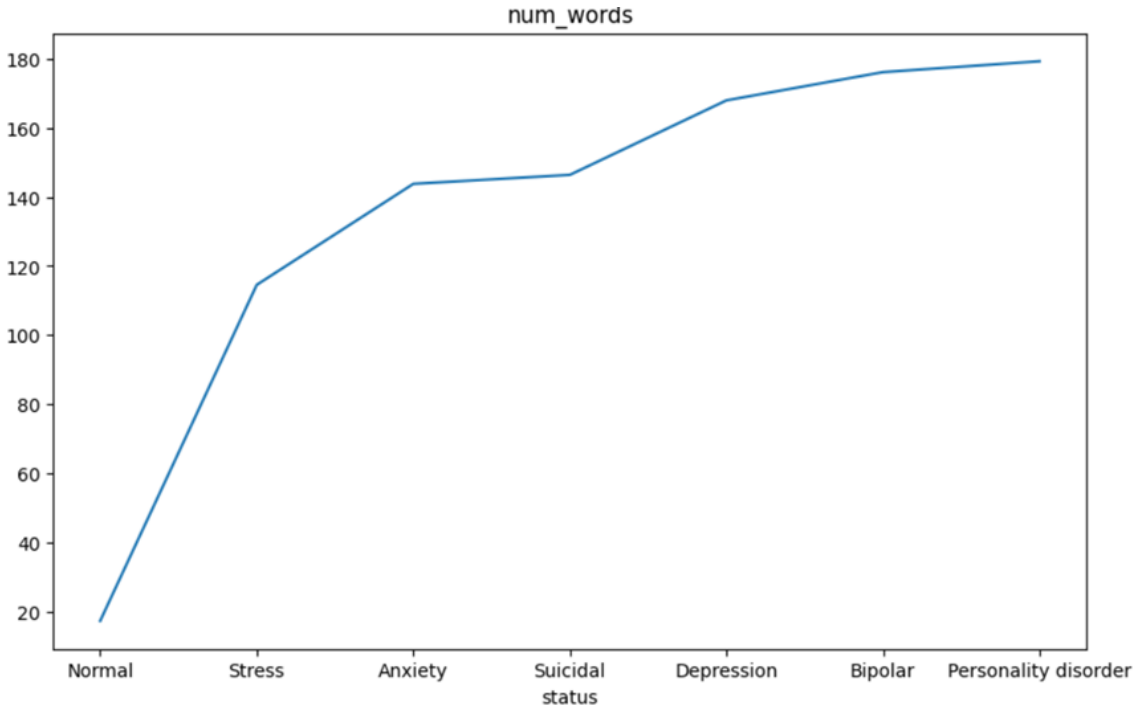

Figure 8: Number of Words by Mental Health Status

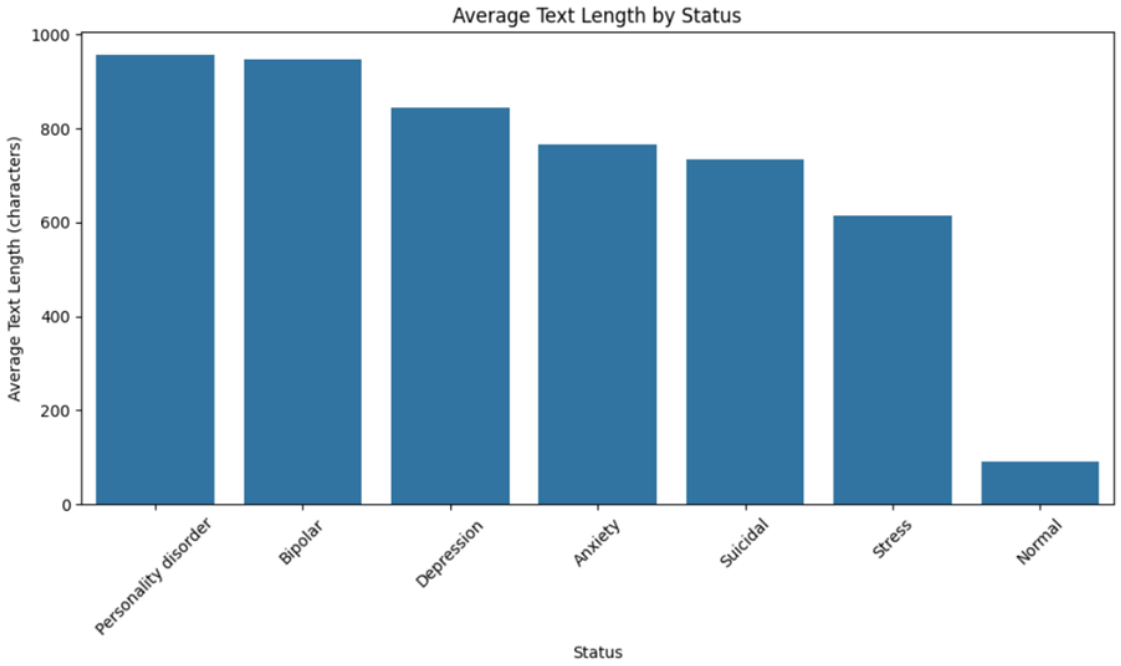

Figure 9: Average Text Length by Mental Health Status

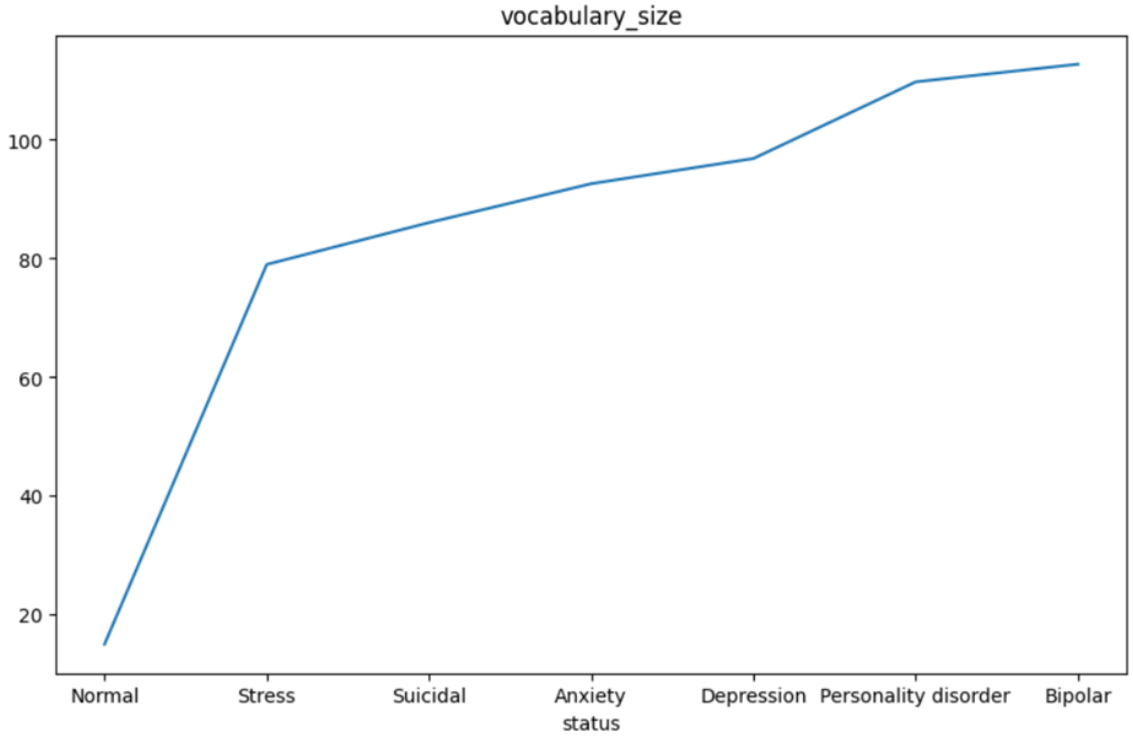

Figure 10: Vocabulary Size by Mental Health Status

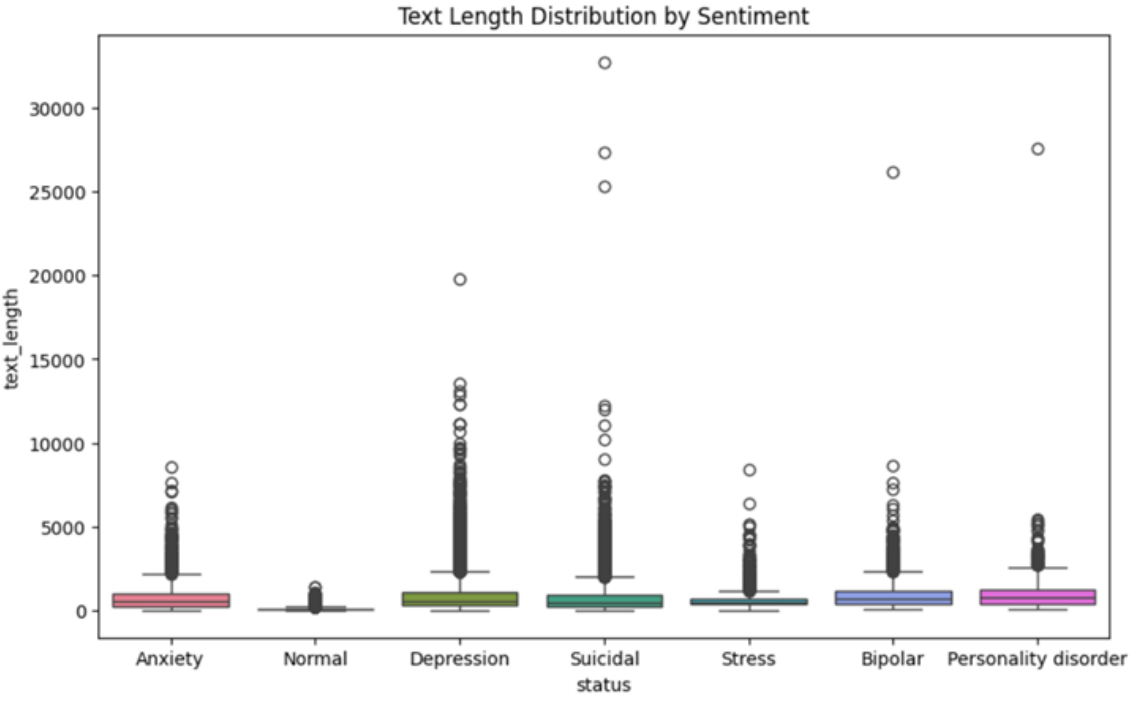

Figure 11: Text Length Distribution by Sentiment (Box Plot)

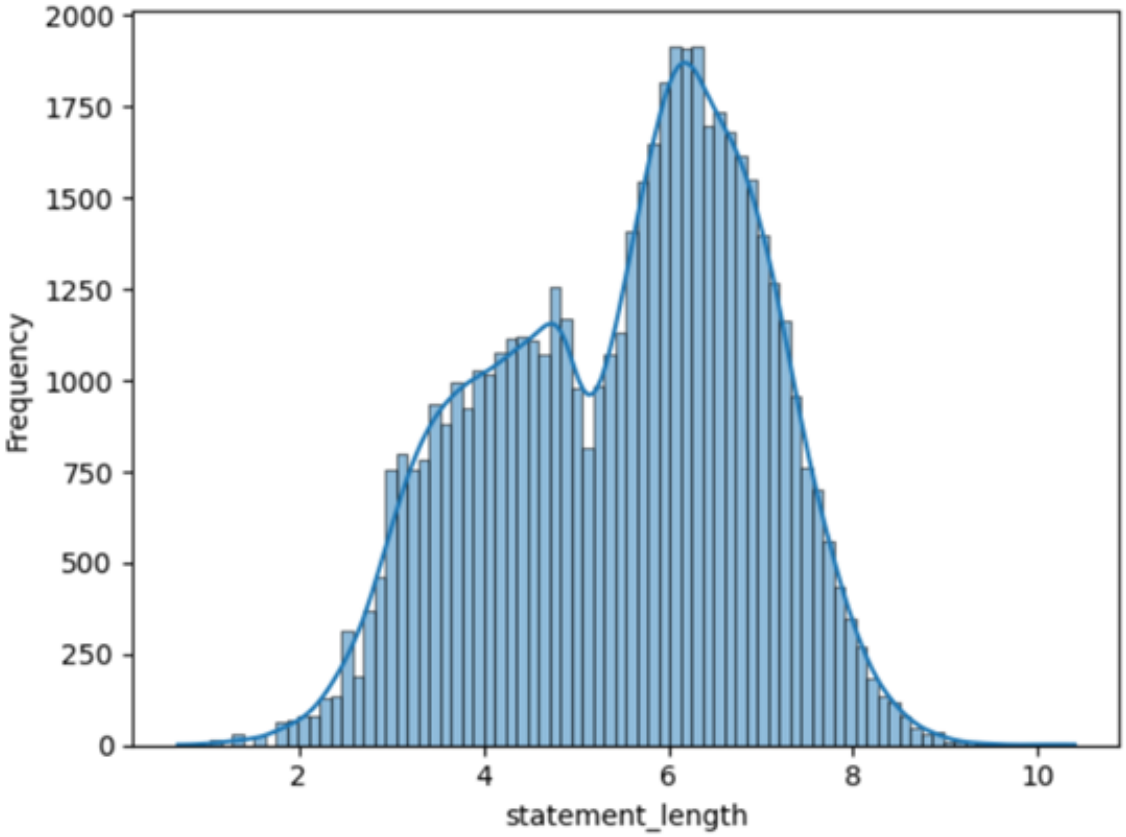

Figure 12: Log-Transformed Statement Length

liers, indicating a broader dispersion. This variability may reflect the complexity and intensity of language related to suicidal ideation. The interquartile range (IQR), represented by the box, further captures within group spread; wider boxes denote greater variability. Outliers, plotted as individual points, highlight exceptionally long or short texts that may merit qualitative follow up. Overall, these distributions underscore the value of text length as a discriminative metric for distinguishing sentiment categories and for refining our understanding of linguistic patterns in psychological discourse. The results can also inform feature selection and the development of robust predictive models in subsequent analyses.

5.3 Log-Transformed Linguistic Features

As shown in Figure 12, we use log transformation to reduce skewness and balance variance in length-based linguistic features. Trustworthy statistical analysis and model development are supported by histograms that display more easily understandable distributions.

Compared to the raw scale length (Figure 12), the log-transformed statement length exhibits a roughly symmetric, single-modal distribution with a significantly smaller right tail. By diminishing skewness and bringing the distribution closer to normal, this adjustment improves compatibility with techniques assuming Gaussian errors.

The log-transformed number of words shows a concentrated central mass and fewer extreme values, indicating reduced positive skew in Figure 13. The smoother density profile suggests improved normality and more stable variance, making this feature more suitable for inferential tests and regression modeling.

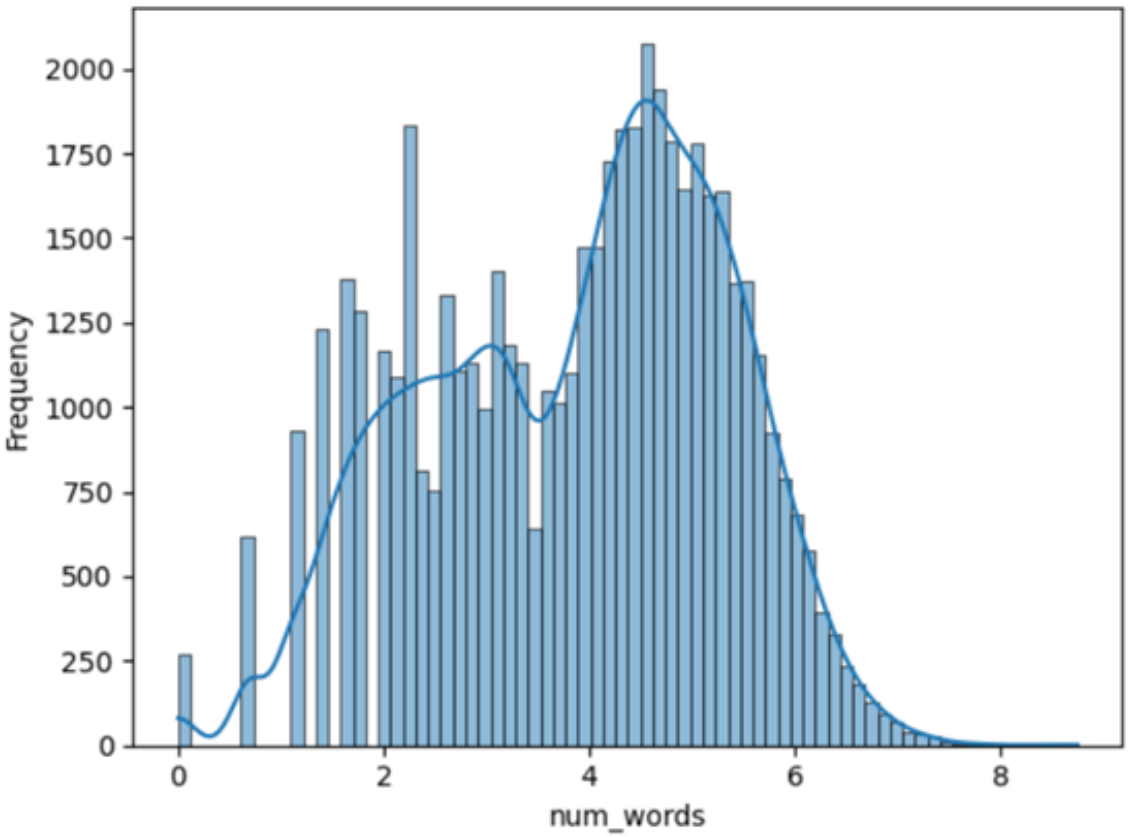

Figure 13: Log-Transformed Number of Words

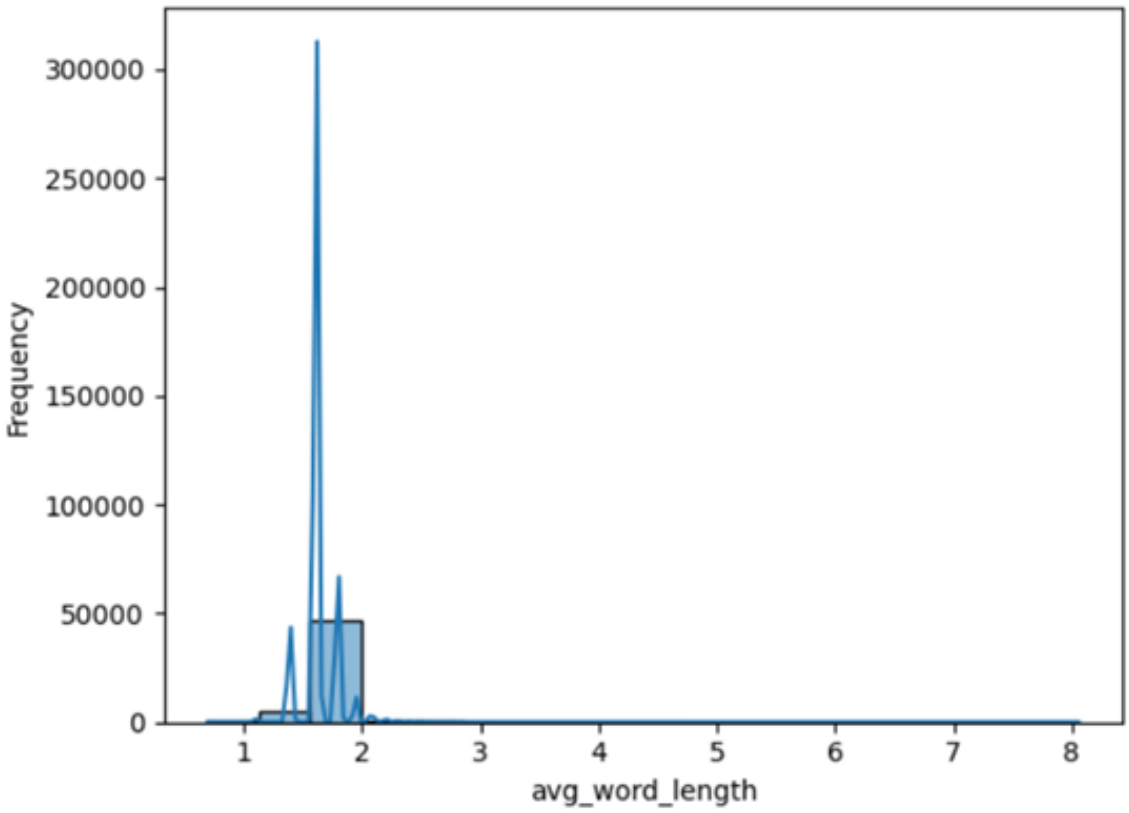

Figure 14: Log-Transformed Average Word Length

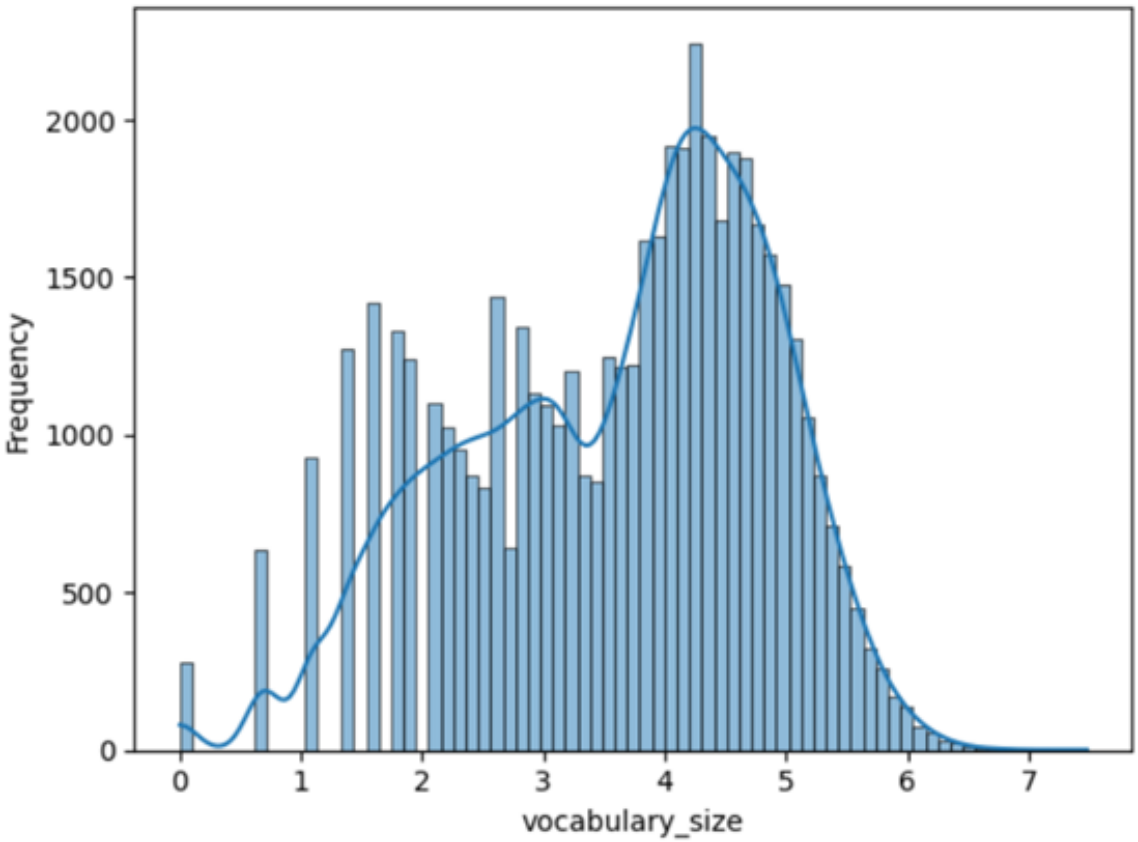

Figure 15: Log-Transformed Vocabulary Size

Despite the transformation, average word length remains heavily skewed, suggesting further preprocessing might be required as indicated in Figure 14. This highlights the complexity of linguistic data and the need for tailored feature engineering strategies.

Vocabulary size responds well to the logarithmic transformation, yielding a more balanced, near-symmetric distribution with reduced influence of extreme observations. This stabilization improves feature reliability and better satisfies assumptions of parametric analyses as shown in Figure 15.

Overall, logarithmic transformation effectively mitigates skewness for several linguistic features and produces distributions that are more amenable to statistical inference and predictive modeling. Nevertheless, the effect is feature-specific; where skewness persists (as with average word length), tailored preprocessing or robust modeling should be applied to preserve validity and generalization.

5.4 Linguistic Feature Analysis

In addition to examining marginal distributions, we assess pairwise Pearson correlations among linguistic features and their association with the encoded mental-health status. These relationships guide feature engineering by highlighting redundancy and potential predictors while guarding against multicollinearity.

The heatmap displays (Figure 16) correlations on a $[-1, 1] \times [-1, 1]$ scale. Length-related measures—`statement_length`, `num_words`, and `vocabulary_size`—are almost perfectly correlated ($r \approx 0.99$–1.00), indicating substantial redundancy. `avg_word_length` shows weak relationships with the other features ($|r| \leq 0.10$).

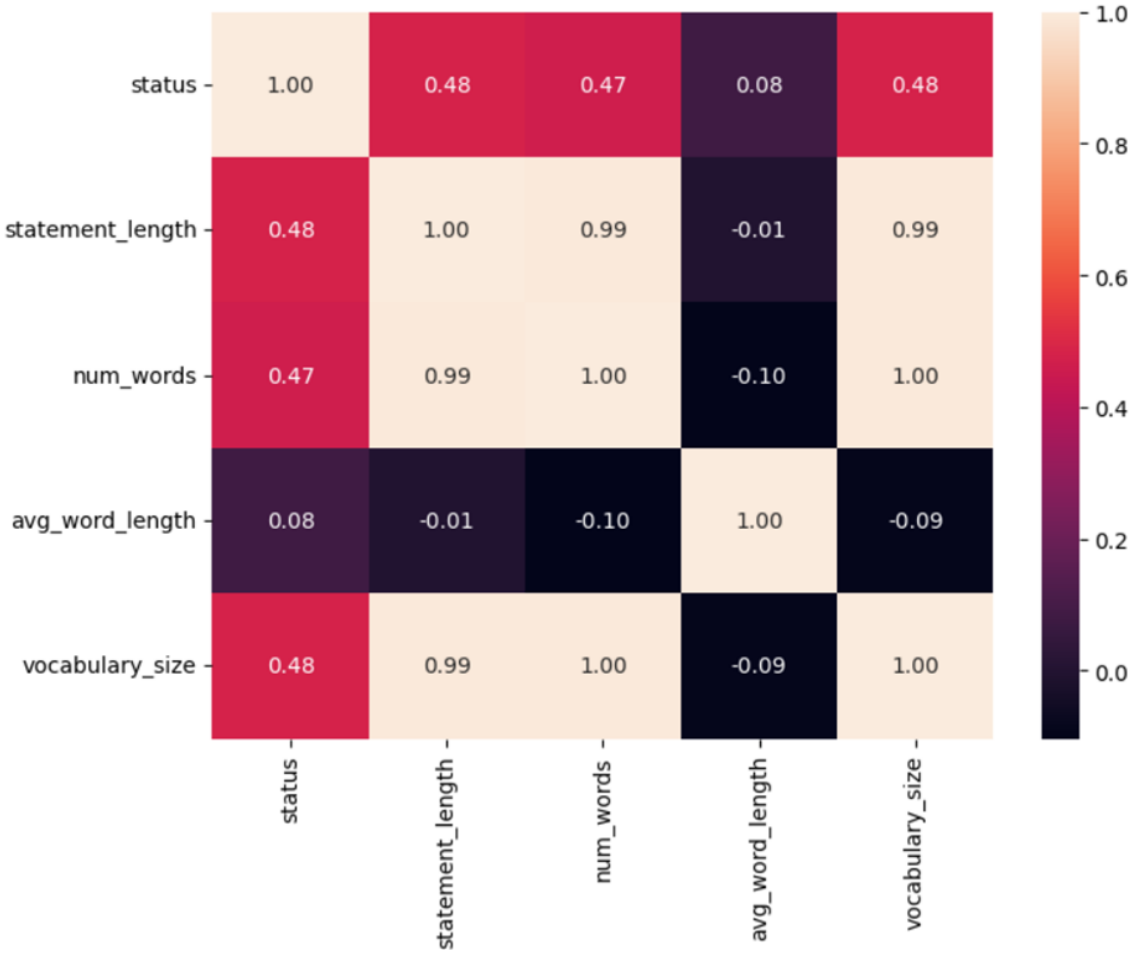

Figure 16: Correlation Heatmap of Linguistic Features

Mental health status exhibits moderate positive correlation ($r \approx 0.48$) with the length-related features and near-zero correlation with `avg_word_length`.

- *Statement Length and Number of Words:* There is a near-perfect correlation (0.99) between statement length and the number of words, indicating that longer statements naturally contain more words. This linear relationship is expected and highlights the importance of considering both features in tandem.
- *Vocabulary Size:* Vocabulary size shows strong correlations with statement length (0.99) and number of words (1.00), suggesting that richer vocabulary is associated with longer statements. This could imply more complex or detailed expressions in certain mental health statuses.
- *Average Word Length:* The average word length has low correlations with other features, indicating it might be less predictive on its own but could still contribute nuanced insights when combined with other features.
- *Mental Health Status:* The correlation of linguistic features with mental health status, though moderate (around 0.48), suggests that these features can be useful predictors when combined in a model. The moderate correlation indicates potential predictive indicators that warrant further exploration.

This correlation analysis informs feature selection by favoring high-signal, low-redundancy predictors. We prioritize one representative length measure, augment with length-normalized lexical features, and retain `avg_word_length` for complementary stylistic information to improve the reliability and accuracy of mental health status classification.

5.5 Data Preparation and Feature Extraction

Following the methodology of Hassan et al. (2025), we implemented a standardized preprocessing pipeline to optimize the performance of both ML and DL *deep learning* models while maintaining consistency across experiments.

The preprocessing steps were:

- *Cleaning the text:* Removing URLs, HTML tags, user mentions, hashtags, special characters, and extraneous whitespace using regular-expression rules to reduce noise and improve downstream modeling.
- *Lowercasing:* Converting all tokens to lowercase to ensure uniform casing and minimize sparsity due to case differences.
- *Stopword removal:* Eliminating high-frequency function words (e.g., "the," "and," "is") using the Natural Language Toolkit (NLTK) stopword list, thereby reducing redundancy and emphasizing semantically informative content.
- *Lemmatization:* Reducing words to their canonical forms (e.g., "running," "ran," "runs" $\rightarrow$ "run") with NLTK's lemmatizer, enhancing feature consistency and improving the quality of subsequent embeddings and feature extraction.

This pipeline prepared high-quality textual data suitable for effective modeling by standardizing inputs, improving consistency, and reducing extraneous variability that can impair model performance (Hassan et al., 2025).

Feature Transformation for Machine Learning Models (MLMs)

TF-IDF Vectorization is frequently used in machine learning models as a crucial step to convert text into structured numerical features. This technique transforms cleaned text data into numerical representations by calculating Term Frequency Inverse Document Frequency (TF-IDF). TF-IDF captures the frequency of keywords while minimizing the significance of words that occur excessively frequently across documents, which could otherwise distort the results (Jalilifard et al., 2021).

To be more detailed, Inverse Document Frequency (IDF) is a key component of TF-IDF. It works with Term Frequency (TF) to weight terms effectively. Essentially, TF-IDF calculates how often a word appears in a specific document compared to its commonality across the entire corpus. Words unique to a few documents get higher TF-IDF scores than common words like "the", "and." The TF-IDF vectorization method can be customized by adjusting parameters like the minimum and maximum document frequency to filter out phrases that are either too common or too rare (Aizawa, 2003; Jalilifard et al., 2021).

Term Frequency (TF) measures how often a term w appears in a document d; it reflects the importance of w within d. The (normalized) term frequency is computed as:

$$\mathrm{TF}(w, d) = \sum_{w' \in d} f_{w',d} \cdot f_{w,d} \tag{1}$$

Let D denote the corpus, $d \in D$ a document, and w a term. Let $f_{w,d}$ be the raw count of w in d, and $df_{w,D}$ the number of documents in D that contain w. We use TF–IDF to weight terms by their within-document salience and rarity in the corpus:

$$\mathrm{tf}(w, d) = \frac{f_{w,d}}{\sum_{w' \in d} f_{w',d}} \tag{2}$$

$$\mathrm{idf}(w, D) = \log\left(\frac{|D|}{df_{w,D}}\right) \tag{3}$$

$$\text{tf-idf}(w, d, D) = \mathrm{tf}(w, d)\,\mathrm{idf}(w, D) \tag{4}$$

Equations (2)–(4) assign higher weights to terms that are frequent in a specific document yet infrequent across the corpus, while ubiquitous function words receive lower weights. The scores reported in this thesis were computed in Python (Jupyter Notebook) using the above definitions.

As mentioned in this paper, TF-IDF vectorization transforms text data into numerical representation, facilitating the application of machine learning models such as logistic regression, SVM and clustering algorithms. This improves the models' ability to understand and learn from textual data (Aizawa, 2003). Overall, the implementation of TF-IDF involves calculating the TF-IDF score for each word in a query across all documents in the corpus. The documents are then ranked based on the sum of the TF-IDF scores of the query words, and the top-ranked documents are returned as the most relevant results.

All things considered, TF-IDF is an elegant and effective method for term weighting in information retrieval, enhancing the relevance of search results by focusing on the unique terms that best describe the content of individual documents.

This chapter profiled the corpus and prepared it for modeling. We documented its multi-source origin and seven-class label space and that showed us a pronounced class imbalance (over representation of *Normal*, *Depression*, *Suicidal*). In addition, we described linguistic structure (length, vocabulary, surface features). WordClouds and distributional plots provided qualitative context, while correlation analyses revealed that length-based features are highly redundant; consequently, one representative length measure is retained and log transforms are used where appropriate to reduce skewness. We also established a reproducible preprocessing pipeline (cleaning, lowercasing, stopword removal, lemmatization) and specified TF–IDF/embedding formulations together with the TF–IDF equations used throughout. These diagnostics motivate stratified splits, class weighting, probability calibration, and class-specific thresholds for risk-sensitive labels (e.g., *Suicidal*), and they justify reporting macro/weighted F1 alongside accuracy.

The next chapter, *Machine Learning Models for Mental Health Detection*, operationalizes these insights. We instantiate comparable pipelines for classical and neural models (LR/SVM, trees/ensembles, CNN/LSTM, BERT), fix train/validation/test protocols and seeds, and evaluate with per-class metrics, confusion matrices, and precision–recall analyses. Where imbalance or overlap is most pronounced, we apply the mitigation strategies identified here (class weights, calibrated scores, decision thresholds) to ensure a fair and risk-aware comparison across models.

6 Machine Learning Models for Mental Health Detection

We compare model families under a unified pipeline to ensure fair, reproducible results. Unless noted otherwise, the following settings apply:

- *Data split:* Stratified train/validation/test partitioning (fixed random seed = 42).
- *Preprocessing:* Text cleaning, lowercasing, stopword removal, and lemmatization. Neural and transformer models utilize their own tokenization schemes.
- *Representations:* TF–IDF (max_features=5000) for linear and tree-based models; We employ averaged Word2Vec embeddings to encode text in classical baseline models; padded token sequences for CNN/LSTM; and end-to-end tokenization for BERT.
- *Optimization:* Validation-driven tuning and early stopping are used where applicable, and L2 regularization is applied to linear models.
- *Imbalance handling:* `class_weight='balanced'` (linear/trees), plus class-specific thresholding and probability calibration for risk-sensitive labels (e.g., *Suicidal*).
- *Evaluation:* Our evaluation metrics comprise accuracy, macro/weighted F1, per-class confusion matrices, and precision–recall analysis; reporting settings are specified to support reproducibility.

We now detail each model, starting with logistic regression.

6.1 Logistic Regression

Logistic Regression for Mental Health Sentiment Analysis

Logistic regression (LR) models the conditional probability of a categorical label given features through a linear score transformed by a logistic/softmax link. Its convex training objective, competitive accuracy on linearly separable representations, fast inference, and transparent coefficients make it well suited for auditable NLP applications in mental health assessment and triage (Bishop, 2006; Hastie et al., 2009; Pedregosa et al., 2011).

Binary classification

Given features $\boldsymbol{x} \in \mathbb{R}^p$, LR links a linear predictor to the probability via the sigmoid:

$$p(y{=}1 \mid \boldsymbol{x}) = \sigma\Big(\boldsymbol{w}^\top \boldsymbol{x} + b\Big) = \frac{1}{1 + \exp\Big(-(\boldsymbol{w}^\top \boldsymbol{x} + b)\Big)} \tag{5}$$

Parameters are estimated by maximum likelihood, equivalently minimizing binary cross-entropy:

$$\mathcal{L}(\boldsymbol{w}, b) = -\sum_{i=1}^{n} \Big[y_i \log \hat{p}_i + (1 - y_i) \log\Big(1 - \hat{p}_i\Big) \Big], \quad \hat{p}_i = \sigma\Big(\boldsymbol{w}^\top \boldsymbol{x}_i + b\Big) \tag{6}$$

With ℓ_2 regularization, the penalized objective is:

$$\min_{\boldsymbol{w},b} \; \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \boldsymbol{w}^\top \boldsymbol{x}_i + b) \; + \; \lambda \|\boldsymbol{w}\|_2^2 \tag{7}$$

Multi-class (softmax)

$$\Pr(y{=}k \mid \boldsymbol{x}) = \frac{\exp\left(\boldsymbol{w}_k^\top \boldsymbol{x} + b_k\right)}{\sum_{j=1}^{K} \exp\left(\boldsymbol{w}_j^\top \boldsymbol{x} + b_j\right)}, \qquad k \in \{1, \ldots, K\} \tag{8}$$

with class-specific parameters $(\boldsymbol{w}_k, b_k)$ and prediction $\arg\max_k \Pr(y{=}k \mid \boldsymbol{x})$. In imbalanced settings common to mental health corpora, using class weights (e.g., inverse-frequency) can improve macro-averaged performance and minority-class recall (Pedregosa et al., 2011). When calibrated probabilities matter (e.g., risk triage), post-hoc calibration and threshold selection on a validation split are recommended (Niculescu-Mizil & Caruana, 2005).

6.1.1 Relation to Sentiment/NLP for Mental Health

In text classification, LR is typically applied to high-dimensional features (e.g., TF–IDF or dense sentence embeddings). Its linear decision surface supports interpretable analyses of which features push predictions toward clinically relevant categories (e.g., depression, anxiety). Combined with careful preprocessing and class weighting, LR offers a strong, transparent baseline for mental health sentiment analysis before moving to more complex, less interpretable models (Bishop, 2006; Hastie et al., 2009; Pedregosa et al., 2011).

6.1.2 Word2Vec Embeddings with Logistic Regression

Bag-of-words features are purely lexical. Word2Vec learns distributed representations by predicting a word from its context (CBOW) or the context from a target word (skip-gram), placing semantically related terms near one another in the embedding space (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013). To obtain fixed-length document vectors for LR, we average in-vocabulary token embeddings and standardize the result, which provides a lightweight semantic summary suitable for linear classification.

We train Word2Vec on the tokenized *training* texts only (to avoid test leakage), using `vector_size=100`, `window=5`, `min_count=2` in Listing 4. A document vector is the mean of its in-vocabulary token vectors; we then apply `StandardScaler` and fit multinomial LR with `solver='lbfgs'`, `class_weight='balanced'`, `max_iter=2000`, `random_state=42`. Listing 4 shows the training code.

Listing 4: Train Word2Vec on tokenized sentences and classify with Logistic Regression

```
import nltk, string, re, numpy as np, os
from nltk.tokenize import word_tokenize
from gensim.models import Word2Vec
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from joblib import dump

nltk.download('punkt', quiet=True)

def clean_text(t):
    t = t.lower()
    t = t.translate(str.maketrans('', '', string.punctuation))
    t = re.sub(r'\s+', ' ', t).strip()
    return t

Xtr_clean = [clean_text(t) for t in X_train_text]
Xte_clean = [clean_text(t) for t in X_test_text]
Xtr_tokens = [word_tokenize(t) for t in Xtr_clean]
Xte_tokens = [word_tokenize(t) for t in Xte_clean]

vec_size = 100
w2v = Word2Vec(sentences=Xtr_tokens, vector_size=vec_size, window
    =5, min_count=2, workers=4)

def sent_vec(tokens):
    iv = [w for w in tokens if w in w2v.wv]
    if not iv: return np.zeros(vec_size, dtype=np.float32)
    return np.mean(w2v.wv[iv], axis=0)

Xtr_w2v = np.vstack([sent_vec(t) for t in Xtr_tokens])
Xte_w2v = np.vstack([sent_vec(t) for t in Xte_tokens])

scaler = StandardScaler()
Xtr_w2v = scaler.fit_transform(Xtr_w2v)
Xte_w2v = scaler.transform(Xte_w2v)

lr_w2v = LogisticRegression(solver='lbfgs', multi_class='
    multinomial',
                            class_weight='balanced', max_iter=2000,
                                random_state=42)
lr_w2v.fit(Xtr_w2v, y_train)
y_pred = lr_w2v.predict(Xte_w2v)

# Optional: save artifacts
dump(w2v, os.path.join(output_dir, "word2vec_model.pkl"))
dump(scaler, os.path.join(output_dir, "w2v_scaler.pkl"))
dump(lr_w2v, os.path.join(output_dir, "w2v_logreg.pkl"))
```

Table 1 summarizes performance. Overall, accuracy is 0.61; macro- and weighted-average F1 are 0.53 and 0.63. Per-class F1 shows heterogeneity: *Normal* 0.74, *Suicidal* 0.66, *Anxiety* 0.60, *Depression* 0.59, *Bipolar* 0.53, with lower values for *Stress* 0.34 and *Personality disorder* 0.29. The confusion matrix (Figure 17) indicates frequent confusions from minority labels into *Normal* or *Depression*, consistent with imbalance and overlapping lexical/semantic cues.

Class	Precision	Recall	F1-score	Support
Anxiety	0.55	0.65	0.60	768
Bipolar	0.45	0.63	0.53	556
Depression	0.73	0.50	0.59	3081
Normal	0.86	0.64	0.74	3269
Personality disorder	0.18	0.65	0.29	215
Stress	0.27	0.47	0.34	517
Suicidal	0.60	0.75	0.66	2131
Accuracy			0.61	10537
Macro avg	0.52	0.61	0.53	10537
Weighted avg	0.68	0.61	0.63	10537

Table 1: Word2Vec + Logistic Regression classification report

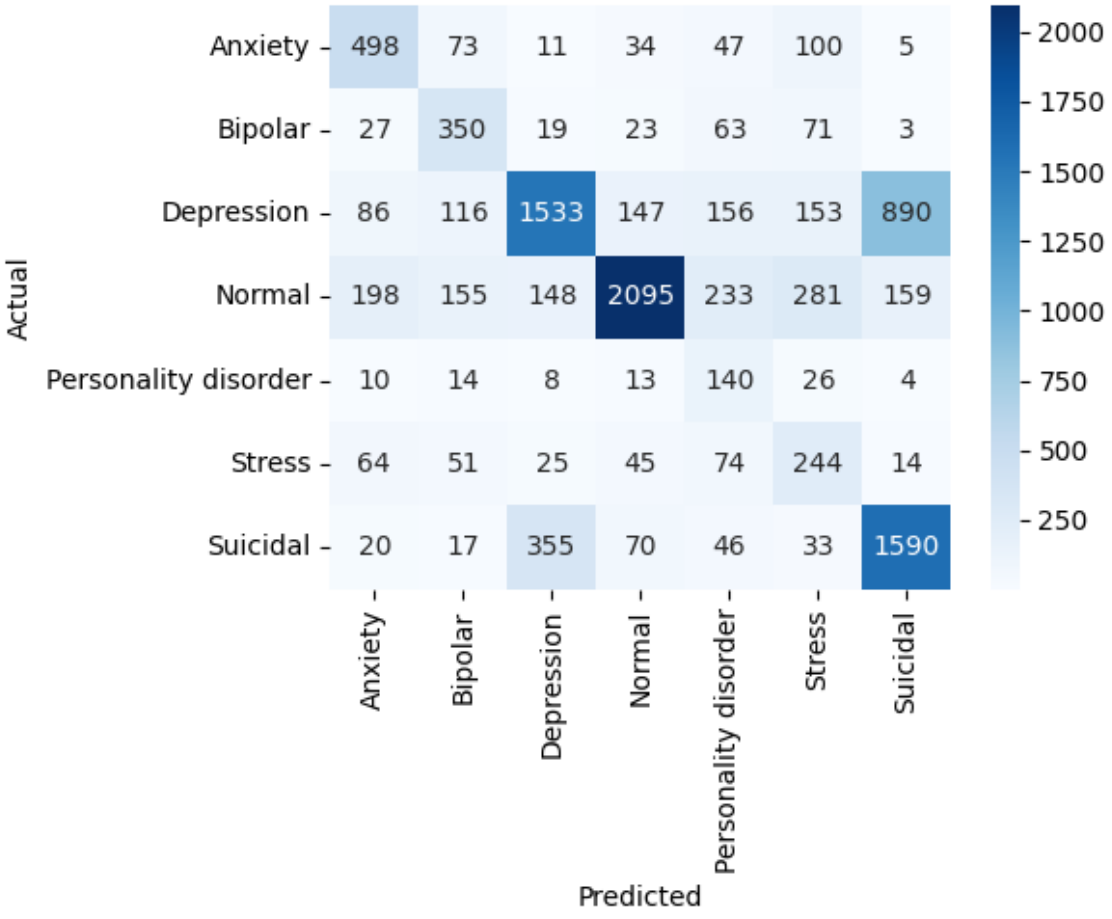

Figure 17: Confusion matrix for Word2Vec + Logistic Regression

Interpretation for mental health text

Averaged Word2Vec + LR is fast, reproducible, and auditable. It captures broad semantic signals (synonymy, paraphrase) more robustly than purely lexical counts while retaining a linear decision surface for interpretability. Its main limitations are loss of word-order/compositional cues due to averaging and sensitivity to label imbalance; probability calibration and class-specific thresholds are advisable when prioritizing recall for high-risk categories (e.g., suicidal intent) (Niculescu-Mizil & Caruana, 2005; Pedregosa et al., 2011).

6.1.3 TF–IDF with Logistic Regression

TF–IDF provides a high-dimensional, sparse representation that emphasizes terms distinctive for a document relative to the corpus, which is effective for short social-media posts and forum entries common in mental health datasets. When paired with multinomial logistic regression (LR) under ℓ_2 regularization, the result is a convex, data-efficient classifier whose coefficients yield transparent lexical contributions. This combination remains a strong baseline for text classification and sentiment analysis: it scales well, is straightforward to calibrate, and supports auditable decisions for safety-critical use cases such as risk screening (Jurafsky & Martin, 2023; Pedregosa, Varoquaux, et al., 2020; Sammut & Webb, 2017).

Let $\boldsymbol{x} \in \mathbb{R}^p$ be the TF–IDF vector for a post. The multinomial LR predicts

$$\Pr(y{=}k \mid \boldsymbol{x}) = \frac{\exp(\boldsymbol{w}_k^\top \boldsymbol{x} + b_k)}{\sum_{j=1}^{K} \exp(\boldsymbol{w}_j^\top \boldsymbol{x} + b_j)}, \qquad k \in \{1, \ldots, K\}. \tag{9}$$

with parameters $\{(\boldsymbol{w}_k, b_k)\}$ learned by maximum likelihood and ℓ_2 regularization (in `scikit-learn`, $C{=}1/\lambda$).

Class imbalance is addressed by inverse-frequency weights (`class_weight=balanced`), which improves macro-averaged performance on minority labels.

As shown in Listing 5, we fit TF–IDF with `max_features=5000` and multinomial LR (L–BFGS, ℓ_2 penalty, `class_weight=balanced`, `max_iter=2000`, `random_state=42`). Listing 5 shows the training and evaluation code.

Listing 5: TF–IDF + Logistic Regression: training and evaluation

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.pipeline import Pipeline
from joblib import dump
import os

tfidf_lr = Pipeline([
    ('tfidf', TfidfVectorizer(max_features=5000)),
    ('lr', LogisticRegression(solver='lbfgs', multi_class='
       multinomial',
                              class_weight='balanced', max_iter
                                 =2000, random_state=42))
])

tfidf_lr.fit(X_train_text, y_train)
y_pred = tfidf_lr.predict(X_test_text)
print("TF-IDF + Logistic Regression")
evaluate_preds(y_test, y_pred, le)

# Save artifacts
dump(tfidf_lr.named_steps['lr'], os.path.join(output_dir, "
   tfidf_logreg.pkl"))
dump(tfidf_lr.named_steps['tfidf'], os.path.join(output_dir, "
   tfidf_vectorizer.pkl"))
```

Table 2 reports the classification metrics; the confusion matrix is shown in Figure 18.

Class	Precision	Recall	F1-score	Support
Anxiety	0.77	0.80	0.78	768
Bipolar	0.72	0.80	0.76	556
Depression	0.80	0.60	0.68	3081
Normal	0.91	0.91	0.91	3269
Personality disorder	0.44	0.74	0.55	215
Stress	0.50	0.72	0.59	517
Suicidal	0.67	0.77	0.71	2131
Accuracy			0.76	10537
Macro avg	0.69	0.76	0.71	10537
Weighted avg	0.78	0.76	0.76	10537

Table 2: TF–IDF + Logistic Regression classification report.

The model achieves strong performance for *Normal* (F1 = 0.91) and good recall for *Suicidal* (0.77), indicating that salient lexical markers are captured by TF–IDF features. A substantial portion of *Depression* is confused with *Suicidal*, suggesting overlapping vocabulary (e.g., hopelessness, self-harm references). Lower scores for *Personality disorder* and *Stress* are consistent with limited support and greater lexical heterogeneity. For risk-sensitive use, threshold tuning and probability calibration can prioritize recall on high-risk classes, while feature enrichment (e.g., character n-grams, domain lexicons, or task-adapted embeddings) may reduce minority-class confusions without sacrificing LR's transparency.

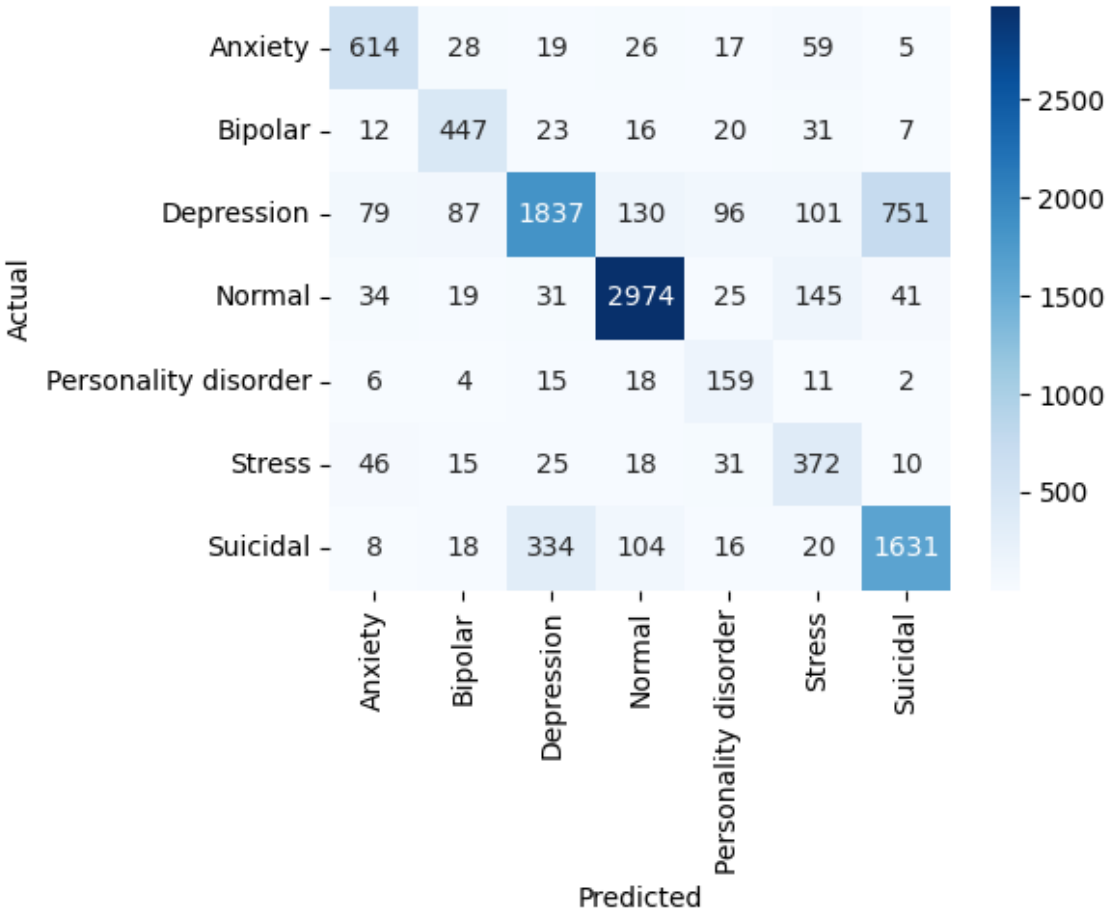

Figure 18: Confusion matrix for TF–IDF + Logistic Regression

6.1.4 Sentence-Transformer Embeddings with Logistic Regression.

Sentence-level embeddings map each text to a dense vector that encodes its meaning in a fixed-dimensional space. We use a compact model from the SENTENCE-TRANSFORMERS family (`all-MiniLM-L6-v2`; 384 dim) trained with contrastive objectives so that semantically related sentences lie close together in the embedding space (Reimers & Gurevych, 2019; W. Wang et al., 2020). Each post is encoded to an ℓ_2-normalized embedding $\mathbf{z}$, and a linear classifier (logistic regression) is fit on top (Pedregosa et al., 2011). The classifier estimates class probabilities via the softmax of a linear score,

$$\Pr(y{=}k \mid \mathbf{z}) = \frac{\exp(\mathbf{w}_k^\top \mathbf{z} + b_k)}{\sum_j \exp(\mathbf{w}_j^\top \mathbf{z} + b_j)}. \tag{10}$$

with class-specific parameters $(\mathbf{w}_k, b_k)$. To compensate for label imbalance, class weights are set to `balanced`.

Texts are lowercased strings and encoded in batches; embeddings are ℓ_2-normalized. We fit `LogisticRegression` (solver=`lbfgs`, multi-class=`multinomial`, max_iter= 5000, `class_weight=balanced`, `random_state`=42) on the training split and evaluate on the held out test set using per-class precision/recall/F1 together with macro- and weighted-average F1 and overall accuracy.

Table 3 presents that the accuracy is 0.73, with macro and weighted F1 of 0.67 and 0.74, respectively. Per-class F1 shows strong performance on *Normal* (0.91) and competitive scores on *Anxiety* (0.75), *Bipolar* (0.71), and *Suicidal* (0.66). *Depression* achieves 0.65 (precision 0.77, recall 0.57). The model attains high recall but modest precision for *Personality disorder* (precision 0.30, recall 0.77, F1 0.43) and improved recall for *Stress* (precision 0.45, recall 0.70, F1 0.55). These patterns indicate that sentence-level semantics markedly improve discrimination over purely lexical baselines for several classes, while minority categories still require careful treatment of class imbalance and decision thresholds.

Class	Precision	Recall	F1-score	Support
Anxiety	0.72	0.79	0.75	768
Bipolar	0.64	0.80	0.71	556
Depression	0.77	0.57	0.65	3081
Normal	0.92	0.90	0.91	3269
Personality disorder	0.30	0.77	0.43	215
Stress	0.45	0.70	0.55	517
Suicidal	0.65	0.67	0.66	2131
Accuracy			0.73	10537
Macro avg	0.64	0.74	0.67	10537
Weighted avg	0.76	0.73	0.74	10537

Table 3: Sentence-Transformers + Logistic Regression classification report

As shown in Figure 19, rows denote actual classes and columns denote predicted classes; darker cells indicate higher counts. The matrix exhibits a clear diagonal, with strongest separation for *Normal*. The largest off-diagonal confusions are between *Depression* and *Suicidal*, while *Personality disorder* and *Stress* are often

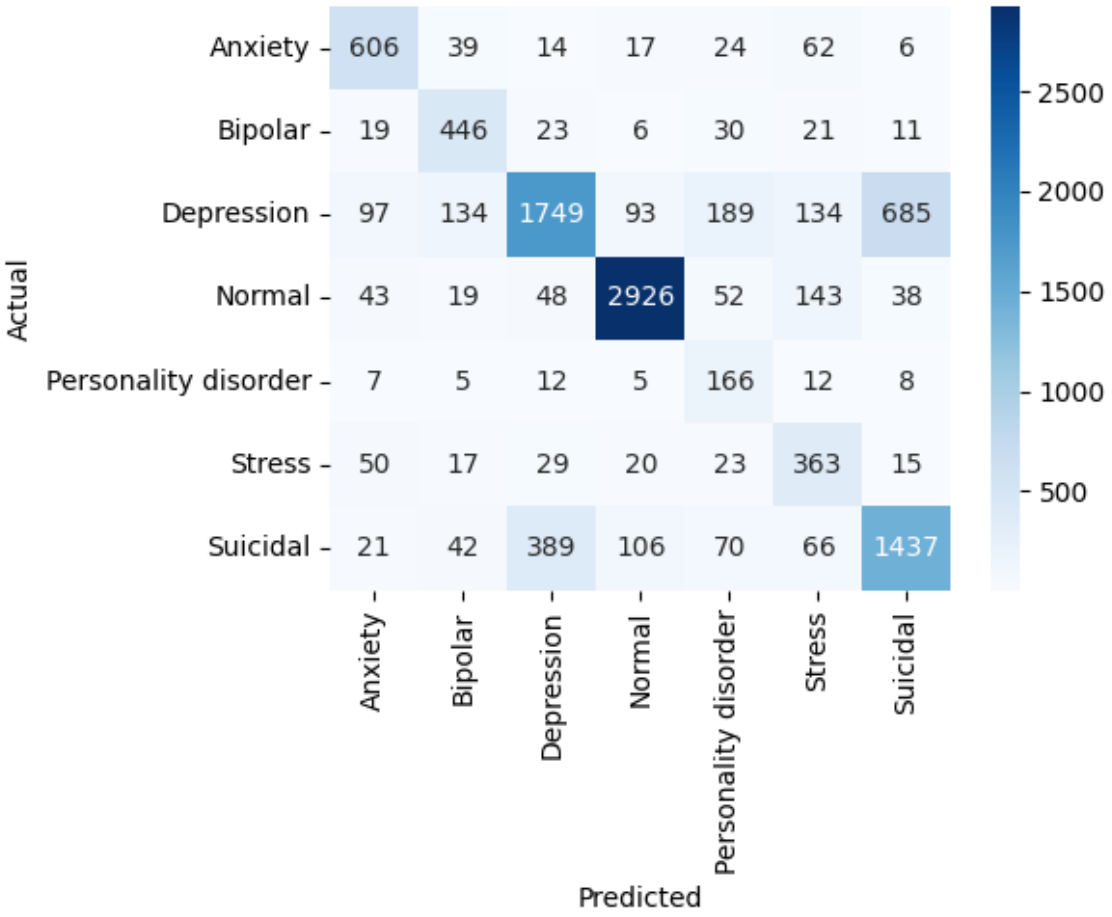

Figure 19: Sentence-Transformers + Logistic Regression

mapped to *Normal/Depression*, reflecting class-imbalance and overlapping language. Per-class recalls from Table 3 are: Normal 90%, Anxiety 79%, Bipolar 80%, Depression 57%, Suicidal 67%, Stress 70%, Personality disorder 77%. These patterns align with the precision/recall trade-offs reported in Table 3 and motivate class-specific thresholds, calibration, and imbalance-aware learning.

Interpretation for Mental Health Text

- **Sentence-level embeddings** aggregate phrasing variants (e.g., paraphrases of distress) and thus provide robustness to surface form, yielding consistent gains on several classes (Reimers & Gurevych, 2019).
- **Minority labels** remain challenging; high recall with relatively low precision (e.g., *Personality disorder*) suggests tuning thresholds and class costs to the application's risk profile.
- **Compared with purely lexical features**, semantic embeddings better capture context and meaning, but can still confuse clinically adjacent categories when language overlaps or when data are sparse.
  - *Stronger encoders or task-adaptive fine-tuning:* Evaluating larger models (e.g., `e5`, `bge`) and consider supervised fine-tuning on the labeled corpus (Reimers & Gurevych, 2019).
  - *Imbalance-aware learning:* Managing with tuning C and class weights; combine with re-sampling or cost-sensitive objectives to raise recall for rare, high-risk labels.
  - *Thresholding and calibration:* Calibrating scores and apply class-specific thresholds to reduce false negatives in safety-critical categories; assess macro/weighted F1 alongside PR curves.
  - *Feature enrichment and error analysis:* Where terminology is highly overlapping, add character n-grams or domain lexicons alongside embeddings and inspect common confusions to refine label definitions and annotations.

Overall, sentence-transformer embeddings paired with logistic regression provide a fast and reproducible semantic baseline. With imbalance handling and task-specific adaptation, the approach achieves substantial improvements suitable for mental health text classification at scale.

6.2 Random Forest

Random Forest (RF) aggregates many decision trees trained on bootstrap samples and random feature subsets, reducing variance relative to a single tree and providing strong out-of-sample performance on high-dimensional text (Breiman, 2001; Pedregosa et al., 2011). RFs are robust to noisy or weakly-informative terms and expose feature importance for auditability.

We fit RF on TF–IDF features in a leakage-safe pipeline. The configuration below uses 300 trees as seen in Listing 6; other hyperparameters were kept at library defaults for a clean baseline.

Table 4 shows us the overall accuracy is 0.70 (macro-F1 0.60, weighted-F1 0.69). *Normal* is strongest (F1 0.88), and *Depression* achieves balanced F1 0.66 with

Listing 6: TF–IDF + Random Forest: training and evaluation

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.ensemble import RandomForestClassifier
from sklearn.pipeline import Pipeline

rf_pipe = Pipeline([
    ('tfidf', TfidfVectorizer(max_features=5000)),
    ('rf', RandomForestClassifier(n_estimators=300, random_state
       =42))
])
rf_pipe.fit(X_train_text, y_train)
y_pred = rf_pipe.predict(X_test_text)
print("TF-IDF + RandomForest")
evaluate_preds(y_test, y_pred, le)
```

Class	Precision	Recall	F1-score	Support
Anxiety	0.91	0.48	0.63	768
Bipolar	0.98	0.46	0.63	556
Depression	0.55	0.81	0.66	3081
Normal	0.82	0.95	0.88	3269
Personality disorder	1.00	0.29	0.45	215
Stress	0.98	0.23	0.37	517
Suicidal	0.73	0.48	0.58	2131
Accuracy			0.70	10537
Macro avg	0.85	0.53	0.60	10537
Weighted avg	0.75	0.70	0.69	10537

Table 4: TF–IDF + Random Forest classification report

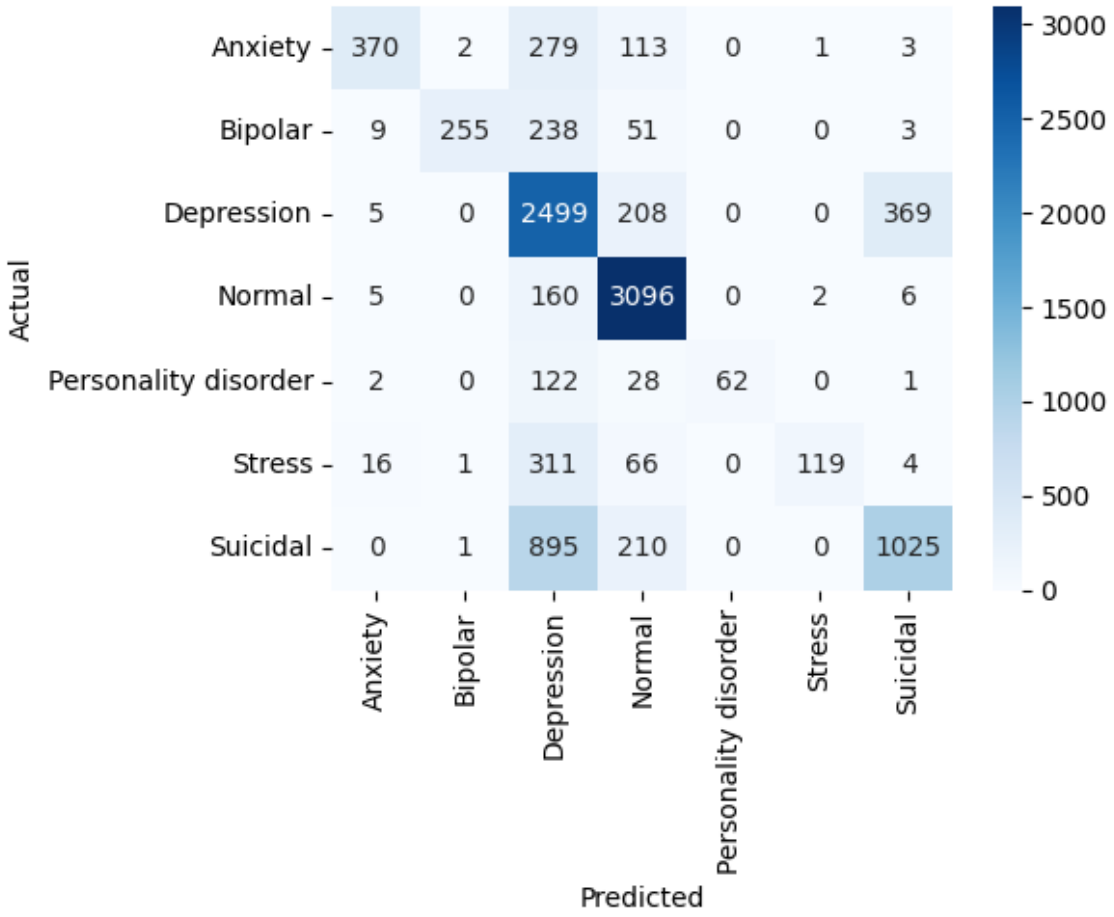

Figure 20: Confusion matrix for Random Forest

higher recall than precision. *Anxiety* and *Bipolar* show high precision but low recall (both F1 0.63), indicating conservative voting that misses positives. Minority labels *Personality disorder* (F1 0.45) and *Stress* (F1 0.37) have low recall despite very high precision, a typical pattern under class imbalance with lexical overlap. *Suicidal* attains F1 0.58; errors concentrate into *Depression* and *Normal*, consistent with the confusion matrix (Figure 20). These profiles suggest that RF captures majority-class cues well and recall on minority. In addition, clinically salient classes requires imbalance-aware training and/or richer representations.

Relation to LSTM and Hybridization

Alongside the tree-based baseline, we considered Long Short-Term Memory (LSTM) networks, which capture word order and long-range dependencies that bag-of-words features do not (Hochreiter & Schmidhuber, 1997). Random Forests (RF) and LSTMs embody complementary inductive biases: RFs exploit high-dimensional lexical signals with strong regularization, whereas LSTMs model sequential structure. Hybrid use is pragmatic in two forms:

(i) Using LSTM derived embeddings as fixed features for an RF; and

(ii) Stacking, where RF and LSTM predictions are combined by a meta classifier.

Note: RF does not "update" neural weights; integration is via feature-based fusion or ensemble learning (e.g., stacking), not joint training.

Limitations and Next Steps

- Tune class-balanced objectives (e.g., `class_weight`, thresholding on validation), increase trees, and explore `max_features`/depth for bias–variance trade-offs.
- Enriching features with character n-grams and domain lexicons; consider LSA/TruncatedSVD before RF to reduce sparsity.
- Calibrating probabilities (e.g., isotonic/Platt) when operating points matter, especially for *Suicidal*.
- Comparing with transformer baselines and evaluating hybrid stacking to raise minority-class recall while maintaining interpretability.

6.3 Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes' theorem and the simplifying assumption that features are conditionally independent given the class label. For an input vector $\mathbf{x} = (x_1, \ldots, x_n)$ and class C_k, the posterior is

$$P(C_k \mid \mathbf{x}) \;\propto\; P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k), \tag{11}$$

and the predicted class is obtained by

$$\hat{k} \;=\; \arg\max_{k} \; P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k). \tag{12}$$

Despite the "naive" independence assumption, Naive Bayes performs well on text classification because bag-of-words features are high-dimensional, sparse, and ap-

proximately independent. It trains in linear time, scales to large corpora, and is robust to many irrelevant features (Pang & Lee, 2008; Ravi & Ravi, 2015).

Class	Precision	Recall	F1-score	Support
Anxiety	0.81	0.57	0.67	768
Bipolar	0.93	0.39	0.55	556
Depression	0.51	0.82	0.63	3081
Normal	0.85	0.82	0.83	3269
Personality disorder	1.00	0.11	0.19	215
Stress	0.90	0.10	0.19	517
Suicidal	0.73	0.53	0.61	2131
Accuracy			0.67	10537
Macro avg	0.82	0.48	0.52	10537
Weighted avg	0.73	0.67	0.66	10537

Table 5: TF–IDF + MultinomialNB classification report (test set).

We employ Multinomial Naive Bayes (MNB) with TF–IDF representations of the statements. MNB models token frequencies (or TF–IDF weights in practice) with additive/Lidstone smoothing; the smoothing hyperparameter $\alpha > 0$ is tuned by cross-validation. The pipeline lowercases text, computes TF–IDF vectors, and then fits the MNB classifier (Pedregosa et al., 2011). Performance is summarized with accuracy and class-wise precision, recall, and F1-score, and confusion matrices are inspected to diagnose systematic errors.

In Table 5, overall accuracy was 0.67; macro- and weighted-average F1 were 0.52 and 0.66, respectively. Class-wise F1-scores were: *Normal* 0.83, *Depression* 0.63 (precision 0.51, recall 0.82), *Anxiety* 0.67, *Bipolar* 0.55, *Suicidal* 0.61, and lower values for *Stress* 0.19 and *Personality disorder* 0.19. The confusion matrix indicates that minority categories are frequently mapped to *Normal* or *Depression*, a pattern typical for MultinomialNB on sparse bag-of-words features under class imbalance (Pang & Lee, 2008; Ravi & Ravi, 2015). While MNB is a strong lexical baseline for text (Pang & Lee, 2008; Ravi & Ravi, 2015), improving minority-class performance typically requires:

(i) Tuning additive smoothing α or using ComplementNB in imbalanced settings (Pedregosa et al., 2011);

(ii) Enriching features beyond unigrams (e.g., character n-grams, bi/tri-grams, sublinear TF scaling) to capture morphology and misspellings (Pang & Lee, 2008);

(iii) Adjusting class priors or applying imbalance-aware sampling; and

(iv) Calibrating scores and, where appropriate, class-specific decision thresholds to trade precision/recall on critical classes.

In general, MNB remains a transparent and scalable reference for TF–IDF, while the error profile suggests feature enrichment and explicit imbalance handling for clinically important minority labels (Calvo et al., 2017; Ravi & Ravi, 2015).

6.4 K–Means Clustering on TF–IDF (Unsupervised Analysis)

Unsupervised clustering provides a label-free view of structure in the corpus, useful for:

(i) Sanity-checking separability of sentiments,

(ii) Surfacing latent regimes (e.g., high-distress vs. neutral), and

(iii) Initializing semi-supervised pipelines.

With sparse lexical features (TF–IDF), K-Means offers a fast, scalable baseline with clear reporting (inertia/WCSS and silhouette) (Pedregosa, Varoquaux, et al., 2020).

K-Means partitions n vectors into k clusters by minimizing within-cluster sum of squares:

$$\min_{\{\mathcal{C}_j,\mu_j\}_{j=1}^{k}} \sum_{j=1}^{k} \sum_{x_i \in \mathcal{C}_j} \|x_i - \mu_j\|^2. \tag{13}$$

Because inertia decreases monotonically in k, model selection uses (i) the elbow in the WCSS–k curve and (ii) the silhouette, which balances compactness and separation (Pedregosa, Varoquaux, et al., 2020; Rousseeuw, 1987). The elbow method is a widely used heuristic for selecting in K-Means clustering (Umarogono et al., 2019), and in our text-based experiments we combine it with silhouette analysis to balance compactness and separation.

Implementation (TF–IDF, elbow/silhouette, and 2D projection)

As implemented in Listing 7, we compute TF–IDF, sweep k to record inertia and cosine-based silhouette (Figure 21), select $k^\star$ by the silhouette maximum, then fit K-Means and visualize clusters in 2D using TruncatedSVD on the sparse matrix (Figure 22).

The elbow curve shows diminishing WCSS returns, and the cosine-based silhouette attains its maximum at k=2 (Figure 21), with values near or below zero thereafter, indicating weak separation. We therefore adopt $k^\star$=2. The 2D TruncatedSVD projection exhibits a coarse bipartition consistent with this choice (Figure 22). Figure 22 visualizes the clustering result in PCA-reduced space, showing a clear bipartition consistent with the optimal $k^\star = 2$ selected via silhouette analysis. This supports the theoretical link between PCA and K-Means clustering, as established by Ding & He (C. Ding & He, 2004). Given the modest silhouettes typical for high-dimensional sparse text, we treat clusters as exploratory structure for seeding semi-supervised workflows, guiding annotation, or motivating feature refinement (e.g., adding character n-grams).

Practical Notes

- *Distance:* Utilizing cosine for silhouette on TF–IDF; values near/under 0 indicate weak structure.
- *Scale:* Preferring `MiniBatchKMeans` (set `n_init=10`, `batch_size`, `max_iter`); fix `random_state`.
- *Dimensionality:* Using *TruncatedSVD* for visualization and/or pre-reduce to

Listing 7: MiniBatchKMeans on TF–IDF with elbow/silhouette (cosine) and TruncatedSVD visualization

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.cluster import MiniBatchKMeans
from sklearn.decomposition import TruncatedSVD
from sklearn.metrics import silhouette_score

tfidf_all = TfidfVectorizer(max_features=5000)
X_all_tfidf = tfidf_all.fit_transform(data['statement'].astype(str))

ks = range(2, 11)
inertia, sil = [], []
for k in ks:
    km = MiniBatchKMeans(
        n_clusters=k, random_state=42, n_init=10,
        batch_size=2048, max_iter=100
    )
    labels_k = km.fit_predict(X_all_tfidf)
    inertia.append(km.inertia_)

    if np.unique(labels_k).size < 2:
        sil.append(np.nan)
        continue
    try:
        s = silhouette_score(
            X_all_tfidf, labels_k, metric='cosine',
            sample_size=min(2000, X_all_tfidf.shape[0]),
            random_state=42
        )
    except Exception:
        s = np.nan
    sil.append(s)

fig, ax = plt.subplots(1, 2, figsize=(10, 4))
ax[0].plot(list(ks), inertia, 'bo-')
ax[0].set_title('Elbow (WCSS)'); ax[0].set_xlabel('k'); ax[0].set_ylabel('WCSS')
ax[1].plot(list(ks), sil, 'go-')
ax[1].set_title('Silhouette (cosine)'); ax[1].set_xlabel('k'); ax[1].set_ylabel('Score')
plt.tight_layout(); plt.show()

k_star = 3 if np.all(np.isnan(np.asarray(sil, float))) else list(ks)[int(np.nanargmax(sil))]
print(f"Chosen k: {k_star}")

km = MiniBatchKMeans(n_clusters=k_star, random_state=42, n_init=10,
                     batch_size=2048, max_iter=100)
labels_km = km.fit_predict(X_all_tfidf)

svd = TruncatedSVD(n_components=2, random_state=42)
Z = svd.fit_transform(X_all_tfidf)

plt.figure(figsize=(6,5))
plt.scatter(Z[:,0], Z[:,1], c=labels_km, s=8, cmap='viridis')
plt.xlabel('SVD1'); plt.ylabel('SVD2'); plt.title(f'K-Means (k={k_star}) in SVD space')
plt.colorbar(label='Cluster'); plt.tight_layout(); plt.show()
```

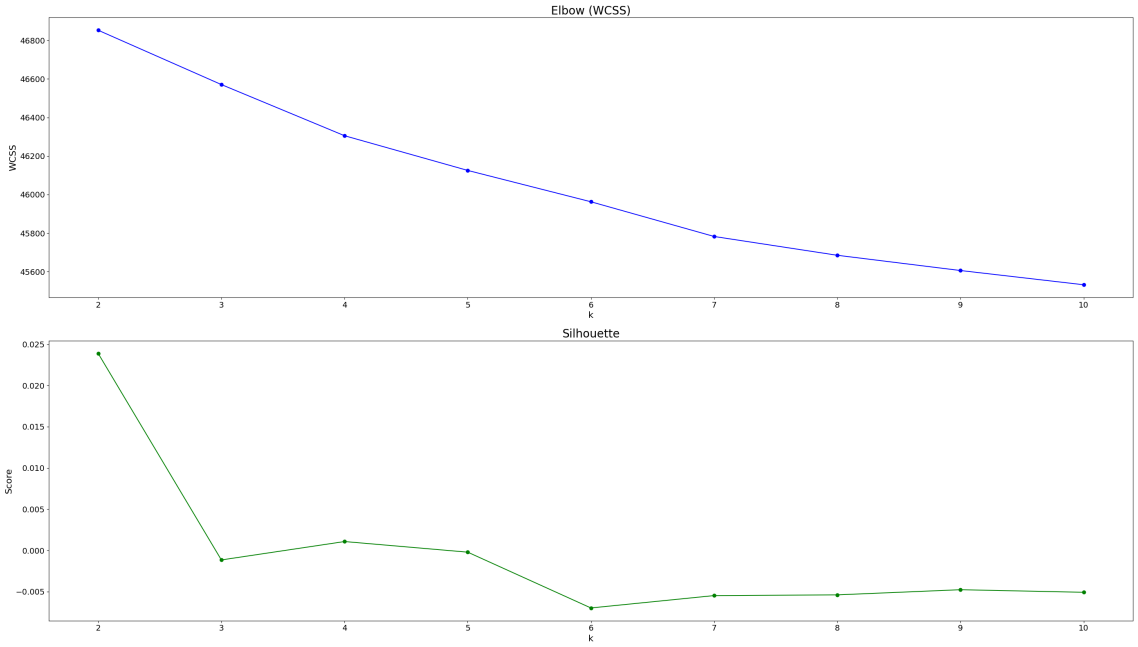

Figure 21: Elbow (WCSS) and silhouette (cosine) across k for TF–IDF features

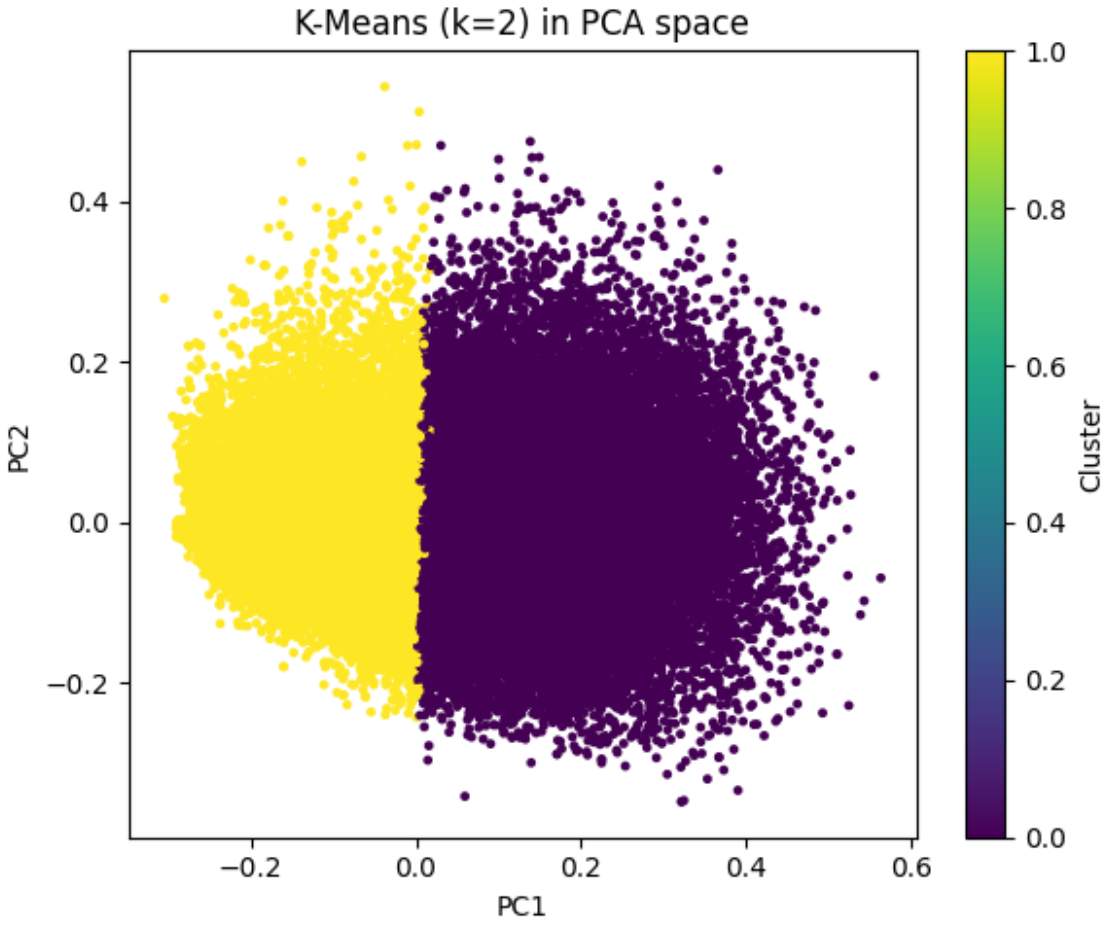

Figure 22: K–Means clustering (k=2) visualized in the first two TruncatedSVD components

50–300 dims to stabilize silhouettes; avoid `.toarray()`.

- *Cost controls:* Sample in `silhouette_score` and guard for single-label failures.
- *Interpretation:* Clusters are exploratory (not clinical labels); use to seed semi-/supervised workflows or guide annotation.
- *Reproducibility/quality:* Log seeds and versions; sanity-check stability across seeds and report a second index (e.g., Calinski–Harabasz or Davies–Bouldin).

We now turn to supervised baselines to quantify separability with labels, starting with Decision Trees (subsection 6.5).

6.5 Decision Tree

Decision trees (DTs) are attractive for mental health sentiment analysis because they produce explicit, human-readable rules that can be audited and natively handle sparse, high-dimensional features such as TF–IDF. Moreover, they require little preprocessing and they train and infer rapidly (Breiman et al., 1984; Pedregosa et al., 2011; Quinlan, 1993). Their transparency is useful in safety-critical contexts (e.g., classification) where one must explain which lexical cues led to a prediction.

A DT partitions the feature space recursively. At each node t, the algorithm chooses the split in characteristics x_j and threshold τ that maximally reduces class impurity:

$$\Delta I(t, j, \tau) = I(t) - \frac{n_L}{n_t} I(t_L) - \frac{n_R}{n_t} I(t_R). \tag{14}$$

Here, n_t is the number of examples at node t, and t_L, t_R are the left/right children.

Common impurity measures are:

$$\text{Gini}(t) = \sum_{k=1}^{K} p_k(1 - p_k), \tag{15}$$

$$\text{Entropy}(t) = -\sum_{k=1}^{K} p_k \log p_k. \tag{16}$$

with p_k the class proportion at node t. Greedy, top-down induction repeats until a stopping rule is met (e.g., no gain, maximum depth, minimum leaf size). To control variance, one uses structural constraints (e.g., `max_depth`, `min_samples_leaf`) or cost-complexity pruning that minimizes empirical loss plus a size penalty (Breiman et al., 1984; Pedregosa et al., 2011). For imbalanced labels, class-weighted impurities up-weight minority classes during splitting.

With TF–IDF or character n-grams, DTs learn concise lexical rules (e.g., presence of suicide-related terms) that are easy to inspect. Limitations are well known: a single tree can overfit sparse text and may form brittle thresholds on rare features; minority classes can suffer if not class-weighted. In practice, DTs offer a transparent baseline; ensembles (Random Forest/Boosting) usually improve accuracy but reduce the interpretability of a single rule (Breiman et al., 1984; Hastie et al., 2009). Post-hoc explanation methods (e.g., SHAP) can quantify feature contributions per prediction (Lundberg & Lee, 2017).

We train a DT on TF–IDF features in a leakage-safe pipeline and account for class imbalance with inverse-frequency weights as shown in Listing 8.

Listing 8: TF–IDF + Decision Tree: training, evaluation, and export

```
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.tree import DecisionTreeClassifier
from joblib import dump
import os

dt_pipe = Pipeline([
    ('tfidf', TfidfVectorizer(max_features=5000)),
    ('dt', DecisionTreeClassifier(
        random_state=42,
        class_weight='balanced',
        max_depth=None,
        min_samples_split=2,
        min_samples_leaf=1
    ))
])

dt_pipe.fit(X_train_text, y_train)
y_pred = dt_pipe.predict(X_test_text)
print("TF-IDF + DecisionTree")
evaluate_preds(y_test, y_pred, le)

dump(dt_pipe.named_steps['dt'],    os.path.join(output_dir, "
    tfidf_decisiontree.pkl"))
dump(dt_pipe.named_steps['tfidf'], os.path.join(output_dir, "
    tfidf_vectorizer_dt.pkl"))
```

Class	Precision	Recall	F1-score	Support
Anxiety	0.59	0.59	0.59	768
Bipolar	0.60	0.62	0.61	556
Depression	0.59	0.56	0.58	3081
Normal	0.84	0.86	0.85	3269
Personality disorder	0.49	0.55	0.52	215
Stress	0.38	0.41	0.40	517
Suicidal	0.51	0.52	0.51	2131
Accuracy			0.64	10537
Macro avg	0.57	0.59	0.58	10537
Weighted avg	0.64	0.64	0.64	10537

Table 6: TF–IDF + Decision Tree classification report

Findings in Decision Tree

- According to Table 6, the comprehensive accuracy 0.64; macro–F1 0.58.
- Strongest on *Normal* (F1 0.85); moderate on *Anxiety*/*Bipolar* ($\approx$ 0.59–0.61); weaker on *Stress* (F1 0.40) and *Suicidal* (F1 0.51).
- Confusions (Figure 23): *Depression* $\leftrightarrow$ *Suicidal*; *Stress* $\rightarrow$ *Normal*, reflecting overlapping lexical markers and label imbalance.
- Auditability: Export root-to-leaf paths as human-readable rules.

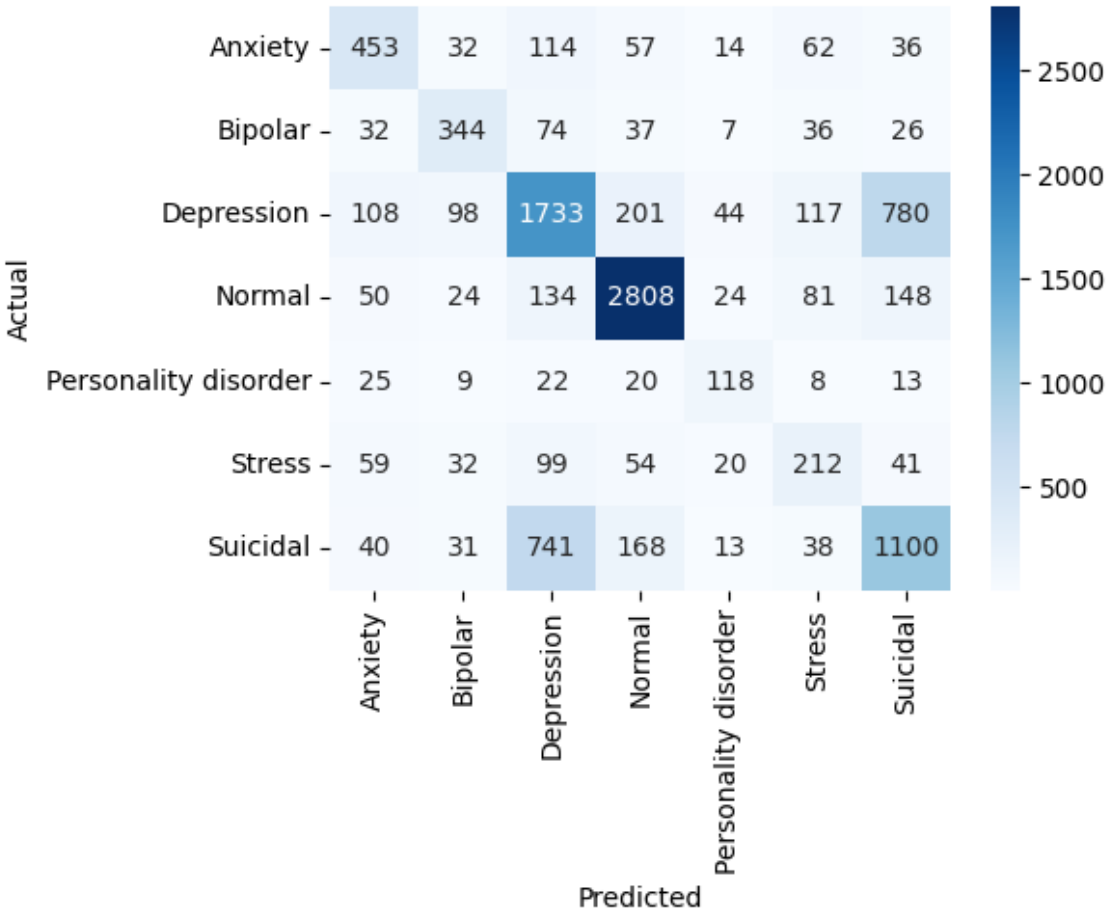

Figure 23: Confusion matrix for TF–IDF + Decision Tree

- For risk-sensitive labels (e.g., *Suicidal*): raise recall via thresholding and cost-sensitive objectives.
- If higher accuracy is required: move to ensembles (Random Forest, Gradient Boosting) or richer representations (e.g., sentence embeddings) while retaining rule summaries for governance.

Practical recommendations:

- Utilizing `class_weight='balanced'` and tune `max_depth` and `min_samples_leaf` to curb overfitting.
- Adding character n-grams in TF–IDF to capture misspellings and morphology.
- Considering post-pruning (`ccp_alpha`) and cross-validated depth to stabilize rules.
- Complementing tree rules with SHAP analyses for per-prediction attributions (Lundberg & Lee, 2017).

6.5.1 Model Interpretability with SHAP

In plain terms, SHAP provides a straightforward narrative that starts with an average base score for a single prediction and shows how the selection of each input feature (such as a feature or word) affects it. Negative nudges oppose the class you defined, while positive nudges support it. Each nudge adds up to the model's exact score for that instance (Lundberg & Lee, 2017).

Formally, SHAP (SHapley Additive exPlanations) assigns to each feature i a contribution ϕ_i derived from Shapley values in cooperative game theory:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[f_{S \cup \{i\}}(x) - f_S(x) \right] \tag{17}$$

so ϕ_i is the average marginal improvement when adding i to every subset S of features. The values satisfy three desirable properties: (i) local additivity $f(x) = \phi_0 + \sum_i \phi_i$, (ii) consistency (if a model relies more on a feature, its ϕ_i does not go down), and (iii) missingness ($\phi_i = 0$ if a feature is unused). Efficient variants exist for trees (TreeSHAP), linear models, and deep nets (Lundberg & Lee, 2017).

How to read SHAP based on the dataset:

- **Baseline ϕ_0:** the average score on a small, stratified background sample from the development data.
- **Sign:** $+\phi_i$ pushes toward the explained class (e.g., Suicidal); $-\phi_i$ pushes away.
- **Magnitude:** $|\phi_i|$ indicates influence strength on that decision.
- **Views:** per-case (why this post was labeled X) and global (aggregate $|\phi_i|$ across many posts to see stable, important cues).

Why use SHAP for mental health sentiment analysis.

- *Case transparency:* reviewers can see which words raised risk and which were protective (e.g., help-seeking language).

- *Model governance:* supports threshold setting for high-risk classes and documentation of model behavior.
- *Model-agnostic:* same explanation interface for LR, RF, CNN/LSTM.

Caveats and good practice:

SHAP explanations are local, relative, and assumption dependent (Lundberg & Lee, 2017). In this thesis we apply the following guardrails.

- *Feature dependence:* When cues co-occur (e.g., "hopeless" with "worthless"), credit is shared across them; TreeSHAP is exact for tree models, while linear/deep explainers approximate conditional effects. Group tokens into meaningful phrases to reduce credit fragmentation (Lundberg & Lee, 2017).
- *Background (baseline) choice:* SHAP compares a post to a background distribution. Different backgrounds shift the baseline ϕ_0 and redistribute ϕ_i. We use a small, stratified sample of the development set and report this choice to keep explanations comparable (Lundberg & Lee, 2017).
- *Target being explained:* Multi-class models require choosing a specific output. We explain the class logit (not post-softmax probability) to avoid probability coupling across classes; signs are read "toward/away from this class."
- *Stability checks:* Tokenization or preprocessing changes can alter attributions. We fix seeds, tokenizer, and vectorizer versions, and we verify that explanations change when the model is randomized (sanity checks) (Adebayo et al., 2018).
- Robustness to gaming Explanations can be manipulated by adversarial preprocessing or spurious shortcuts; we cross-validate attributions on held-out data and avoid using explanations alone for model selection (Slack et al., 2020).
- *Class imbalance and rarity:* Rare but high-risk terms can dominate local ϕ_i. We interpret SHAP alongside calibrated precision–recall and per-class error analysis.
- *Privacy and ethics:* Personally Identifiable Information (PII) is hidden from view, and gathering plots are preferred over raw text where practical; explanations aid in audit and triage but are not clinical recommendations.

Qualitative SHAP-style Explanations

The following anonymized excerpts (with a content warning for self-harm mentions) illustrate the direction of influential cues without reporting numeric values (Lundberg & Lee, 2017; Sarkar, 2024). "Direction" denotes a plausible push for the class: + toward, − away.

Ex. 1 (label: *Suicidal*).

"Why continue living if life is only going to deteriorate."

Cue	Dir.	Rationale
"why continue living"	+	Direct challenge to continuing life; strong intent cue.
"deteriorate"	+	Catastrophic outlook amplifies risk language.
Neutral punctuation	0	No clear directional effect.

Ex. 2 (boundary: *Depression* vs. *Suicidal*).

"Perpetually alone. It's no one's fault but my own."

Cue	Dir.	Rationale
"perpetually alone"	+ (*Depression*)	Persistent loneliness/anhedonia signals low mood.
"fault ... my own"	+ (*Depression*)	Guilt/self-blame typical in depressive language.
No explicit self-harm phrase	− (*Suicidal*)	Lacks intent terms; reduces *Suicidal*.

Ex. 3 (label: *Depression*).

"Stuck in a loop. Everything is the same."

Cue	Dir.	Rationale
"stuck in a loop"	+	Rumination/learned helplessness framing.
"everything is the same"	+	Flat affect, loss of novelty/interest.
No help-seeking cue	−	Missing protective language lowers confidence.

Ex. 4 (boundary: *Depression* with protective cue).

"I have a goal to reach, but I am too broken."

Cue	Dir.	Rationale
"too broken"	+ (*Depression*)	Strong self-devaluation.
"goal to reach"	− (*Depression*)	Future orientation/help-seeking is protective.
Contrastive "but"	mixed	Competing signals remain salient.

Ex. 5 (label: *Suicidal*).

"Well, you will die anyway someday, why . . . "

Cue	Dir.	Rationale
"will die anyway"	+	Normalization of death; risk-raising cue.
Rhetorical "why"	+	Challenges reasons to continue; intent-adjacent.
Hesitation "..."	0	Minimal directional effect.

Ex. 6 (label: *Anxiety*).

"Head injury worries? I constantly worry about head ..."

Cue	Dir.	Rationale
"constantly worry"	+	Persistent, hard-to-control worry characteristic of anxiety.
"head injury worries?"	+	Health-anxiety focus on feared harm.
Question framing "?"	0	Punctuation itself is neutral.

Ex. 7 (label: *Depression*).

"Just lost my job today. It has not been a great week."

Cue	Dir.	Rationale
"lost my job"	+	Salient negative life event linked to low mood.
"not been a great week"	+	Global negative evaluation/low affect.
No explicit help-seeking cue	−	Absence of protective cue lowers confidence.

Ex. 8 (label: *Stress*).

"He turns it on first thing in the morning and turns it on ..."

Cue	Dir.	Rationale
"first thing in the morning"	+	Early, automatic engagement suggests heightened arousal.
"turns it on ..." (repeated exposure)	+	Ongoing exposure to a stressor (e.g., news/social media).
No relaxation/offset cue	−	Lack of counter-regulation signal.

These tables mirror SHAP's per-feature "push" explanation without computing numeric values. If SHAP is later computed, replace the Dir. column with signed contributions ϕ_i and include the baseline note to recover the exact score.

6.5.2 Support Vector Machines

Support Vector Machines (SVMs) are supervised learners for classification and regression that estimate a decision boundary maximizing the margin between classes. With soft-margin optimization, misclassifications are controlled by a regularization

constant, and non-linear relations are modeled via kernel functions that implicitly map inputs to higher-dimensional feature spaces (Cortes & Vapnik, 1995). In contrast to probabilistic approaches such as logistic regression, which directly estimate class membership probabilities, SVMs rely on kernelized large-margin separation to handle both linear and non-linear structure (Z. Ding et al., 2025).

Given the high dimensionality and sparsity of text, we evaluated both linear SVMs and non-linear SVMs with a radial basis function (RBF) kernel. Model selection was based on the weighted F1-score to account for class imbalance. Hyperparameters, including the regularization strength C, class weighting, and the RBF shape parameter γ-were tuned via grid search (Z. Ding et al., 2025). Final models were implemented with scikit-learn's Support Vector Classifier (`SVC`) and `LinearSVC` (Pedregosa et al., 2011). For multi-class prediction, the default one-vs-one (OvO) scheme was adopted: a binary classifier is trained for each class pair, and the final label is determined by majority voting across pairwise decisions (Z. Ding et al., 2025).

Implementation of SVM for Mental Health Detection on Social Media

We trained a linear Support Vector Machine using a pipeline with TF–IDF features and `LinearSVC` to classify mental health related posts. Texts were cast to strings, transformed with `TfidfVectorizer` capped at 5,000 features, and fed to `LinearSVC` with default settings (C=1.0, L_2 penalty, squared-hinge loss, one-vs-rest reduction). The model was fit on the training split and evaluated on the held-out test set using precision, recall, F1 (per class, macro, weighted), and accuracy via `classification_report`. This configuration targets sparse, high-dimensional text and offers strong performance with efficient training.

Class	Precision	Recall	F1-score	Support
Anxiety	0.82	0.77	0.80	768
Bipolar	0.84	0.76	0.80	556
Depression	0.72	0.72	0.72	3081
Normal	0.88	0.96	0.91	3269
Personality disorder	0.81	0.59	0.68	215
Stress	0.69	0.48	0.56	517
Suicidal	0.69	0.69	0.69	2131
Accuracy			0.78	10537
Macro avg	0.78	0.71	0.74	10537
Weighted avg	0.77	0.78	0.77	10537

Table 7: TF–IDF + LinearSVC classification report (test set).

Based on Table 7, the overall accuracy was 0.78, with macro and weighted F1 of 0.74 and 0.77. Per-class F1 was highest for *Normal* (0.91) and strong for *Anxiety* and *Bipolar* (both 0.80). *Depression* reached 0.72, *Suicidal* 0.69, *Personality disorder* 0.68, and *Stress* was lowest (0.56). The confusion matrix shows concentrated errors for *Depression* $\leftrightarrow$ *Suicidal*, *Stress* $\rightarrow$ *Normal*/*Depression*, and *Personality disorder* $\rightarrow$ *Normal*, consistent with imbalance and overlapping lexical cues.

As shown in Figure 24, rows are actual classes and columns are predicted counts (darker = more). A strong diagonal is evident—especially for *Normal*—with good

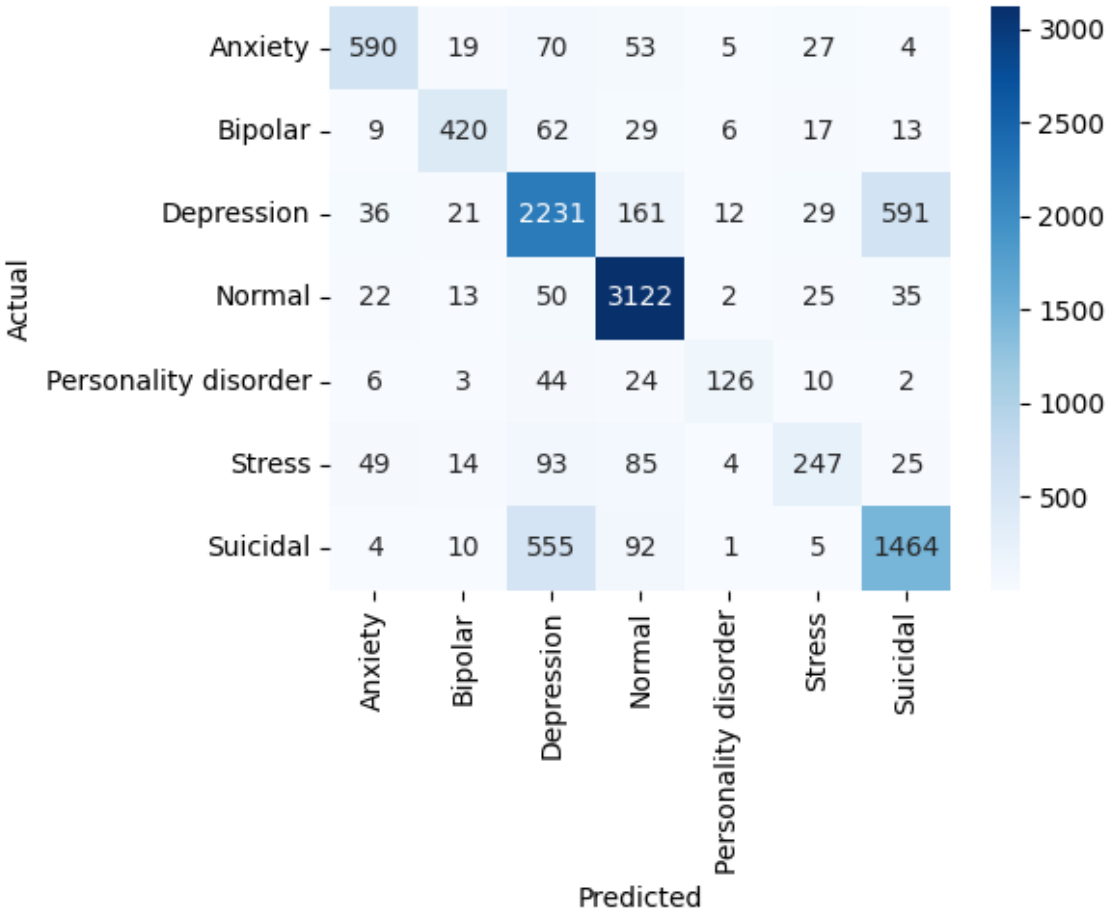

Figure 24: SVM Confusion Matrix

separation for *Anxiety* and *Bipolar*. The largest off-diagonal errors occur for *Depression* $\leftrightarrow$ *Suicidal*, and for *Stress* and *Personality disorder* mapping to *Normal*/*Depression*, consistent with class imbalance and overlapping vocabulary. Per-class recalls from Table 7 are: Normal 96%, Anxiety 77%, Bipolar 76%, Depression 72%, Suicidal 69%, Personality disorder 59%, Stress 48%. These patterns suggest that additional imbalance handling (e.g., resampling or class-specific thresholds), character n-grams, and probability calibration could further reduce the dominant off-diagonal confusions.

Practical Considerations and Next Steps

- Tune hyperparameters (C, class weights) and consider `class_weight=balanced` or resampling to mitigate imbalance.
- Enrich features with character n-grams and domain lexicons; explore bigram/trigram n_gram ranges.
- For potential non-linear boundaries, evaluate `SVC` with RBF kernel on a reduced feature space, or apply dimensionality reduction (e.g., PCA on dense embeddings).
- Perform targeted error analysis on classes with lower F1 to refine labels and features.

This setup demonstrates that a linear SVM paired with TF–IDF is an effective baseline for mental health related conversation classification, while offering clear avenues for further improvement.

6.6 Convolutional Neural Networks (Conv1D) for Mental Health Sentiment Analysis

Convolutional neural networks (CNNs) for text slide small filters over token embeddings to detect local n-gram patterns (e.g., "feel so hopeless"), then aggregate the strongest signals via pooling.This produces position-tolerant short phrase detectors, is parameter-efficient, and trains fast, making it optimal for short, noisy, and highly lexical social media posts (Bucur et al., 2024; Minaee et al., 2021; Shen et al., 2022).

Let $E \in \mathbb{R}^{L\times d}$ be the embedding matrix for a post of length L (padded/truncated) with embedding size d. A width-w filter $F \in \mathbb{R}^{w\times d}$ produces a feature map

$$c_i = \sigma\Big(\langle F,\, E_{i:i+w-1}\rangle + b\Big), \qquad i = 1, \ldots, L - w + 1. \tag{18}$$

followed by global max pooling $z = \max_i c_i$ to capture the strongest activation of pattern F. Stacking multiple filters (varying w) gives a feature vector z fed to dense layers and a softmax over classes. Dropout regularizes the dense layers.

As seen in Listing 9, we tokenize to a fixed vocabulary, pad sequences to length 100, and train a 1D-CNN with Rectified Linear Unit (ReLU), global-max pooling, dropout, and a softmax head. The configuration mirrors widely used text-CNN baselines and keeps the pipeline lightweight and reproducible.

Listing 9: Conv1D text CNN for mental-health sentiment classification

```
# CNN (Conv1D) classifier on tokenized sequences
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Embedding, Conv1D,
    GlobalMaxPooling1D, Dense, Dropout
from tensorflow.keras.optimizers import Adam
import numpy as np

max_words, max_len = 10000, 100

# (Re)create tokenizer/padded sequences if not present
try:
    tok, Xtr_pad, Xte_pad
except NameError:
    tok = Tokenizer(num_words=max_words, oov_token='<OOV>')
    tok.fit_on_texts(X_train_text)
    Xtr_seq = tok.texts_to_sequences(X_train_text)
    Xte_seq = tok.texts_to_sequences(X_test_text)
    Xtr_pad = pad_sequences(Xtr_seq, maxlen=max_len, padding='post'
        )
    Xte_pad = pad_sequences(Xte_seq, maxlen=max_len, padding='post'
        )

cnn = Sequential([
    Embedding(input_dim=max_words, output_dim=128, input_length=
        max_len),
    Conv1D(filters=128, kernel_size=5, activation='relu'),
    GlobalMaxPooling1D(),
    Dropout(0.5),
    Dense(64, activation='relu'),
    Dense(len(label_names), activation='softmax')
])
cnn.compile(loss='sparse_categorical_crossentropy',
            optimizer=Adam(1e-3),
            metrics=['accuracy'])

hist_cnn = cnn.fit(Xtr_pad, y_train, epochs=5, batch_size=32,
                   validation_split=0.2, verbose=1)

y_pred_cnn = np.argmax(cnn.predict(Xte_pad), axis=-1)
print("CNN (Conv1D)")
evaluate_preds(y_test, y_pred_cnn, le)
```

As shown in Table 8, the general accuracy is 0.76 (macro-F1 0.68; weighted-F1 0.76). The model performs very well on *Normal* (F1 0.92) and achieves balanced scores on *Depression* (F1 0.72) and *Suicidal* (F1 0.68). *Bipolar* reaches F1 0.71, while *Personality disorder* remains challenging (F1 0.47)—consistent with minority support and lexical heterogeneity. The confusion matrix (Figure 25) shows residual confusion between *Depression* and *Suicidal*, typical when local phrase patterns overlap. Compared with linear TF–IDF baselines, the CNN benefits from position-tolerant phrase detectors and learns robust cues without manual n-gram engineering; compared with heavy transformer models, it remains fast, light-weight, and easy to tune.

Class	Precision	Recall	F1-score	Support
Anxiety	0.78	0.70	0.74	768
Bipolar	0.69	0.74	0.71	556
Depression	0.71	0.73	0.72	3081
Normal	0.93	0.92	0.92	3269
Personality disorder	0.56	0.41	0.47	215
Stress	0.55	0.54	0.55	517
Suicidal	0.67	0.68	0.68	2131
Accuracy			0.76	10537
Macro avg	0.70	0.67	0.68	10537
Weighted avg	0.76	0.76	0.76	10537

Table 8: Conv1D CNN classification report (test set).

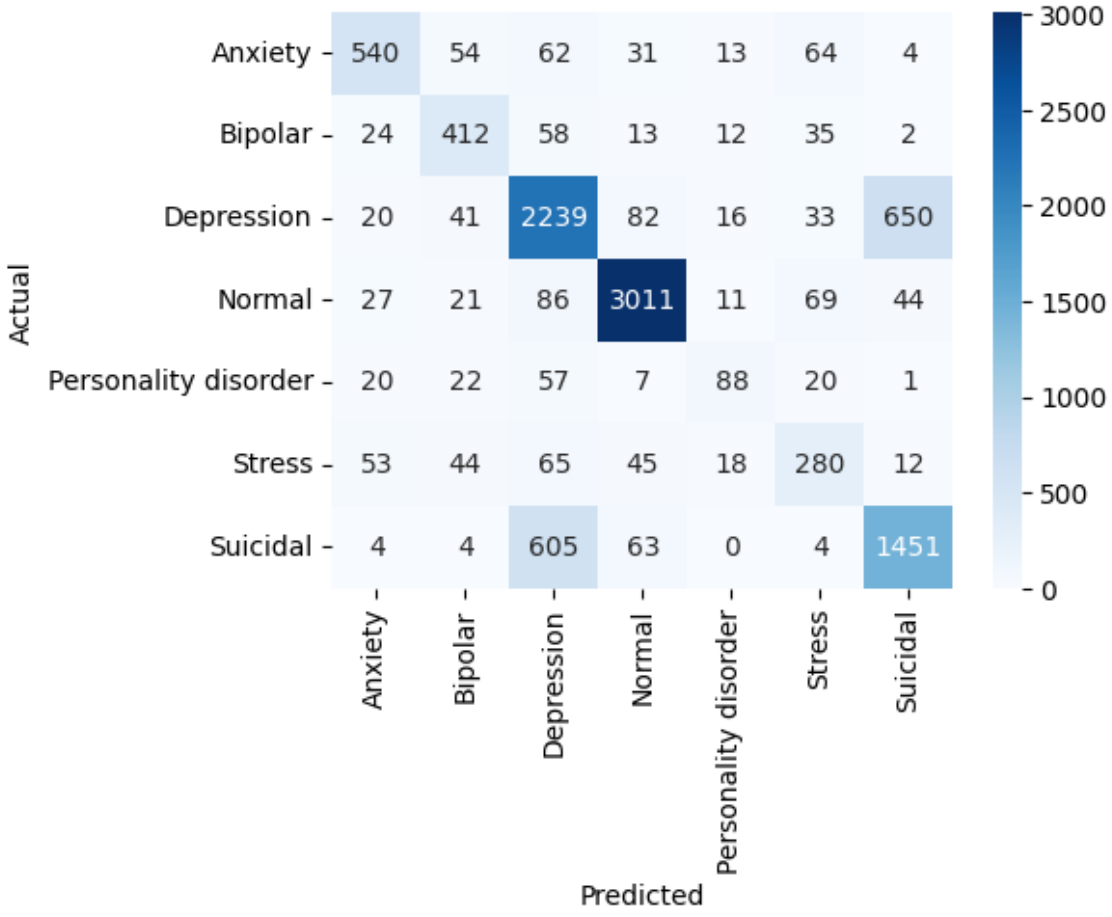

Figure 25: Confusion matrix for the Conv1D CNN

Why CNNs are a good fit:

- **Local phrase detectors**: Short danger cues (negation, self-harm phrases) are detected by filters, which are translation-tolerant by pooling.
- **Efficiency**: Large social media corpora and iterative analysis are well suited for small footprints and fast training/inference.
- **Regularization**: On noisy text, dropout and max-pooling diminish overfitting.
- **Extensibility**: can incorporate subword/character channels to handle misspellings; can be stacked with BiLSTM/transformers if higher capacity is needed.

Recommendations:

- Tuning filter sizes (e.g., kernel sizes 3/4/5 in parallel) and the number of filters; add character-level Conv1D for robustness to misspellings.
- Utilizing class weights or focal loss to raise recall on minority classes; add early stopping on macro-F1.
- Initializing embeddings with task relevant pretrained vectors (e.g., domain GloVe/fastText) and fine-tune.
- For interpretability, reporting saliency/Grad-CAM over tokens to highlight phrases influencing predictions.

6.7 Bidirectional Encoder Representations from Transformers (BERT)

BERT is a transformer-based language model for natural language processing tasks. In the mental health context, BERT can analyze and interpret text (e.g., social media posts or clinical notes) to identify and classify mental health conditions. By modeling bidirectional context and nuanced language, BERT supports early identification and intervention (Devlin et al., 2019).

Our BERT configuration follows common practice; we also reviewed public Kaggle implementations (Saha, 2024).

Analyzing sentiment analysis related to mental health using BERT, a deep learning model, involves several key steps. The data is split into training and validation sets and tokenized using BERT's tokenizer. A custom dataset class is defined for handling data batches. The BERT model is initialized for sequence classification, and training is conducted over 10 epochs using the AdamW optimizer. The training loop updates model weights based on the loss calculated from predictions versus actual labels.

After training, the model is evaluated using accuracy, precision, recall, and F1-score metrics, and a confusion matrix is plotted to visualize prediction performance across different mental health categories. The final accuracy achieved is approximately 0.83 as shown in Table 9.

As shown in Figure 26, rows are actual classes and columns are predicted classes; cells show counts (darker = more). The matrix has a strong diagonal, indicating good discrimination overall. Largest confusions occur between *Depression* and

Table 9: BERT classification report

Category	Precision	Recall	F1-Score	Support
Normal	0.95	0.96	0.96	3269
Depression	0.79	0.75	0.77	3081
Suicidal	0.69	0.75	0.72	2131
Anxiety	0.91	0.84	0.88	768
Stress	0.70	0.80	0.75	517
Bipolar	0.91	0.86	0.88	556
Personality disorder	0.69	0.76	0.72	215
Accuracy		0.83		10537
Macro avg	0.81	0.82	0.81	10537
Weighted avg	0.83	0.83	0.83	10537

Suicidal (Dep→Sui: 667; Sui→Dep: 497). Per-class recalls: Normal 3139/3269 (96%), Depression 2300/3081 (75%), Suicidal 1588/2131 (75%), Anxiety 648/768 (84%), Stress 413/517 (80%), Bipolar 476/556 (86%), Personality disorder 163/215 (76%). Errors for minority labels mainly map to Normal/Depression, suggesting class-imbalance effects; targeted rebalancing and threshold tuning may reduce these off-diagonal confusions.

This chapter compared classical and neural approaches under a unified, risk-aware protocol. Key takeaways:

- Performance pattern: BERT yielded the strongest overall results; among non-transformers, linear TF–IDF margins (SVM, LR) were consistently competitive; CNN/LSTM were solid on short texts; tree-based and Naive Bayes baselines trailed but remained interpretable.
- *Imbalance handling:* Stratified splits, `class_weight='balanced'`, and class-specific thresholds improved minority/risk-class recall (e.g., *Suicidal*); calibrated probabilities aided operating-point selection.
- *Feature/representation effects:* TF–IDF remained a strong lexical baseline; sentence-transformer embeddings improved context sensitivity; averaged Word2Vec under-performed due to loss of word order.
- *Interpretability vs. accuracy:* Linear models and trees support inspection and governance; higher-capacity models trade some transparency for gains on overlapping classes.
- *Reproducibility:* Fixed seeds, shared preprocessing, and consistent metrics (accuracy, macro/weighted F1, PR curves) enable fair comparison.

The limitations observed—lexical overlap between adjacent clinical labels, sparse minority classes, and the need for calibrated, auditable decisions—motivate evaluating *Large Language Models* that capture richer semantics, support few-shot adaptation, and can generate rationales. The next chapter positions LLMs relative to our supervised baselines and covers prompting, parameter-efficient tuning, retrieval augmentation, calibration/uncertainty, and the safety, privacy, and governance requirements for mental health applications.

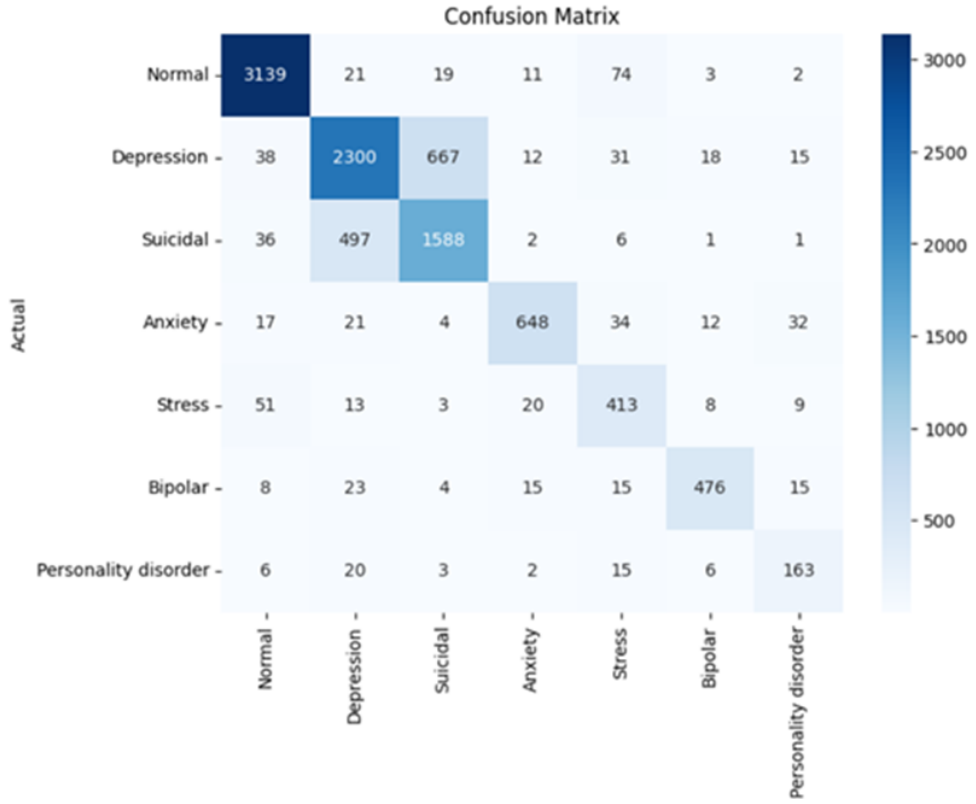

Figure 26: BERT Confusion Matrix

7 Large Language Models

Large language models (LLMs) such as GPT3.5, GPT4o, Llama2, and Llama3 demonstrate strong capabilities in natural language understanding and generation (K. Chen et al., 2024). Trained on web-scale corpora, they can be adapted to downstream tasks including sentiment analysis and text classification, making them promising tools for automated psychological stress detection (Ge et al., 2025). In this domain, LLMs offer distinctive advantages: They can generate rationales that support predictions that are just partially understandable and generalize to new contexts with limited supervision (Ge et al., 2025; X. Wang et al., 2024). This chapter investigates the advantages LLMs offer, the challenges they face, and their potential impact on mental health interventions.

7.1 Introduction to LLMs

LLMs possess generative properties that allow them to provide explanations for their predictions, enhancing the transparency of the decision-making process. This semi-interpretability is a significant advancement, building trust and improving the reliability of stress detection models (X. Wang et al., 2024). For instance, LLMs can analyze user generated content on social media platforms to identify emotional distress. By examining the language used in tweets or posts, LLMs can detect signs of depression, such as frequent use of negative affective vocabulary (e.g., “despair”, “pain”), self defeating comments, and discussions of self injury (Ge et al., 2025).

Large language models are effective at processing large-scale, heterogeneous text from platforms such as Twitter, Reddit, blogs, and forums. Through self-attention,

they capture long-range dependencies and nuanced semantics that underlie complex judgments in text categorization, sentiment analysis, and psychologically relevant assessments (Vaswani et al., 2017). Long-context and memory-augmented approaches enable LLMs to retain and reuse temporally dispersed cues across multi-session dialogues and can support analyses of longitudinal language change within domain-specific monitoring pipelines (Bucur et al., 2024; Tavakoli et al., 2025).

For governance and procurement, we refer to the FAITA–Mental Health rubric (subsection 7.2) when discussing deployment of LLM-based tools.

7.2 Evaluating AI Mental Health Tools

LLM tools for mental health are expanding quickly, but their safety and quality are uneven. To bring consistency, Golden and Aboujaoude (2024) propose FAITA–Mental Health, a practical rubric for today's AI landscape. It adapts pre-LLM app frameworks and rates six areas (total score 0–24). In brief: *credibility* asks whether a tool has a clear goal, uses validated methods, and keeps people engaged; *user experience* looks at how well it adapts to the person, the naturalness of exchanges, and whether there are easy ways to give feedback or get help; *user agency* covers control of personal data and features that build self-management rather than dependency; *equity and inclusion* examine cultural fit and active bias mitigation; *transparency* requires plain-language disclosure of ownership, funding, development, and governance; and *crisis management* checks whether the tool recognizes emergencies and routes users to appropriate local resources with follow through.

In practice, FAITA can serve as a buying or research checklist alongside version/change logs, hallucination and guardrail tests, and privacy/security reviews. As the authors note, the scale still needs formal validation (e.g., inter-rater reliability and links to clinical or safety outcomes), so it should complement rather than replacing risk management, bias audits, and human oversight (Golden & Aboujaoude, 2024).

This systems view has direct implications for digital and AI tools used in mental health. Tools should be co-designed with people with lived experience, advance access to community based care, embed human-rights safeguards (e.g., non-coercion, transparency), address equity and social determinants, and support prevention and promotion as well as treatment. Therefore, we assess AI tools against these criteria in addition to model accuracy (Freeman, 2022).

7.3 Model Families and Positioning: GPT, Llama, and Mistral

LLMs differ in training data scale, tokenizer design, parameter count, and alignment methods. For proprietary models, GPT–3.5 and GPT–4/4o are widely used production systems; open models such as Llama 3 and Mistral offer strong alternatives with local deployability. While this thesis did not empirically benchmark proprietary APIs due to cost, privacy, and comparability constraints, we reference their public model cards to contextualize capability tiers (Meta AI, 2024; Mistral AI, 2024; OpenAI, 2023, 2024). In the mental health setting, practical considerations include guardrail maturity, privacy posture (on-device vs. cloud), latency, and cost per decision, which often dominate raw accuracy when designing triage workflows.

7.4 Training and Alignment of LLMs for Classification

Modern LLMs are trained in three stages:

(i) Unsupervised pretraining on large corpora via next-token prediction;

(ii) Task adaptation via supervised fine-tuning (SFT); and

(iii) Post-training alignment to human preferences and safety norms (e.g., Reinforcement learning from human feedback (RLHF) or direct preference optimization).

SFT exposes models to instruction-style inputs and task exemplars; RLHF optimizes responses using learned reward models that reflect human judgments of helpfulness and harmlessness (Bai et al., 2022; Ouyang et al., 2022). For mental health applications, alignment is not merely stylistic: it shapes denials around high-risk content, crisis guidance, and mitigation. However, preference datasets might contain cultural or demographic biases; consequently, domain-specific assessment and clinician-in-the-loop review remain necessary. In-context scenarios and prompt style interact with instruction following for zero/few-shot classification; attentive prompt design can significantly impact macro-F1 without changing model weights.

7.5 Prompting Strategies for Sentiment and Risk Classification

We consider three prompting regimes:

(i) Zero-shot with a calibrated instruction and constrained JSON output;

(ii) Few-shot with representative class exemplars (balanced across minority labels); and

(iii) Chain-of-thought (CoT) with rationale suppression at output (to avoid leaking sensitive content).

In safety-critical use, we favor JSON-only answers and deterministic decoding (low temperature) to reduce output variance. Prompt policies should explicitly define label schema, tie-breaking rules, and abstention (`"uncertain"`) to support selective prediction and human review. We recommend logging prompts and versions for auditability and assessing prompt robustness (paraphrase, adversarial cues) during validation.

7.6 Parameter-Efficient Adaptation (LoRA/PEFT)

Full fine-tuning of large models is costly. Parameter-efficient fine-tuning (PEFT) methods such as Low-Rank Adaptation (LoRA) insert trainable low-rank matrices into attention/MLP layers while freezing the backbone (Hu et al., 2022). PEFT typically updates $\ll 1\%$ of parameters, reducing compute, storage, and the risk of catastrophic forgetting. For mental health sentiment tasks, PEFT on curated, de-identified text can close much of the gap to full fine-tuning while easing deployment constraints. As Hu et al. (2022) mentioned, PEFT also enables label-space updates (e.g., adding a new risk category) with minimal retraining. When combined with instruction-tuned bases, PEFT tends to improve consistency and calibration over pure prompting baselines.

7.7 Retrieval-Augmented Classification (RAG)

RAG conditions the model on retrieved passages from a vetted corpus, improving faithfulness and reducing hallucination (Lewis et al., 2020). For classification, retrieval can surface policy snippets (e.g., crisis guidance, referral protocols) or domain definitions that help disambiguate adjacent labels (Depression vs. Suicidal). In mental health contexts, the retrieval index must exclude personal data and adhere to privacy and licensing constraints. RAG can also document the rationale provenance (which sources informed the label), aiding audit and clinician review.

7.8 Uncertainty, Calibration, and Selective Prediction

Risk-aware deployment requires calibrated confidence. While decoder likelihoods are not probabilities over labels, one can map scores to calibrated confidences using validation data (e.g., temperature scaling for logit-based heads, Platt or isotonic calibration for classifiers) (Guo et al., 2017). The Brier score summarizes probabilistic accuracy; precision–recall curves remain preferred under class imbalance. For high-stakes labels (e.g., Suicidal), selective prediction abstains when confidence falls below a threshold, routing cases to human review. Conformal prediction further provides finite-sample coverage guarantees for prediction sets under exchangeability (Vovk et al., 2005). We recommend reporting coverage vs. set size and auditing failure modes by class.

7.9 Safety, Privacy, and Governance for LLMs in Mental Health

Safety extends beyond content filtering. Recent studies highlight over-trust, unclear data handling, sensitivity of emotional disclosures, and risks across data, model, user, and context (Baidal et al., 2025; Kwesi et al., 2025). Governance should ensure:

(i) *Scope and boundaries (context):* Stating the intended use (e.g., screening, psycho-education); explicitly exclude diagnosis/treatment; detect off-label use and providing clear escalation.

(ii) *Audit and evaluation (model/user):* Auditing bias and performance across subgroups; testing consistency across prompts and languages; training crisis-risk scenarios; monitor outcomes after deployment.

(iii) *Data governance (data):* Tracking data provenance; collect only what is necessary for the stated purpose; mitigating privacy-extraction and linkability risks; document retention and sharing.

(iv) Privacy-by-default (data): Avoiding persistent identifiers; disabling the log retention and model-training on user content by default where feasible; utilizing short-lived encrypted storage and prefering on-device processing when possible.

(v) *Up-to-date knowledge (data/model):* Using retrieval-augmented answers grounded in dated, citable sources; maintaining and documenting a regular knowledge-refresh process.

(vi) *Robustness and safety operations (model):* Red-team for poisoning, jailbreaks, and prompt-injection; run adversarial tests; keep an incident-response playbook for harmful outputs and data events.

(vii) *User safeguards (user):* Providing contextual privacy notices for sensitive content; offering simple delete/export and opt-out controls; add guardrails against manipulation; design to reduce over-reliance; support "private mode" sessions.

(viii) *Explainability and transparency (model/context):* Publishing plain-language policies and technical artifacts (model/prompt cards, data-flow and retention diagrams, evaluation reports).

(ix) *Human oversight (context):* Routing high-risk interactions to qualified humans; defining governance roles; keep decision logs for accountability.

Clinical judgment ought to be supplemented by these systems, not replaced. For responsible deployment, institutional review, ongoing monitoring, and written process controls are essential (Baidal et al., 2025; Kwesi et al., 2025).

Trustworthy-AI lens

Complementing the above controls, we apply five trustworthy AI principles (Übellacker, 2024):

(i) *Beneficence:* Limiting scope to supportive, non-clinical use; measurement of benefit and harm; avoiding unwarranted efficacy claims.

(ii) *Non-maleficence:* Running red-teaming; maintaining crisis-escalation pathways; block unsafe patterns; conduct regular safety drills.

(iii) *Autonomy:* Keeping a human in the loop for high-risk cases; obtain clear, granular consent; provide easy opt-out, deletion, and data portability.

(iv) *Justice:* Auditing subgroup fairness; ensure accessible UX; reviewing equity in refusal and triage behavior.

(v) *Explicability:* Providing model/prompt cards; showing data-flow and retention diagrams; publishing versioned evaluation reports for review.

7.10 Cost, Latency, and Deployment Trade-offs

Model choice entails operational trade-offs. Proprietary APIs (e.g., GPT–4/4o) offer strong capability and mature guardrails at a recurring per-token cost; open-weight models (e.g., Llama 3, Mistral/Mixtral) enable private inference and tighter latency control on suitable hardware. For batch triage, feasibility is driven by throughput (requests/sec) and cost per 1k tokens; for interactive agents, tail latency and consistent, policy-aligned refusal behavior are critical. Parameter-efficient tuning and quantization reduce serving cost, and retrieval can shrink prompts to control context-window spend.

In practice, empirical profiling shows that sparse Mixture-of-Experts (MoE) models can match dense accuracy while supporting larger batches and higher single-GPU throughput. The MoE layer dominates runtime; batch size is bounded mainly by GPU memory and sequence length; and throughput typically shifts from memory- to compute-bound as batch size increases. An analytical model further estimates

Table 10: LLM families at a glance (capabilities and deployment posture; sources: model cards).

Model family	Access	Notable strengths	Deployment
GPT–3.5 / GPT–4/4o	Proprietary API	Strong instruction following; mature safety tooling; long context (4o)	Cloud API
Llama 3 (8B/70B)	Open weights	Competitive instruction models; local fine-tuning via PEFT	On-premises / cloud
Mistral / Mixtral	Open weights	Efficient small/mixture models; good latency	On-premises / cloud
Gemini 1.5	Proprietary API	Multimodal; long context	Cloud API

fine-tuning cost across GPUs (e.g., A40/A100/H100), helping select a cost-efficient setup (Xia et al., 2024).

7.11 Reference Model Families and Sources

We cite representative proprietary and open LLM families for context (GPT–3.5, GPT–4/4o; Llama–2/3; Mistral) but do not run new API-based experiments due to cost, privacy, and comparability constraints. Specifications are taken from public model cards (Meta AI, 2024; Mistral AI, 2024; OpenAI, 2023, 2024). Table 10 follows public results that are not directly comparable to our supervised experiments without matched data, prompts, and evaluation protocols.

7.12 Leveraging GPT-2 for Text Generation in Mental Health Support

GPT-2 (Generative Pretrained Transformer 2) is a generative language model that employs the transformer architecture to produce fluent, context-sensitive continuations from short prompts (Radford et al., 2019; Vaswani et al., 2017). Pretraining on web-scale corpora equips GPT-2 with broad lexical and syntactic knowledge, enabling strong zero-shot and few-shot performance on completion, summarization, and dialogue. Although surpassed in scale by recent models, GPT-2 remains a practical baseline for domain adaptation and controlled generation.

For mental health related prompts, A pretrained GPT-2 model and tokenizer have been used to generate continuations; the primary instance was taken from a dataset that focused on anxiety. The resulting text illustrates topical coherence and salient, contextually relevant language. Prompting with "Tips for managing anxiety" demonstrated the model's ability to produce concise, actionable suggestions in a

self-help style. These demonstrations underline GPT-2's potential to support content creation for psycho-education and low-intensity digital interventions (Radford et al., 2019; Vaswani et al., 2017).

Domain adaptation and safety

We develop the model for fine-tuning on a curated corpus of mental health statements to improve domain alignment. Fine-tuning minimizes irrelevant or off-topic outputs and aligns word choices, tone, and content with domain standards (e.g., psycho-educational accuracy, supportive style). The application should include adaptable safety mechanisms (request limits, response filtering), rigorous human review, and evaluation against clinical style criteria to reduce risks such as hallucinations, overgeneralization, or inappropriate advice. Above all, AI-generated information should be used to support professional treatment; open communication with crisis resources and clinician oversight are essential.

Evidence of applicability

Recent applications suggest that transformer-based models, including GPT-2, can be adapted for mental health contexts ranging from informational chat-bots to therapy-adjacent support tools, provided that domain-specific training and guardrails are in place (Mohssen & Safi, 2024). Within these bounds, GPT-2 offers a tractable platform for research prototypes and for studying the effect of targeted fine-tuning on generation quality, specificity, and safety in anxiety-management content.

Beyond GPT-2

Newer instruction following LLMs (GPT-3/3.5/4) can accelerate qualitative coding and theory-building workflows (structured prompting, JSON outputs, RAG, long-context). We treat these as tooling rather than clinical evidence, while domain claims remain grounded in mental health specific sources elsewhere (Übellacker, 2024).

Next, we present LLM assisted triage examples drawn from the Kaggle dataset analyzed in this thesis. In our implementation, prompts are instrumented to return schema-validated JSON fields (label, risk, rationale, escalation) for downstream modules.

7.13 LLM-Assisted Triage Examples (Illustrative; Not a Substitute for Clinical Care)

The following responses were generated by ChatGPT (OpenAI, 2025)

Anxiety:

User input: "I haven't slept well for 2 days; I'm restless. Why huh?"

Interpretation: Sleep disturbance plausibly linked to stress or anxiety; alternatively, positive arousal/overstimulation may contribute.

Response characteristics: Brief, empathetic normalization; suggest sleep-hygiene strategies and stress-reduction techniques; encourage monitoring and, if persistent/worsening, consultation with a professional.

Suicidality:

User input: "My idea of hell would be if there were a heaven... I really want

to be dead..."

Response characteristics: Prioritize safety; express empathy without minimization; advise immediate help-seeking (trusted contacts, crisis lines, emergency services if at risk); avoid detailed means discussion; encourage follow-up with mental health professionals. *Note:* High-risk content requires escalation pathways; automated support must be complemented by human intervention.

Depression:

User input: "I'm a pathetic piece of trash... my weekend is just sleeping, phone, one meal."

Interpretation: Marked self-deprecation, hopelessness, and behavioral withdrawal consistent with depressive symptoms.

Response characteristics: Validating distress; counter global negative self-judgments; suggest small, achievable activities and social contact; recommend professional evaluation if symptoms persist or impair functioning.

Bipolar:

User input: "Feeling like I'm being watched/judged constantly ... it usually feels invasive ... sometimes I weirdly enjoy it ... I'm not really sure how to deal with it."

Interpretation: Heightened self-referential salience and ideas-of-reference–like experiences can occur with anxiety or during mood elevation (hypomania/-mania), where arousal, distractibility, and attention to external cues increase; mixed feelings (distress yet occasional enjoyment) align with fluctuating mood and reward sensitivity.

Response characteristics: Validating the distress and confusion; encourage sleep regularity and daily routine; use grounding/attention-shifting (5–4–3–2–1 senses, paced breathing, descriptive labeling); reduce triggers (caffeine, alcohol/THC); track mood and sleep. Seek timely professional assessment if symptoms persist or are accompanied by decreased need for sleep, racing thoughts, grandiosity, marked impulsivity, or psychotic-like features. *Note:* If perceptions escalate to voices/commands, strong paranoia about harm, or any self-harm intent, prioritize immediate safety and crisis resources.

Normal:

User input: "my macbook just froze luckily i was able to take a screen shot of my paper and retyped the end of it i submitted my paper min late."

Interpretation: An everyday tech failure with acute frustration and time pressure; content is situational and not indicative of a mental health concern.

Response characteristics: Acknowledge the hassle and stress; suggest practical safeguards (autosave, frequent versions, cloud sync, scheduled backups, submission time buffer); recommend quick recovery steps (forced restart, updates, disk checks) and short decompression after submission (hydration, brief walk, breathing). Encourage self-compassion for minor delays; if tech issues or deadline stress recur frequently, plan workflow and contingency buffers.

Stress:

User input: "I'm 29 and I have great difficulty relaxing."

Interpretation: Ongoing hyperarousal and trouble down-regulating; often linked to workload, rumination, sleep debt, or stimulant use. Screen for duration, impairment, muscle tension, irritability, and sleep disruption; differentiate from generalized anxiety if worry is pervasive and hard to control.

Response characteristics: Validate the experience; suggest brief, evidence-based down-regulation: diaphragmatic breathing (inhale 4s, exhale 6s) 5–10 minutes, progressive muscle relaxation, and brief mindfulness. Recommend sleep hygiene (regular schedule, evening screen/caffeine limits), micro-breaks and boundary setting, scheduled worry time, and regular physical activity. If symptoms persist >2 weeks, markedly impair functioning, or co-occur with panic/insomnia/depressive symptoms, seek professional evaluation.

Personality disorder:

User input: "Just got home from a party ... I stood awkwardly the whole time, felt hyper-aware of eyes on me, avoid dancing/games/karaoke because I fear being judged and rejected. I wish I could let loose and have fun."

Interpretation: Strong self-consciousness, fear of negative evaluation, and rigid avoidance of self-expression; substance use (alcohol/THC) did not reduce hyper-arousal. This presentation overlaps with social anxiety and avoidant/self-critical trait patterns; dataset label does not imply diagnosis.

Response characteristics: Validate the distress; reduce safety behaviors (hiding, avoiding eye contact, constant self-monitoring). Use graded exposure to valued actions (e.g., greet one person, join a short game, sing one chorus) with pre-planned coping (paced breathing, attention shifting to external details). Practice compassionate self-talk and cognitive reappraisal of perfectionistic standards ("good enough participation over performance"). Limit alcohol/Tetrahydrocannabinol (THC) if they worsen monitoring or rebound anxiety. Set small approach goals, progress of the debrief, and reward effort. If longstanding and impairing, consider CBT/DBT or schema-focused therapy; seek professional evaluation.

LLMs can surface plausible interpretations and provide supportive, psychoeducational guidance. They must operate within safety guardrails and clearly direct users to professional care; they are adjuncts—not replacements—for clinicians.

7.14 Challenges and Limitations

Despite advances, foundation LLMs are optimized for broad objectives and often lack theory-guided, domain-specific priors for psychological applications (X. Wang et al., 2024). Performance can be brittle: cue-driven detectors (e.g., BERT/RoBERTa variants) may overfit lexical shortcuts and generalize poorly across contexts and populations (Ge et al., 2025). Clinical use raises safety, bias, and calibration concerns that require explicit guardrails and human oversight. Compute and environmental costs of training/fine-tuning remain substantial; parameter-efficient tuning (e.g., adapters/LoRA), transfer learning, and domain adaptation can improve validity while reducing resource use (X. Wang et al., 2024). Addressing these limitations is critical for trustworthy, scalable, and clinically actionable stress-detection systems.

Our models are cross-sectional and focus on system-level interactions; we do not capture PPCT-style proximal processes or chronosystem dynamics, which limits

causal and developmental inference (Eriksson et al., 2018).

7.15 Ethical Considerations

Use a precautionary, clinician-in-the-loop approach grounded in non-maleficence (Übellacker, 2024). Systems should be transparent about capabilities and limits, provide traceable rationales where feasible, and document data sources and evaluation procedures. Clinical judgment should not be replaced by LLMs, but supported; high-risk content (such as suicidal thoughts) should be immediately referred to crisis services, and experienced professionals should continue to make final decisions regarding diagnosis and treatment (Y. Liu & Wang, 2025; Übellacker, 2024). Regular bias checks, robust data set governance, and reducing unequal performance are essential for fairness. Data minimization, secure management, and regulatory compliance are essential for privacy and consent. Continuous post-release monitoring, audit trails, and transparent accountability for model changes and failures are essential for responsible deployment.

7.16 Evaluation Metrics

Robust evaluation requires metrics beyond overall accuracy under class imbalance (Ge et al., 2025). Models should report precision, recall (sensitivity), specificity, and F1-score. Calibration (e.g., Brier score) and per class analyses further clarify failure modes and clinical usability.

Foundational Model Paradigms

- **BERT** (Bidirectional Encoder Representations from Transformers): trained with bidirectional masked-language modeling; excels at contextual understanding for classification, QA, and NER.
- **GPT** (Generative Pretrained Transformer): auto-regressive, left-to-right training confers strong text generation and continuation abilities suitable for dialogue and creative or assistive writing.

These paradigms underpin subsequent LLM developments and benchmarks (X. Chen & Lin, 2025).

7.17 Future Directions

- Combining mental health detection with event detection so that social media signals map to real-world risks (e.g., crises), enabling earlier alerts and triage (Ge et al., 2025).
- *Personalization with retrieval-augmented generation (RAG)*: tailoring assessments and guidance by retrieving domain knowledge and relational context (X. Chen & Lin, 2025; Ge et al., 2025).
- *Explainable LLM-based data synthesis:* Generating privacy-preserving synthetic corpora while studying explanation fidelity and reliability (Ge et al., 2025).
- *Theory-informed and hybrid modeling:* Embedding cognitive/clinical theory and combine semantic embeddings with statistical features (e.g., TF–IDF) (X. Chen & Lin, 2025; X. Wang et al., 2024).

- *Resource-efficient adaptation:* Transferring domain adaptation, learning and parameter-efficient tuning in order to reduce compute and environmental costs (X. Wang et al., 2024).
- *Ethics, safety, and governance:* Enforcing transparency, human oversight, bias auditing, and privacy protections across the lifecycle (Übellacker, 2024).
- *Cross-platform and multimodal fusion:* Aggregating of signals across platforms and modalities (text, imagery, metadata) (X. Chen & Lin, 2025; Ge et al., 2025).
- *Longitudinal validation:* Tracking within user language changes over time to assess trajectory and intervention impact (X. Wang et al., 2024).

Although LLMs show promise for nuanced language understanding, the standard use for classification remains limited; integrating of complementary models and multi-modal data, alongside rigorous ethical safeguards, is essential for accurate, trustworthy, and clinically relevant deployment (X. Chen & Lin, 2025; Ge et al., 2025; Übellacker, 2024; X. Wang et al., 2024).

7.18 Computing Environment

Experiments were executed in Jupyter (Notebook/Lab) on a Lenovo ThinkPad running Windows 11 with an Intel Core i7-8550U (4c/8t, base 1.99 GHz), 16 GB RAM, Intel UHD Graphics 620 (integrated iGPU; shared memory 7.9 GB; no CUDA), and a 512 GB NVMe SSD; Python 3.11.9. All training and evaluation ran on CPU (no discrete GPU acceleration)

7.19 Challenges Associated with Current Models

During data testing and model training, several practical constraints emerged when implementing architectures such as BERT, GPT-2, and RoBERTa. Architectural complexity and high computational demands were the main causes of this; assessment and training cycles frequently took longer than five minutes per iteration, which hindered experimentation and hyperparameter adjustment. A primary bottleneck involved software compatibility: a Python 3.11.9 environment proved incompatible with key dependencies (notably PyTorch and packages in the transformers stack). Standardizing to Python 3.10.17 and aligning PyTorch, CUDA (where applicable), and transformers versions resolved conflicts and restored a stable pipeline. Operational mitigations included isolated virtual environments, pinned package versions, mixed precision where hardware permitted, and caching datasets/models. With these adjustments, training and evaluation became more robust and time efficient.

This chapter positioned large language models (LLMs) for mental health text tasks, outlining:

(i) Capabilities - rich semantics, few-shot generalization, rationale generation;

(ii) Practical adaptation routes - prompting, PEFT/LoRA, retrieval-augmentation;

(iii) Risk-aware operation - calibration, uncertainty, selective prediction; and

(iv) Governance requirements - safety, privacy, fairness, transparency.

We also mentioned the necessity for clinical workflow deployment and the existing restrictions (dataset shift, brittleness, computation).

We now turn to the *Findings* to quantify performance. Using the common protocol established earlier, we report accuracy and macro/weighted F1, confusion matrices, and precision–recall analyses to compare classical baselines (TF–IDF + linear models), neural models (CNN/LSTM), and transformer encoders (BERT), highlighting trade-offs between accuracy, robustness, and interpretability.

8 Findings

The performance of the candidate models was evaluated on a common held out split using two complementary metrics—accuracy and weighted F1 in order to reflect overall correctness and the balance between precision and recall under class imbalance. The results are summarized in Table 11.

Model	Accuracy	F1-score (weighted)
Logistic Regression	0.76	0.76
Decision Tree	0.64	0.64
Random Forest	0.70	0.69
Naive Bayes	0.67	0.66
SVM (LinearSVC)	0.78	0.77
LSTM	0.73	0.73
CNN (Conv1D)	0.76	0.76
BERT	0.85	0.84
TF–IDF + Logistic Regression	0.76	0.76
Word2Vec + Logistic Regression	0.61	0.63
Sentence-Transformers + Logistic Regression	0.73	0.74

Table 11: Comparative Summary Table

Across models as seen in Table 11, the transformer baseline (BERT) yields the strongest results (accuracy 0.85, weighted F1 0.84). Among non-transformer approaches, linear margins on TF–IDF are consistently strong: SVM reaches 0.78/0.77 and TF–IDF + logistic regression 0.76/0.76, closely matched by the CNN (0.76/0.76). The LSTM attains 0.73/0.73 under the short-sequence setup used here. Tree-based learners are moderate (random forest 0.70/0.69, naive Bayes 0.67/0.66, decision tree 0.64/0.64). Among embedding baselines, sentence-transformer embeddings with logistic regression are competitive but trail the best linear model (0.73/0.74), while average Word2Vec embeddings perform lowest (0.61/0.63).

Interpretation

(i) Lexical TF–IDF with linear decision functions remains a strong, transparent baseline.

(ii) Contextual encoders (BERT) provide the best balance overall.

(iii) Sentence-level embeddings without task adaptation are competitive but do not surpass the top TF–IDF margin; simple Word2Vec averaging is least effective.

(iv) For minority, clinically salient labels, adopt explicit imbalance mitigation (class weights/thresholding), probability calibration, and targeted error analysis of adjacent classes (e.g., *Depression* vs. *Suicidal*).

8.1 Visualization of Results

The bar charts in Figure 27 summarize accuracy and weighted F1, underlying the gap between contextual encoders and simpler baselines and the strong performance of linear TF–IDF margins.

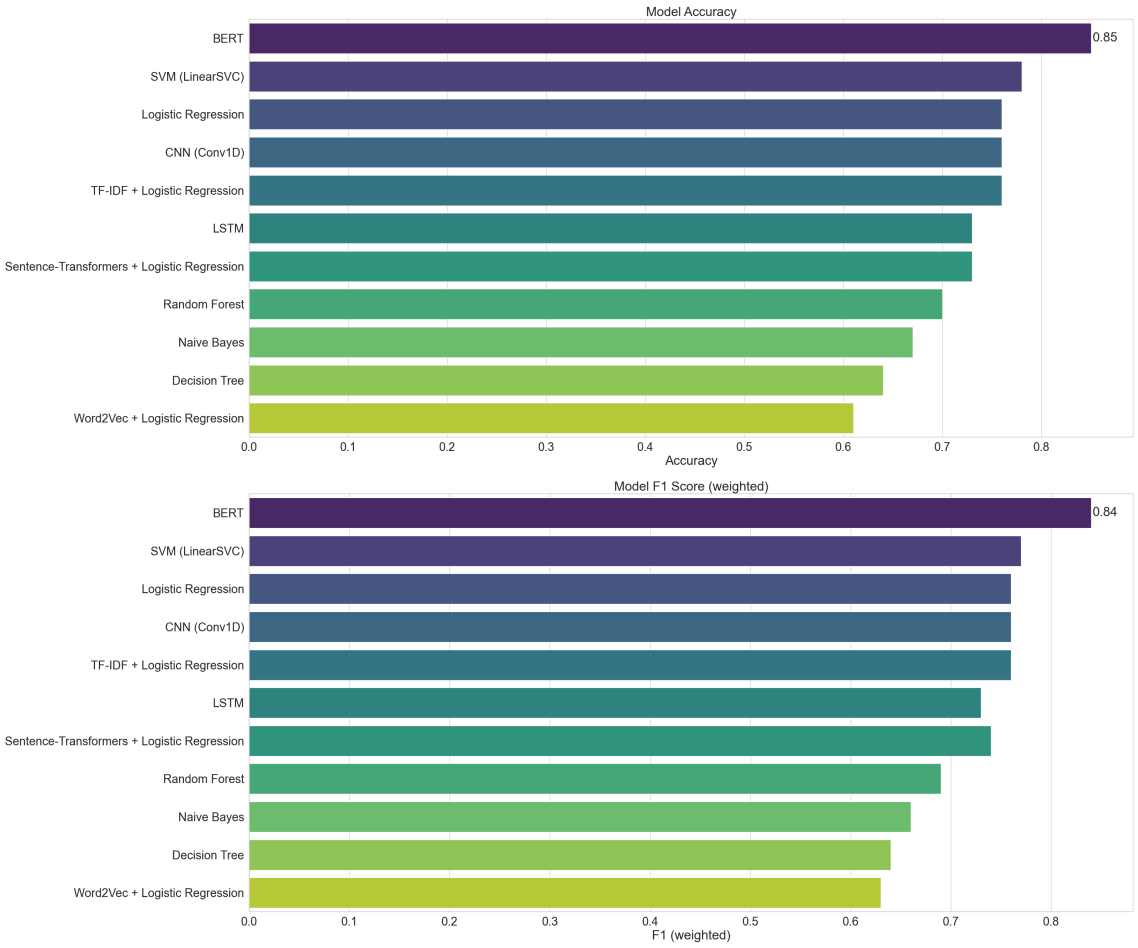

Figure 27: Model comparison bars

9 Conclusion

This thesis examined how Natural Language Processing can support mental health assessment from social media text by comparing classical ML models with deep neural architectures under a unified, reproducible pipeline. A curated, multi-source corpus was preprocessed, exploratory analyses informed feature design (including log transforms for skewed counts), and models were evaluated on a held out split using accuracy and F1 (macro and weighted) to account for class imbalance.

9.1 Summary of Findings

Contextual encoders delivered the strongest overall performance: BERT achieved the top results on our test set (accuracy 0.85, weighted F1 0.84). Among non-transformer approaches, linear margins on TF–IDF are consistently strong: SVM reached 0.78/0.77 (accuracy/F1), and TF–IDF + logistic regression 0.76/0.76, closely matched by the CNN baseline 0.76/0.76. The LSTM attained 0.73/0.73 under the short-sequence setup used here. Tree-based learners were moderate (random forest 0.70/0.69, decision tree 0.64/0.64, naive Bayes 0.67/0.66). Among embedding baselines, sentence-transformer embeddings with logistic regression were competitive (0.73/0.74) but did not surpass the best TF–IDF margin, while average Word2Vec embeddings performed lowest (0.61/0.63). These results indicate that context-aware encoders (transformers) and strong lexical baselines (linear TF–IDF) are most effective on this corpus, whereas simple embedding averages under-represent clinically salient nuance.

9.2 Contributions

(i) A transparent, apples-to-apples comparison across classical and deep models under consistent preprocessing and metrics;

(ii) Evidence that TF–IDF remains a strong, low-cost representation for linear learners on sparse text;

(iii) Feature analysis shows that length measurements are largely redundant and should be combined or normalized, but at the same time, stylistic proxies (e.g., average word length) add complementary signal; and

(iv) A reproducible pipeline with documented artifacts for assisting additional research and implementation.

9.3 Limitations

This study has several limitations. First, class imbalance and possible label noise reduce recall for minority classes and can bias aggregate scores. Second, because the dataset spans multiple platforms and time periods, external validity to new communities, languages, or time windows will require domain adaptation and prospective validation. Third, the computational demands of large transformer models constrain real time or edge deployment. Finally, ethical considerations, including privacy, informed consent, fairness across populations, and the risk of harm from misclassification, necessitate human oversight, transparent documentation, and formal governance.

9.4 Future Work

Priorities include:

(1) Stronger imbalance treatment (class-weighted losses, calibrated thresholds, resampling),

(2) Domain adaptation and continual learning for distribution shift,

(3) Hybrid feature design that fuses contextual embeddings with targeted linguistic features,

(4) Interpretable analyses (e.g., SHAP/attention) to surface actionable cues,

(5) Resource-efficient adaptation of large models (parameter-efficient tuning), multilingual/multimodal extensions, and

(6) Rigorous ethical safeguards with clinician-in-the-loop evaluation and clear escalation pathways for high-risk content.

9.5 Closing Remarks

NLP provides a reliable path to scalable and timely support for mental health assessment. With appropriate safeguards, contextual models can augment clinical judgment without replacing it by improving early screening and long term monitoring. The comparative evidence and pipeline developed here form a practical basis for future interdisciplinary work focused on higher accuracy, greater robustness, and responsible deployment.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/1810.03292

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, *39*(1), 45–65. https://doi.org/10.1016/S0306-4573(02)00021-3

Alahmadi, K., Alharbi, S., Chen, J., & Wang, X. (2025). Generalizing sentiment analysis: A review of progress, challenges, and emerging directions. *Social Network Analysis and Mining*, *15*(45). https://doi.org/10.1007/s13278-025-01234-5

Amartey, R. T. (2023, August). Natural language processing for mental health diagnostics. https://doi.org/10.13140/RG.2.2.35011.21283

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders (DSM-5-TR)*. American Psychiatric Publishing.

Angermeyer, M. C., & Matschinger, H. (2003). Public beliefs about schizophrenia and depression: Similarities and differences. *Social Psychiatry and Psychiatric Epidemiology*, *38*, 526–534. https://doi.org/10.1007/s00127-003-0676-6

Angermeyer, M. C., & Matschinger, H. (2004). Public attitudes to people with depression: Have there been any changes over the last decade? *Journal of Affective Disorders*, *83*(2-3), 177–182. https://doi.org/10.1016/j.jad.2004.08.001

Antonovsky, A. (1996). The salutogenic model as a theory to guide health promotion. *Health Promotion International*, *11*(1), 11–18. https://doi.org/10.1093/heapro/11.1.11

Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv:2212.08073*. https://arxiv.org/abs/2212.08073

Baidal, M., Derner, E., & Oliver, N. (2025). Guardians of trust: Risks and opportunities for LLMs in mental health. *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, 11–22.

Bailey, E., Boland, A., Bell, I., Nicholas, J., La Sala, L., & Robinson, J. (2022). The mental health and social media use of young Australians during the COVID-19 pandemic. *International Journal of Environmental Research and Public Health*, *19*(3), 1077. https://doi.org/10.3390/ijerph19031077

Bateman, A. W., & Krawitz, R. (2013). *Borderline personality disorder: An evidence-based guide for generalist mental health professionals*. Oxford University Press. https://doi.org/10.1093/med:psych/9780199644209.001.0001

Best, P., Manktelow, R., & Taylor, B. (2014). Online communication, social media and adolescent wellbeing: A systematic narrative review. *Children and Youth Services Review*, *41*, 27–36.

Bhugra, D., Till, A., & Sartorius, N. (2012). What is mental health? *International Journal of Social Psychiatry*, *59*(1). https://doi.org/10.1177/0020764012463315

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bisson, J. I., Ehlers, A., Matthews, R., Pilling, S., Richards, D., & Turner, S. (2018). Psychological treatments for chronic post-traumatic stress disorder: Systematic review and meta-analysis [Published online 02 January 2018]. *The British Journal of Psychiatry*. https://doi.org/10.1192/bjp.2017.27

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

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Wadsworth.

Brough, P., Biggs, A., Blaxland, M., & et al. (2020). Work-life balance in the digital era: The role of icts. *Annual Review of Organizational Psychology and Organizational Behavior*, *7*, 291–324.

Bruce, M. J., Pagán, A. F., & Acierno, R. (2025). State of the science: Evidence-based treatments for posttraumatic stress disorder delivered via telehealth. *Journal of Traumatic Stress*, *38*, 5–15. https://doi.org/10.1002/jts.23074

Bucur, A. M., Moldovan, A. C., Parvatikar, K., Zampieri, M., KhudaBukhsh, A. R., & Dinu, L. P. (2024). On the state of NLP approaches to modeling depression in social media: A post-COVID-19 outlook. https://arxiv.org/abs/2410.08793

Byrne, S., & McLean, N. (2002). Elite athletes: Effects of the pressure to be thin. *Journal of Science and Medicine in Sport*, *5*(2), 80–94.

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

Carolan, S., Harris, P. R., & Cavanagh, K. (2017). Ehealth interventions for mental health in the workplace: A systematic review. *Journal of Medical Internet Research*, *19*(10), e323. https://doi.org/10.2196/jmir.6396

Carr, C. T., & Hayes, R. A. (2015). Social media: Defining, developing, and divining. *Atlantic Journal of Communication*, *23*(1), 46–65. https://doi.org/10.1080/15456870.2015.972282

Charuvastra, A., & Cloitre, M. (2008). Social bonds and posttraumatic stress disorder. *Annual Review of Psychology*, *59*, 301–328. https://doi.org/10.1146/annurev.psych.58.110405.085650

Chauhan, V., Sharma, A., & Sikka, G. (2021). The emergence of social media data and sentiment analysis in election prediction: A review. *Computers in Human Behavior Reports*, *3*, 100073. https://doi.org/10.1016/j.chbr.2021.100073

Chen, K., Lim, N., Lee, C., & Guerzhoy, M. (2024). Exploring complex mental health symptoms via classifying social media data with explainable LLMs. https://arxiv.org/abs/2412.10414

Chen, L., Ho, S. S., & Lwin, M. O. (2017). A meta-analysis of factors predicting cyberbullying perpetration and victimization: From the social cognitive and media effects approach. *New Media & Society*, *19*(8), 1194–1213. https://doi.org/10.1177/1461444816634037

Chen, X., & Lin, X. (2025). Generating medically informed explanations for depression detection using LLMs. https://arxiv.org/pdf/2503.14671

Chugh, S. (2020). Mental health and covid. *Journal of Medical Evidence*, *1*(2), 159–160. https://doi.org/10.4103/JME.JME_186_20

Colom, F., Vieta, E., Martinez-Aran, A., Reinares, M., Goikolea, J. M., Benabarre, A., Torrent, C., Comes, M., & Corominas, J. (2003). A randomized trial on the efficacy of group psychoeducation in the prophylaxis of recurrences in bipolar patients whose disease is in remission. *Archives of General Psychiatry*, *60*(4), 402–407. https://doi.org/10.1001/archpsyc.60.4.402

Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural Language Processing of Social Media as Screening for Suicide Risk [CC BY-NC 4.0]. *Biomedical Informatics Insights*, *10*, 1–11. https://doi.org/10.1177/1178222618792860

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Cristea, I. A., et al. (2015). The effectiveness of cognitive behavioral therapy for anxiety disorders: A meta-analysis. *Clinical Psychology Review*, *37*, 36–50.

Deng, F., & Jiang, X. (2022). Effects of human versus virtual human influencers on the appearance anxiety of social media users. *Journal of Retailing and Consumer Services*, *65*, 103233. https://doi.org/10.1016/j.jretconser.2022.103233

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 4171–4186. https://arxiv.org/abs/1810.04805

Ding, C., & He, X. (2004). K-Means Clustering via Principal Component Analysis. *Proceedings of the 21st International Conference on Machine Learning (ICML 2004)*, 29. https://doi.org/10.1145/1015330.1015408

Ding, Z., Wang, Z., Zhang, Y., Cao, Y., Liu, Y., Shen, X., Tian, Y., & Dai, J. (2025). Efficient or powerful? Trade-offs between machine learning and deep learning for mental illness detection on social media [arXiv preprint]. https://arxiv.org/abs/2503.01082

Eather, N., Morgan, P. J., Lubans, D. R., Lonsdale, C., & Ntoumanis, N. (2023). Sport and physical activity as part of a mental health recovery and prevention strategy [CC BY 4.0]. *Frontiers in Sports and Active Living*, *5*, 1223459. https://doi.org/10.3389/fspor.2023.1223459

Echezarraga, A., Calvete, E., González-Pinto, A. M., & Las Hayas, C. (2018). Resilience dimensions and mental health outcomes in bipolar disorder in a follow-up study. *Stress and Health*, *34*, 115–126. https://doi.org/10.1002/smi.2767

Engel, G. L. (1977). The need for a new medical model: A challenge for biomedicine. *Science*, *196*(4286), 129–136. https://doi.org/10.1126/science.847460

Eriksen, K., & Kress, V. E. (2008). Gender and diagnosis: Struggles and suggestions for counselors. *Journal of Counseling & Development*, *86*(2), 152–162. https://doi.org/10.1002/j.1556-6678.2008.tb00495.x

Eriksson, M., Ghazinour, M., & Hammarström, A. (2018). Different uses of Bronfenbrenner's ecological theory in public mental health research: Value for policy and practice. *Social Theory & Health*, *16*, 414–433. https://doi.org/10.1057/s41285-018-0065-6

Eurofound and the International Labour Office. (2021). *Telework and ICT-based mobile work: Flexible working in the digital age*. Publications Office of the European Union/ILO. https://www.eurofound.europa.eu/

Faraone, S. V., Asherson, P., Banaschewski, T., & et al. (2021). Attention-deficit/hyperactivity disorder. *Nature Reviews Disease Primers*, *7*(1), 42.

Finkelstein, J., & Lapshin, O. (2007). Reducing depression stigma using a web-based program. *International Journal of Medical Informatics*, *76*(10), 726–734. https://doi.org/10.1016/j.ijmedinf.2006.07.004

Frank, E., Kupfer, D. J., Thase, M. E., Mallinger, A. G., Swartz, H. A., Fagiolini, A. M., Grochocinski, V. J., Houck, P., Scott, J., Thompson, W., & Monk, T. (2005). Two-year outcomes for interpersonal and social rhythm therapy in individuals with bipolar disorder. *Archives of General Psychiatry*, *62*(9), 996–1004. https://doi.org/10.1001/archpsyc.62.9.996

Freeman, M. (2022). The world mental health report: Transforming mental health for all. *World Psychiatry*. https://doi.org/10.1002/wps.21018

Furnham, A., Lee, V., & Kolzeev, V. (2015). Mental health literacy and borderline personality disorder (BPD): What do the public make of those with BPD? *Social Psychiatry and Psychiatric Epidemiology*, *50*, 317–324. https://doi.org/10.1007/s00127-014-0936-7

Fuss, J., Briken, P., & Klein, V. (2018). Gender bias in clinicians' pathologization of atypical sexuality: A randomized controlled trial with mental health professionals. *Scientific Reports*, *8*, 3715. https://doi.org/10.1038/s41598-018-22108-z

GBD 2019 Diseases and Injuries Collaborators. (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019. *The Lancet*, *396*(10258), 1204–1222.

Ge, Z., Hu, N., Li, D., Wang, Y., Qi, S., Xu, Y., Shi, H., & Zhang, J. (2025). A Survey of Large Language Models in Mental Health Disorder Detection on Social Media. https://arxiv.org/abs/2504.02800

Generalised anxiety disorder and panic disorder in adults: Management (Clinical guideline No. CG113). (2011). National Institute for Health and Care Excellence (NICE). Retrieved October 17, 2025, from https://www.nice.org.uk/guidance/cg113

Golden, A., & Aboujaoude, E. (2024). The framework for AI tool assessment in mental health (FAITA – mental health): A scale for evaluating AI-powered mental health tools [Proposes a 6-dimension, 0–24 framework for assessing AI mental health tools].

Grande, I., Berk, M., Birmaher, B., & Vieta, E. (2016). Bipolar disorder. *The Lancet*, *387*(10027), 1561–1572.

Grant, B. F., Chou, S. P., Goldstein, R. B., Huang, B., Stinson, F. S., Saha, T. D., Smith, S. M., Dawson, D. A., Pulay, A. J., Pickering, R. P., & Ruan, W. J. (2008). Prevalence, correlates, disability, and comorbidity of DSM-IV borderline personality disorder. *Journal of Clinical Psychiatry*, *69*, 533–545.

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2019). Studying mental health through social media: A review. *npj Digital Medicine*, *2*, 54. https://doi.org/10.1038/s41746-019-0148-9

Guo, C., et al. (2017). On calibration of modern neural networks. *ICML*. https://arxiv.org/abs/1706.04599

Gusmão, R., Quintaão, S., McDaid, D., Arensman, E., Van Audenhove, C., Coffey, C., Värnik, A., Värnik, P., Coyne, J., & Hegerl, U. (2013). Antidepressant utilization and suicide in Europe: An ecological multi-national study. *PLoS ONE*, *8*(6), e66455. https://doi.org/10.1371/journal.pone.0066455

Gutner, C. A., Rizvi, S. L., Monson, C. M., & Resick, P. A. (2006). Changes in coping strategies, relationship to the perpetrator, and posttraumatic distress in female crime victims. *Journal of Traumatic Stress*, *19*(6), 813–823. https://doi.org/10.1002/jts.20147

Halioua, R., Koehler, K., Claussen, M. C., Lichtenstein, M. B., & Wasserfurth, P. (2025). The intersection of mental health and sports nutrition. *Sports Psychiatry*, *4*(2), 49–50. https://doi.org/10.1024/2674-0052/a000110

Halioua, R., Wasserfurth, P., Toepffer, D., Claussen, M. C., & Koehler, K. (2024). Exploring the relationship between low energy availability, depression and eating disorders in female athletes: A cross-sectional study. *BMJ Open Sport & Exercise Medicine*, *10*, e002035.

Hassan, N. K., Brighton, T. E., & Ogunrinde, V. (2025, May). Text preprocessing techniques in python for effective NLP pipelines. https://www.researchgate.net/publication/392164197_Text_Preprocessing_Techniques_in_Python_for_Effective_NLP_Pipelines

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd). Springer. https://hastie.su.domains/ElemStatLearn/

Hengle, A., Kulkarni, A., Patankar, S., Chandrasekaran, M., D'Silva, S., Jacob, J., & Gupta, R. (2024). Still not quite there! evaluating large language models for comorbid mental health diagnosis [License: CC BY 4.0]. https://arxiv.org/abs/2410.03908

Hinduja, S., & Patchin, J. W. (2008). Cyberbullying: An exploratory analysis of factors related to offending and victimization. *Deviant Behavior*, *29*(2), 129–156. https://doi.org/10.1080/01639620701457816

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hofmann, S. G., Asnaani, A., Vonk, I. J. J., Sawyer, A. T., & Fang, A. (2012). The efficacy of cognitive behavioral therapy: A meta-analytic review. *Cognitive Therapy and Research*, *36*(5), 427–440.

Hu, E. J., et al. (2022). LoRA: Low-Rank Adaptation of Large Language Models. *ICLR*. https://arxiv.org/abs/2106.09685

International classification of diseases 11th revision (ICD-11) [Online version]. (2019). World Health Organization. https://icd.who.int/browse/2025-01/mms/en#334423054

Jalilifard, A., Caridá, V. F., Mansano, A. F., Cristo, R. S., & da Fonseca, F. P. C. (2021). Semantic sensitive tf-idf to determine word relevance in documents. *Advances in Computing and Network Communications*, *736*, 327–337. https://doi.org/10.1007/978-981-33-6987-0_27

Jauhar, S., Johnstone, M., & McKenna, P. J. (2022). Schizophrenia. *The Lancet*, *399*(10323), 473–486. https://doi.org/10.1016/S0140-6736(21)01730-X

Jorm, A. F. (2012). Mental health literacy: Empowering the community to take action for better mental health. *American Psychologist*, *67*(3), 231–243. https://doi.org/10.1037/a0025957

Jorm, A. F., & Oh, E. (2009). Desire for social distance from people with mental disorders: A review. *Australian and New Zealand Journal of Psychiatry*, *43*(3), 183–200. https://doi.org/10.1080/00048670802653349

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing (3rd ed. draft)* [Draft chapters]. Retrieved October 7, 2025, from https://web.stanford.edu/~jurafsky/slp3/

Karasek Jr, R. A. (1979). Job demands, job decision latitude, and mental strain. *Administrative Science Quarterly*, *24*(2), 285–308.

Kiekens, G., et al. (2020). Non-suicidal self-injury: Prevalence, risk factors, and treatment. *The Lancet Psychiatry*, *7*(8), 654–667.

Kowalski, R. M., & Limber, S. P. (2013). Psychological, physical, and academic correlates of cyberbullying and traditional bullying. *Journal of Adolescent Health*, *53*(1), S13–S20. https://doi.org/10.1016/j.jadohealth.2012.09.018

Kwesi, J., Cao, J., Manchanda, R., & Emami-Naeini, P. (2025). Exploring user security and privacy attitudes and concerns toward the use of general-purpose LLM chatbots for mental health. *arXiv*. https://arxiv.org/abs/2507.10695

La Torre, G., Esposito, A., Sciarra, I., & Chiappetta, M. (2019). Technostress, a systematic review: Health risks and solutions. *International Journal of En-*

vironmental Research and Public Health, *16*(21), 3862. https://doi.org/10.3390/ijerph16213862

Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP. *NeurIPS*. https://arxiv.org/abs/2005.11401

Linehan, M. M. (2015). *DBT skills training manual* (2nd ed.). Guilford Press.

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Liu, Y., & Wang, F. (2025). Investigating the interpretability of ChatGPT in mental health counseling: An analysis of AI-generated content differentiation [In press]. *Computer Methods and Programs in Biomedicine*. https://www.sciencedirect.com/science/article/abs/pii/S0169260725002810

Livesley, W. J., Dimaggio, G., & Clarkin, J. F. (2018). Integrated treatment: A conceptual framework for an evidence-based approach to the treatment of personality disorder. *Journal of Personality Disorders*, *32*(1), 58–87.

Lundberg, S. M., & Lee, S.-I. (2017). A Unified approach to interpreting Model Predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*, 4765–4774. https://arxiv.org/abs/1705.07874

Lüttke, S., Hautzinger, M., & Fuhr, K. (2018). E-Health in Diagnostik und Therapie psychischer Störungen: Werden Therapeuten bald überflüssig? *Bundesgesundheitsblatt – Gesundheitsforschung – Gesundheitsschutz*, *61*(3), 263–270. https://doi.org/10.1007/s00103-017-2684-9

Marmot, M. (2005). Social determinants of health inequalities. *The Lancet*, *365*(9464), 1099–1104. https://doi.org/10.1016/S0140-6736(05)71146-6

Marumoto, T., Monma, T., Sawae, Y., & Takeda, F. (2025). Mental health and psychosocial status in mothers of children with attention deficit hyperactivity disorder (ADHD): Differences by maternal ADHD tendencies [Published online 12 April 2023]. *Child Psychiatry & Human Development*, *56*, 34–42. https://doi.org/10.1007/s10578-023-01532-x

McEwen, B. S. (2004). Protection and damage from acute and chronic stress: Allostasis and allostatic load. *Annals of the New York Academy of Sciences*, *1032*, 1–7. https://doi.org/10.1196/annals.1314.001

McGrath, J., Saha, S., Chant, D., & Welham, J. (2008). Schizophrenia: A concise overview of incidence, prevalence, and mortality. *Epidemiologic Reviews*, *30*, 67–76. https://doi.org/10.1093/epirev/mxn001

Meijerink, J., Keegan, A., Bondarouk, T., & Kuhn, K. M. (2021). Algorithmic management: Foundations, examples, and research agenda. *Human Resource Management Review*, *31*(4), 100780. https://doi.org/10.1016/j.hrmr.2020.100780

Messenger, J. C. (2019). *Telework in the 21st century: An evolutionary perspective*. Edward Elgar.

Meta AI. (2024). *Introducing Llama 3*. Retrieved October 8, 2025, from https://ai.meta.com/blog/llama-3/

Meyer, B., & Marks, R. (2018). Cognitive distortions in schizophrenia: A review of the literature. *Psychological Medicine*, *48*(9), 1398–1408. https://doi.org/10.1017/S003329171700303X

Miklowitz, D. J., George, E. L., Richards, J. A., Simoneau, T. L., & Suddath, R. L. (2003). A randomized study of family-focused psychoeducation and pharmacotherapy in the outpatient management of bipolar disorder. *Archives of*

General Psychiatry, *60*(9), 904–912. https://doi.org/10.1001/archpsyc.60.9.904

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Proc. ICLR (Workshop Track)*. https://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Proc. NeurIPS*.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning–based text classification: A comprehensive review. *ACM Computing Surveys*, *54*(3), 1–40. https://dl.acm.org/doi/10.1145/3439726

Mistral AI. (2024). *Mistral models documentation*. Retrieved October 8, 2025, from https://docs.mistral.ai/models/

Mohammad, S. M. (2024). WorryWords: Norms of anxiety association for over 44k English words. https://arxiv.org/abs/2411.03966

Mohssen, H. M., & Safi, H. H. (2024). Chatbot in the E-Service of Mental Health using the reprogramming of the GPT-2 Model [Informatics Institute for Postgraduate Studies, Iraqi Commission for Computers & Informatics, Baghdad, Iraq]. *Journal of International Crisis and Risk Communication Research*, *7*(S9).

Morgan, D. D., Higgins, C. D., & Rogers, C. R. (2025). The role of parent's mental health in shaping perceptions of adolescent mental health experiences after covid-19. *Family Relations*, *74*(2), 583–601. https://doi.org/10.1111/fare.13123

Mountjoy, M., et al. (2018). International Olympic Committee (IOC) consensus statement on Relative Energy Deficiency in Sport (RED-S): 2018 update. *International Journal of Sport Nutrition and Exercise Metabolism*, *28*(4), 316–331. https://doi.org/10.1123/ijsnem.2018-0136

Mountjoy, M., Ackerman, K. E., Bailey, D. M., Burke, L. M., Constantini, N., Hackney, A. C., Heikura, I. A., Melin, A., Pensgaard, A. M., Stellingwerff, T., Sundgot-Borgen, J. K., Torstveit, M. K., Jacobsen, A. U., Verhagen, E., Budgett, R., Engebretsen, L., & Erdener, U. (2023). 2023 International Olympic Committee's (IOC) consensus statement on Relative Energy Deficiency in Sport (REDs). *British Journal of Sports Medicine*, *58*(1), 3–24. https://doi.org/10.1136/bjsports-2023-106994

Mountjoy, M., Sundgot-Borgen, J., Burke, L., & et al. (2014). The IOC consensus statement: Beyond the female athlete triad – relative energy deficiency in sport (RED-S). *British Journal of Sports Medicine*, *48*(7), 491–497.

Mullins, J. T., & White, C. (2019). Temperature and mental health: Evidence from the spectrum of mental health outcomes. *Journal of Health Economics*, *68*, 102240. https://doi.org/10.1016/j.jhealeco.2019.102240

National Collaborating Centre for Mental Health. (2011). *Common mental health disorders: Identification and pathways to care*. The British Psychological Society; The Royal College of Psychiatrists. https://www.nice.org.uk/guidance/cg123

Nattiv, A., Loucks, A. B., Manore, M. M., Sanborn, C. F., Sundgot-Borgen, J., & Warren, M. P. (2007). The Female Athlete Triad [Position Stand of the American College of Sports Medicine]. *Medicine & Science in Sports & Exercise*, *39*(10), 1867–1882. https://doi.org/10.1249/mss.0b013e318149f111

Navarro, R., Yubero, S., & Larrañaga, E. (Eds.). (2016). *Cyberbullying across the globe: Gender, family, and mental health* [ISBN: 9783319255517]. Springer. https://doi.org/10.1007/978-3-319-25552-4

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. *International Conference on Machine Learning (ICML)*, 625–632. https://dl.acm.org/doi/10.1145/1102351.1102430

OECD. (2020). *Productivity gains from teleworking in the post-covid-19 era: How can public policies make it happen?* OECD Policy Responses to Coronavirus (COVID-19). https://www.oecd.org/

OpenAI. (2023). *OpenAI Models: GPT-3.5*. Retrieved October 8, 2025, from https://platform.openai.com/docs/models#gpt-3-5

OpenAI. (2024). *OpenAI Models: GPT-4o and GPT-4o mini*. Retrieved October 8, 2025, from https://platform.openai.com/docs/models#gpt-4o-and-gpt-4o-mini

OpenAI. (2025). ChatGPT (GPT-5.1). https://chat.openai.com/

Orben, A. (2020). Teenagers, screens and social media: A narrative review of reviews. *The Oxford Review of Education*, *46*(3), 347–362.

Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. *Nature Human Behaviour*, *3*, 173–182.

Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *NeurIPS*. https://arxiv.org/abs/2203.02155

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1–135.

Pedregosa, F., Varoquaux, G., et al. (2020). Scikit-learn: Machine learning in python (documentation). https://scikit-learn.org/stable/

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Pennebaker, J. W., & Chung, C. K. (2011). Expressive writing: Connections to physical and mental health. In J. W. Pennebaker (Ed.), *The Oxford handbook of health psychology* (pp. 417–437). Oxford University Press.

Pensgaard, A. M., Sundgot-Borgen, J., Edwards, C., Jacobsen, A. U., & Mountjoy, M. (2023). Intersection of mental health issues and relative energy deficiency in sport (RED-S): A narrative review. *British Journal of Sports Medicine*, *57*(17), 1127–1135.

Pliszka, S. R. (2007). Practice parameter for the assessment and treatment of children and adolescents with ADHD. *Journal of the American Academy of Child & Adolescent Psychiatry*, *46*(7), 894–921. https://doi.org/10.1097/chi.0b013e318054e724

Preventing suicide: A global imperative. (2014). https://www.who.int/publications/i/item/9789241564779

Prince, M., Patel, V., Saxena, S., Maj, M., Maselko, J., Phillips, M. R., & Rahman, A. (2007). No health without mental health. *The Lancet*, *370*(9590), 859–877. https://doi.org/10.1016/S0140-6736(07)61238-0

Prins, M. A., Verhaak, P. F. M., Bensing, J. M., & van der Meer, K. (2008). Health beliefs and perceived need for mental health care of anxiety and depression — the patients' perspective explored. *Clinical Psychology Review*, *28*(6), 1038–1058. https://doi.org/10.1016/j.cpr.2008.02.009

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners [OpenAI Technical Report]. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Ravi, K., & Ravi, V. (2015). A Survey on Opinion Mining and Sentiment Analysis: Tasks, Approaches and Applications. *Knowledge-Based Systems*, *89*, 14–46.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *Proc. EMNLP*. https://arxiv.org/abs/1908.10084

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Ruiz, M. T., & Verbrugge, L. M. (1997). A two-way view of gender bias in medicine. *Journal of Epidemiology and Community Health*, *51*(2), 106–109. https://doi.org/10.1136/jech.51.2.106

Ryff, C. D. (1989). Happiness is everything, or is it? explorations on the meaning of psychological well-being. *Journal of Personality and Social Psychology*, *57*(6), 1069–1081. https://doi.org/10.1037/0022-3514.57.6.1069

Saha, S. (2024). *BERT: Sentiment analysis for mental health (kaggle notebook)* [Kaggle notebook referenced for BERT implementation details]. Retrieved October 8, 2025, from https://www.kaggle.com/code/shuvosaha1234/bert-sentiment-analysis-for-mental-health

Sammut, C., & Webb, G. I. (2017). TF-IDF. In *Encyclopedia of machine learning and data mining*. Springer. https://openlibrary.org/books/OL26836425M/Encyclopedia_of_Machine_Learning_and_Data_Mining

Sarkar, S. (2024). *Sentiment analysis for mental health — dataset* [Kaggle project page (code section)]. Retrieved October 8, 2025, from https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health/code

Schulz, J. M., White, J., Edwards, C., Liu, S., & Thornton, J. S. (2025). Responding to relative energy deficiency in sport (RED-S): A multidisciplinary care pathway for safe return to sport. *Sports Psychiatry*, *4*(2), 95–107.

Schumm, J. A., Briggs-Phillips, M., & Hobfoll, S. E. (2006). Cumulative interpersonal traumas and social support as risk and resiliency factors in predicting PTSD and depression among inner-city women. *Journal of Traumatic Stress*, *19*(6), 825–836. https://doi.org/10.1002/jts.20159

Shen, Y., et al. (2022). *A survey on deep learning for sentiment analysis*. Retrieved October 8, 2025, from https://arxiv.org/abs/2205.12247

Siegrist, J., & Li, J. (2017). Work stress and health in the context of job demands-resources and effort–reward imbalance. *International Journal of Environmental Research and Public Health*, *14*(11), 1373. https://doi.org/10.3390/ijerph14111373

Siegrist, J., & Wahrendorf, M. (2016). *Work stress and health in a globalized economy: The model of effort–reward imbalance*. Springer. https://doi.org/10.1007/978-3-319-32937-6

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. https://dl.acm.org/doi/10.1145/3375627.3375830

Solstad, B. E., Fahrenholtz, I. L., Melin, A., Garthe, I., & Torstveit, M. K. (2025). Participant evaluations of the fuel intervention designed for female endurance

athletes at risk of RED-S: A mixed methods approach. *Sports Psychiatry*, *4*(2), 51–70.

Sparks, B., Zidenberg, A. M., & Olver, M. E. (2024). An exploratory study of incels' dating app experiences, mental health, and relational well-being. *The Journal of Sex Research*, *61*(7), 1001–1012. https://doi.org/10.1080/00224499.2023.2249775

Størbø, T., et al. (2020). Psychosocial therapies for borderline personality disorder: A systematic review and meta-analysis. *JAMA Psychiatry*, *77*(2), 127–137.

Sundgot-Borgen, J., & Torstveit, M. K. (2004). Prevalence of eating disorders in elite athletes is higher than in the general population. *Clinical Journal of Sport Medicine*, *14*(1), 25–32.

Tang, C. (2024, July). *Knowledge enhanced natural language generation* [Doctoral dissertation, University of Surrey]. https://doi.org/10.15126/thesis.901167

Tarafdar, M., Cooper, C. L., & Stich, L. (2019). The dark side of digitalization: Technostress and its consequences. *Information Systems Journal*, *29*(6), 1259–1279.

Tarafdar, M., Tu, Q., Ragu-Nathan, T. S., & Ragu-Nathan, B. S. (2007). The impact of technostress on productivity. *Journal of Management Information Systems*, *24*(1), 301–328.

Tavakoli, M., Salemi, A., Ye, C., Abdalla, M., Zamani, H., & Mitchell, J. R. (2025). Beyond a million tokens: Benchmarking and enhancing long-term memory in LLMs. *arXiv*. https://arxiv.org/abs/2510.27246

Theorell, T., Hammarström, A., Aronsson, G., & et al. (2015). A systematic review including meta-analysis of work environment and depressive symptoms. *BMC Public Health*, *15*(1), 738. https://doi.org/10.1186/s12889-015-1954-4

Tiet, Q. Q., Rosen, C., Cavella, S., Moos, R. H., Finney, J. W., & Yesavage, J. (2006). Coping, symptoms, and functioning outcomes of patients with posttraumatic stress disorder. *Journal of Traumatic Stress*, *19*(6), 799–811. https://doi.org/10.1002/jts.20149

Übellacker, T. (2024). Academiaos: Automating grounded theory development in qualitative research with large language models. https://arxiv.org/pdf/2403.08844

Umarogono, E., Suseno, J. E., & Gunawan, V. S. K. (2019). K-Means clustering optimization using the Elbow Method and Early Centroid determination based on mean and median formula. *Proceedings of the 2nd International Seminar on Science and Technology (ISSTEC 2019)*, *474*. https://doi.org/10.2991/isstec-19.2019.52

Valkenburg, P. M. (2017). Understanding self-effects in social media. *Human Communication Research*, *43*(4), 477–490. https://doi.org/10.1111/hcre.12113

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Proc. NeurIPS*. https://arxiv.org/abs/1706.03762

Verduyn, P., Ybarra, O., Résibois, M., Jonides, J., & Kross, E. (2017). Do social network sites enhance or undermine subjective well-being? A critical review. *Social Issues and Policy Review*, *11*(1), 274–302.

Vogt, D., Erbes, C. R., & Polusny, M. A. (2017). Role of social context in posttraumatic stress disorder (PTSD). *Current Opinion in Psychology*, *14*, 138–142. https://doi.org/10.1016/j.copsyc.2017.01.006

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: Deep Self-Attention distillation for Task-Agnostic compression of Pre-Trained Transformers. https://arxiv.org/abs/2002.10957

Wang, X., Gao, B., Dai, Y., Cao, L., Zhao, L., Yang, Y., & Clifton, D. (2024). Cognition Chain for explainable psychological stress detection on social media. https://arxiv.org/abs/2412.14009

WHO & ILO. (2022). Mental health at work: Policy brief. *World Health Organization and International Labour Organization*. https://www.who.int/publications/i/item/9789240053052

Willard, N. (2006). *Cyberbullying and cyberthreats*. Center for Safe; Responsible Internet Use. https://www.researchgate.net/publication/307981861_Cyberbullying_and_cyberthreats

Wood, A. J., Graham, M., Lehdonvirta, V., & Hjorth, I. (2019). Good gig, bad gig: Autonomy and algorithmic control in platform work. *Work, Employment and Society*, *33*(1), 56–75.

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

Xia, Y., Kim, J., Chen, Y., Ye, H., Kundu, S., Hao, C., & Talati, N. (2024). Understanding the performance and estimating the cost of LLM fine-tuning. https://arxiv.org/abs/2408.04693

Yen, S., Peters, J. R., Nishar, S., Grilo, C. M., Sanislow, C. A., Shea, M. T., Zanarini, M. C., McGlashan, T. H., Morey, L. C., & Skodol, A. E. (2021). Association of borderline personality disorder criteria with suicide attempts: Findings from the collaborative longitudinal study of personality disorders over 10 years of follow-up. *JAMA Psychiatry*, *78*(2), 187–194. https://doi.org/10.1001/jamapsychiatry.2020.3598

Zalsman, G., et al. (2016). Suicide prevention strategies revisited: 10-year systematic review. *The Lancet Psychiatry*, *3*(7), 646–659.

Zeitel-Bank, N., & Tat, U. (2014). Social media and its effects on individuals and social systems [Retrieved from ResearchGate or conference proceedings]. *Proceedings of the International Conference on Economic Sciences and Business Administration*.

Zhu, Y., Li, W., O'Brien, J. E., & Liu, T. (2021). Parent–child attachment moderates the associations between cyberbullying victimization and adolescents' health/mental health problems: An exploration of cyberbullying victimization among Chinese adolescents. *Journal of Interpersonal Violence*, *36*(17-18), NP9272–NP9298. https://doi.org/10.1177/0886260519854559