AI as a Mirror to the Mind: Analysing Mental Health Narratives in the Social Media Landscape

by

Saima Rani
Master of Computer Science

Thesis submitted for the fulfilment of the requirements for the degree of
Master Of Research

Victoria University, Australia
The Institute for Sustainable Industries and Liveable Cities (ISILC)

August 2024

Abstract

The intersection of mental health challenges and social media's broad reach has become a key public health research focus, particularly due to the profound global impacts of the COVID-19 pandemic on mental well-being. This thesis advances mental health research by developing a comprehensive framework for analysing mental health narratives from over one million Reddit posts during pre-pandemic, pandemic, and post-pandemic period. Our novel data annotation approach, integrated with Natural Language Processing (NLP) techniques, uncovers foundational triggers of mental health issues as expressed through social media discourse.

Two primary objectives are central to our study: (1) compiling an extensive corpus from social media posts to serve as a benchmark dataset for analysing mental health-related expressions, and (2) employing NLP-based framework for systematically identifying and analysing underlying factors contributing to mental health conditions. This dual approach not only facilitated the categorisation of mental health triggers but also enabled us to explore Artificial Intelligence (AI) models' alignment with human judgement, assessing AI's capability in interpreting complex mental health dialogues.

Our mixed-methods approach, combining quantitative analysis for identifying temporal trends in mental health discourse during the pandemic era, with qualitative thematic analysis of selected posts for in-depth exploration, offers a nuanced perspective of the mental health landscape. Key findings demonstrate that machine learning (ML) can approximate human decision-making in mental health assessment, revealing critical gaps that AI must bridge to fully capture the nuanced linguistic and emotional contexts of mental health discussions.

The public availability of our dataset, including an annotated subset as detailed in the data availability section, is poised to be a significant resource for training ML models. This work enriches academic discourse around mental health diagnostics and intervention, offering actionable insights for effective public health strategies. By delineating the role of AI and NLP, this thesis highlights their potential in advancing our understanding of mental health dynamics, ultimately promoting a more empathetic and effective approach to public health.

Declaration of Authorship

I, **Saima Rani**, declare that the Master of Research thesis entitled '**AI as a Mirror to the Mind: Analyzing Mental Health Narratives in the Social Media Landscape**' is no more than 50,000 words in length including quotes and exclusive of tables, figures, appendices, bibliography, references and footnotes. This thesis contains no material that has been submitted previously, in whole or in part, for the award of any other academic degree or diploma. Except where otherwise indicated, this thesis is my own work.

I have conducted my research in alignment with the Australian Code for the Responsible Conduct of Research and Victoria University's Higher Degree by Research Policy and Procedures.

Signed: [redacted]

Date: 17-04-2024

Ethics Declaration "All research procedures reported in the thesis were approved by the Victoria University Human Research Ethics Committee on 29 May 2023 (**Approval ID: HRE23-005**)."

Signed: [redacted]

Date: 17-04-2024

Submitting Thesis with Publication

Declaration by **Saima Rani**:
Signature: [redacted] Date: 17-04-2024

Chapter No.	**Publication Title**	**Status**	**Publication Details**
4	From Posts to Knowledge: Annotating a Pandemic-Era Reddit Dataset to Navigate Mental Health Narratives	Published	Feb 2024, Applied Sciences 14, no. 4: 1547
5,6	AI as the Interpreter: Deciphering Mental Health Stories in the Social Media Realm	Submitted	April 2024, Journal of Medical Internet Research (JMIR)

Declaration of Co Authorship Appendix: A

Acknowledgements

First and foremost, with the deepest sense of gratitude towards the Almighty, Allah, the Most Gracious and Merciful, whose infinite blessings and guidance have illuminated my path throughout this journey. It is with profound reverence that I acknowledge His omnipresent support, which has empowered me to undertake and fulfil this academic endeavour.

At the very heart of my journey lies my husband, the bedrock of my support system. His unwavering faith in me and relentless encouragement have been nothing short of inspirational. His unconditional love, patience, and motivation have been the cornerstone of my perseverance. During the peaks and troughs of this journey, his belief in my capabilities never waned; instead, he propelled me forward, urging me to embark on this degree. For his love, sacrifice, and the countless ways he has uplifted me, I am eternally grateful.

Special acknowledgement is due to Dr. Khanadkar, whose mentorship and expertise have been nothing short of a blessing. Dr. Khanadkar has been a guiding light throughout this journey, offering tireless support, invaluable advice, and profound insights at every stage. His readiness to answer even the most basic questions with patience and clarity made the daunting task of research approachable and meaningful. Dr. Khanadkar's role as a supervisor extended beyond academic guidance; his empathy and support during challenging times were remarkable. His dedication to my success and well-being has left an indelible mark on my professional and personal development.

I also wish to express my appreciation to Dr. Sudha for her guidance at the outset of my research journey, particularly in the collection of data, which laid the groundwork for my study.

I am also immensely grateful to Dr. Manjula for her invaluable insights into mental health research and for her generous sharing of expertise. Her willingness to train me in annotation and to introduce me to new knowledge areas, always being available and responsive, has enriched my research experience profoundly.

To Dr. Ross, Dr. Ron, and Dr. Deborah, I extend my heartfelt thanks for being the pillars of my research journey. Your foundational support and encouragement have been crucial in navigating this academic venture. Dr. Deborah, in particular, deserves additional gratitude for her assistance with the ethics application process. Her guidance was instrumental in navigating the complex ethical considerations of my study, ensuring its integrity and compliance.

Furthermore, I am grateful for the support and resources provided by the staff at Victoria University. Their assistance played a crucial role in facilitating my research activities and enhancing my academic experience.

A heartfelt Thanks to my family, whose love and understanding have been unwavering support of this journey. A special word of gratitude to my mother-in-law, whose encouragement have been invaluable. Her understanding and support, especially when I missed calls or postponed conversations, has not only been gracious but profoundly uplifting. My sister and brother-in-law deserve a distinctive acknowledgement for their culinary magic, providing comfort and nourishment during those endless hours spent in front of the computer.Their thoughtfulness has been a delicious and much-needed respite from the rigors of research. And to my nieces, whose anticipation for the completion of this degree reminds me of the joy and excitement that awaits beyond the academic endeavour. Their appreciation for every milestone achieved has made me feel as though I've accomplished something truly significant, infusing this journey with moments of lightheartedness and joy. Their eager waiting has been a constant reminder of the importance of balance and the value of dedicating time to family and loved ones. To my entire family, who adjusted their plans and supported my commitment to this research, I am eternally grateful.

Despite the collective efforts and the achievements celebrated within these pages, it's with a heavy heart I reflect on the ongoing struggles and unbearable pain faced by the people in Gaza. The stark contrast between the pursuit of academic and professional milestones and the harsh realities of conflict is deeply troubling. It serves as a sobering reminder that while we aim to reach beyond the skies in our endeavours, the shadow of unfairness cast by such conflicts remains a stark reality. This juxtaposition has been a source of profound disturbance and pain, underscoring the urgent need for empathy, understanding, and action in our global community to address and heal the wounds of such injustices.

In conclusion, this thesis is a mosaic of the contributions, support, and faith of all the individuals mentioned above. Your collective belief in me has been the wind beneath my wings, enabling me to soar higher than I ever imagined. Thank you, from the bottom of my heart.

List of Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
AutoML	Automated Machine Learning
HRE	Human Research Ethics
IAA	Inter-Annotator Agreement
IRR	Inter-Rater Reliability
LIME	Local Interpretable Model-agnostic Explanations
LLM	Large Language Model
MHEC	Mental Health Evaluation Corpus
MH-RAC	Mental Health-Reddit Annotated Corpus
ML	Machine Learning
NLP	Natural Language Processing
PaLM	Pathways Language Model
PRAW	Python Reddit API Wrapper
RMHD	Reddit Mental Health Dataset
SHAP	SHapley Additive ExPlanations
STAT	Semantic Text Annotation Tool

Contents

List of Figures

List of Tables

List of Publications

1. Rani, Saima, Khandakar Ahmed, and Sudha Subramani. 2024. **"From Posts to Knowledge: Annotating a Pandemic-Era Reddit Dataset to Navigate Mental Health Narratives"** Applied Sciences 14, no. 4: 1547. https://doi.org/10.3390/app14041547

2. Rani, S.; Ahmed, K.; O'Connor, M. 2024 **"AI as the Interpreter: Deciphering Mental Health Stories in the Social Media Realm,"** Journal of Medical Internet Research (JMIR). (Submitted)

Chapter 1

Introduction

In an era where AI and mental health intersect with unprecedented complexity, this thesis investigates the power and limitations of AI in sifting through mental health narratives on social media. Mental health, as defined by the World Health Organisation (WHO), is a state of well-being in which an individual realises his or her own abilities, can cope with the normal stresses of life, can work productively, and is able to make a contribution to their community [1]. This exploration emphasises mental health narratives as detailed personal accounts shared on social media that provide deeper insights into individual mental experiences, distinguished from more general or superficial posts. However, the term "post" is used synonymously with "mental health narratives" in this thesis. Chapter 1 sets the stage for this exploration, tracing the evolution of AI as a vital tool in mental health diagnostics and establishing the study's aims and motivations. It presents the guiding hypothesis and research questions, highlighting the potential contributions to knowledge in integrating AI with mental health care.

The significance of this research is articulated, emphasising its relevance to societal well-being. The scope of the study is clearly delineated, covering data sources, the study's time frame, the involvement of domain expert, the technological tools used, and the study's methodological boundaries. It also addresses potential limitations and ethical considerations, ensuring a responsible approach to research. Finally, the thesis outline provides a road map of the chapters that follow, setting the stage for a deep dive into the complexities and potential of AI in enhancing our understanding of mental health narratives on social media.

1.1 Background: Tracing the Pathways

Mental health disorder is one of the most common illnesses worldwide [2], which has the strong correlation with suicide [3]. According to WHO mental health issues cost approximately USD 2.5 trillion in 2010 with an estimated increase of USD 6.0 trillion by 2030, as more than 350 million people are impacted by depression [4]. In Australia only, every 1 in 2 face mental illness and approximately 3000 people end their lives [5]. From 2011 to 2021, suicide rates in males and females increased from 16.2 to 18.6 and from 5.1 to 5.8 deaths per one hundred thousand respectively [6]. These statistics highlight the global impact of mental health problems and underscore the necessity for new modes of intervention and prevention strategies to mitigate mental illness and reduce suicides.

Globally, approximately one in eight individuals grapple with some form of mental disorder, including but not limited to anxiety, depression, Post-Traumatic Stress Disorder(PTSD), and bipolar disorder [7]. The genesis of these disorders often extends beyond the surface, rooted deeply in complex causes. These disorders are not merely the product of immediate or obvious factors but are often tethered to profound root causes that require careful exploration and understanding. For instance, anxiety is frequently a culmination of past life experiences, inherent personality traits, or recent events, rather than merely a reaction to immediate stressors. This complexity underlines the importance of a comprehensive approach to treatment and management, which goes beyond addressing symptoms or immediate triggers to uncovering and addressing root causes.

Recognising the multifaceted nature of mental health disorders, Health Direct Australia identifies several key root causes [8].These include:

- Personality Factors;
- Drug and Alcohol Abuse;
- Trauma and Stress Factors;
- Early Life Environment;
- Biological Factors;
- Genetic Factors.

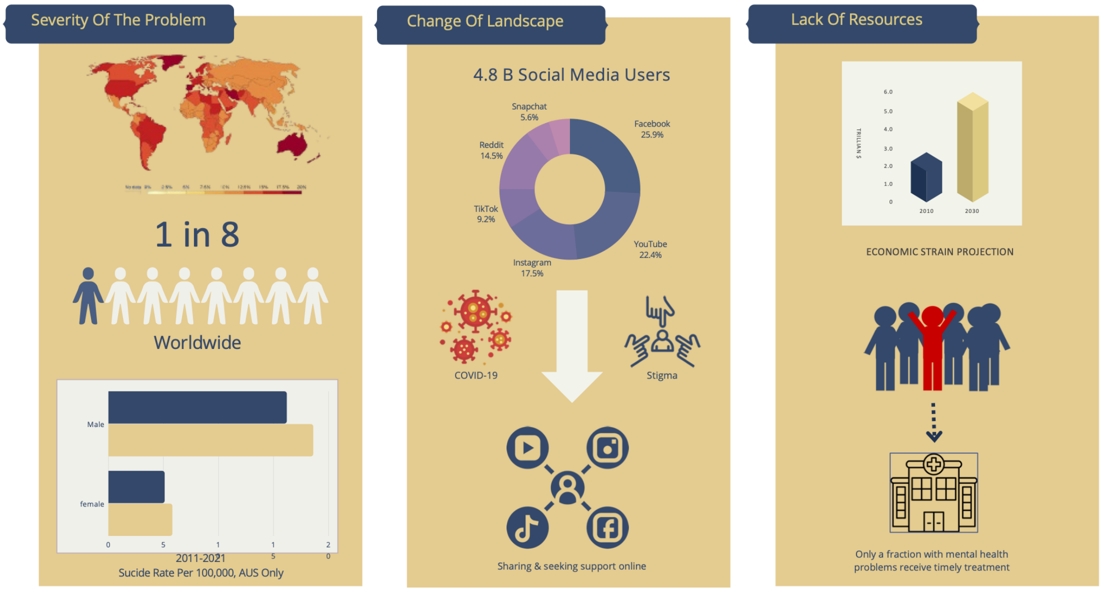


FIGURE 1.1: Why This Research Matters : Rising From The Challenges

These factors (Figure 1.1) illustrate the necessity of a comprehensive treatment approach and underscore the significance of our research in identifying these root causes through social media analysis.

Our research introduces a novel approach by leveraging the vast repository of personal narratives on Reddit. We are creating a comprehensive corpus to capture diverse experiences expressed in social media posts and developing an innovative NLP-based method to systematically identify and analyse these root causes. While existing research has made substantial strides in detecting mental health issues using ML on social media [9, 10], a gap remains in providing actionable information about root cause identification. Our study aims to bridge this gap, offering invaluable insights into the triggers of mental health disorders.

The early detection of root causes is crucial for preventive measures [11]. Traditional methods of diagnosing mental health disorders, often employed in clinical settings, include the use of structured interviews, psychological testing, and symptom observation. These approaches, though widely recognised, present several limitations. For instance, they can be influenced by subjective biases of the clinicians, require significant time and resources, and may not effectively capture the complexity and variability of mental health conditions across different populations. Additionally, such methods may not adequately account for cultural and social factors that influence mental health, leading to potential misdiagnoses or under diagnoses [12]. These methods typically detect illnesses after development, with a significant portion of affected individuals not seeking treatment [13]. The rising frequency of undetected cases and the limitations of current diagnostic approaches highlight the urgent need for more proactive strategies [14].

The advent of social media offers a revolutionary perspective in analysing mental health disorders. In today's interconnected world, the visibility of mental health concerns is unprecedented, with social media platforms serving as a unique lens into the daily stresses and worries of individuals globally. This shift underlines the importance of understanding the root causes of mental health issues beyond surface-level symptoms, emphasising the need for a comprehensive approach that delves into the complex web of factors influencing mental health, including personality traits, early life experiences, and recent events.

By assembling a unique corpus of social media text and employing NLP-based methods, our research aims to systematically identify and analyse the underlying factors of mental health disorders. This innovative approach not only advances the understanding of mental health detection through ML but also fills a critical gap in actionable root cause identification, offering invaluable insights into the triggers of mental health disorders. The alarming rates of suicide, undetected cases, and the lack of effective early prevention strategies further underline the urgency and potential impact of our study.

The innovative approach of our study is segmented into two distinct but interconnected parts, each designed to leverage the power of social media data in understanding mental health disorders. The first part focuses on creating a comprehensive corpus of social media text, meticulously annotated to capture the intricacies and nuances of discussions surrounding mental health. This corpus is not just a collection of data; it is a window into the lived experiences of individuals, offering unprecedented insights into the root causes of mental health issues. By assembling this corpus, our research aims to pave the way for advanced models that can more effectively identify and understand the multifaceted nature of mental health disorders.

In parallel, the second part of our research introduces a sophisticated framework that employs cutting-edge NLP. We then compared the models' predictions to the discerning assessments of experienced human evaluators to delve into the intricacies present within mental health narratives. This framework isn't merely a tool for data analysis; it's a means to bring structure and clarity to the vast, often chaotic world of social media discussions. By classifying and analysing the data within our corpus, the framework seeks to uncover the underlying patterns and factors contributing to mental health disorders. Moreover, the analytical capabilities of our framework offer valuable insights into societal sentiments and concerns.

In sum, our research represents a bold step forward in the intersection of technology, mental health narratives. By harnessing the vast potential of social media data through AI, we aim to uncover the root causes of mental health disorders, offering new pathways for diagnosis. The breadth of its applications underscores the transformative potential of our work, promising not only to advance our comprehension of mental health disorders but also to empower more informed, effective responses to the psychological challenges facing society today.

1.2 Aim and Motivation

The overarching aim of our research is to devise a framework that navigates the multifaceted origins of mental health disorders by leveraging the untapped potential of social media data. Recognising the complexity of these disorders, our approach seeks to construct a semantically rich corpus from firsthand narratives shared across various social media platforms. This endeavour is particularly focused on capturing the evolving sentiments surrounding mental health in different temporal pre-pandemic, mid-pandemic, and post-pandemic context. By harnessing NLP, we seek to tap into this vast repository of personal narratives, thereby analysing mental health sentiments in depth.

Our objectives extend beyond merely assembling this corpus. A thorough review of existing literature reveals that while previous studies have effectively employed ML to detect mental health issues in social media content, there is a notable gap in these methodologies: the in-depth analysis of root causes behind these disorders is often overlooked. Recognising this, our research sets out on a unique path to expand upon these existing methods. We focus not just on detection but also on the identification and understanding of the underlying root causes of mental health disorders.

This dual approach, combining the creation of a semantically profound corpus with a subset of meticulously annotated dataset, promises to provide actionable insights into the triggers and origins of mental health issues as reflected in social media narratives.

By doing so, we aspire to pave the way for more accurate diagnostic pathways and more effective intervention strategies. Ultimately, this study is designed to contribute to the development of a more nuanced understanding of mental health disorders, facilitating early diagnosis and tailored treatment approaches that account for the complexity of these conditions.

Our motivation stems from several key observations and gaps in the current scenario:

Prevalence and Impact: Recognising the widespread impact of mental health issues worldwide and their significant correlation with increased suicide rates and economic burdens, there is a critical need for innovative approaches to mental health diagnosis and treatment.

Limitations of Traditional Methods: Traditional diagnostic methods often fail to detect mental health disorders at an early stage, leading to a high number of undetected cases and individuals not receiving necessary treatment. This gap highlights the urgent need for more proactive and comprehensive diagnostic tools.

The Role of Social Media: The ubiquitous nature of social media and its role as a reflection of societal and individual mental states present a novel opportunity for mental health research. Social media offers a rich dataset for analysing the root causes of mental health issues, providing a new avenue for early detection and intervention.

Advancements in NLP and ML: The advancements in NLP and ML offer unprecedented opportunities for analysing complex datasets. We are motivated to leverage these technologies to interpret social media data in ways that were not previously possible, with the goal of filling gaps in root cause analysis and the early detection of mental health disorders.

Potential for Broad Impact: The possibility of applying our research findings across various sectors, including healthcare, public health policy, and even economic and political analysis, motivates us to explore this untapped potential. The ability to understand and monitor societal sentiments and the psychological impacts of significant events can inform more effective public health strategies and interventions.

1.3 The Compass Points

1.3.1 Hypothesis

Following the elucidation of our aims and motivations, this section delves into the core hypotheses and research questions that will guide our investigative journey. Grounded in the premise that social media narratives hold untapped potential for understanding mental health disorders, our research seeks to bridge the gap between the wealth of unstructured data available on these platforms and the nuanced understanding required to address mental health issues effectively. By employing advanced AI and NLP technologies, we strive to decode the complex interplay between societal changes, individual experiences, and their manifestations in mental health trends. The following hypotheses and research questions are formulated to systematically explore this intricate landscape, aiming to uncover insights that are both profound and actionable. Through this endeavour, we aspire not only to advance the field of mental health diagnostics but also to contribute to the development of more empathetic, responsive, and tailored intervention strategies.

Hypothesis 1: Social media platforms provide a rich source for understanding the root causes of mental health issues. The format of social media posts enables in-depth textual sharing of mental health narratives,thereby creating a dataset not attainable through traditional surveys. This wealth of textual data of human emotions, offers a unique opportunity to develop NLP models capable of accurately identifying mental health root causes. Furthermore, we hypothesise that these root causes may demonstrate shifts across different time periods, particularly pre-pandemic, pandemic, and post-pandemic. By utilising a meticulously annotated subset of this dataset, we aim to trace and accurately identify the evolution of these root causes over time with enhanced precision.

Hypothesis 2: NLP will show a significant correlation with human judgements in identifying the root causes of mental health issues from social media posts. This correlation may vary across the expertise levels of human raters, from domain experts to laypersons, suggesting that AI models can effectively augment human expertise by providing preliminary assessments. This points to areas where AI can complement human cognitive and emotional understanding and where it may require further refinement.

Hypothesis 3: The interpretability of AI models in discerning mental health issues from social media will exhibit distinct patterns of alignment and divergence from human cognitive processes. Through a comprehensive performance analysis of AI's error patterns and interpretability discrepancies, we expect to identify critical areas where AI either surpasses or falls short of human-like reasoning. This exploration will not only reveal the nuanced capabilities of AI in understanding complex mental health narratives but also highlight specific domains for enhancement. The outcome will guide the refinement of AI models to achieve a deeper, more nuanced comprehension of mental health narratives, aiming for a synergy that enhances both AI's diagnostic precision and its ethical integration into mental health care.

1.3.2 Research Questions

RQ-1: How can a semantically comprehensive corpus be created that reflects mental health sentiments across pre-pandemic, pandemic, and post-pandemic periods? and what methodologies can be employed to accurately annotate and categorise mental health-related discussions on social media?

RQ-2: How effectively can advanced AI models identify the root causes of mental health issues from social media posts, and to what extent do their classifications align with the nuanced judgements made by human raters of varying expertise in mental health?

RQ-3: What aspects of AI model interpretability are crucial for extracting meaningful insights from mental health discussions on social media, and what key patterns of errors and discrepancies in AI model performance can inform the enhancement of AI proficiency and the development of more effective mental health diagnostic tools?

Through this rigorous inquiry, we endeavour to illuminate the pathways through which AI and social media can be collaboratively leveraged to advance our understanding and treatment of mental health disorders, ultimately fostering a more informed, nuanced, and compassionate approach to mental health care.

1.4 Contribution to Knowledge

This research strive to make substantive contributions to the existing body of knowledge in the intersection of mental health diagnostics, artificial intelligence, and social media analytics. By addressing the hypotheses and research questions previously articulated, we anticipate advancing the field in the following critical areas:

Firstly, the development of a semantically comprehensive corpus, designed to encapsulate the myriad of factors and sentiments related to mental health across different temporal phases (pre-pandemic, pandemic, and post-pandemic), offers a valuable resource for researchers and practitioners. This corpus will not only serve as a tool for understanding the longitudinal effects of global crises on mental health but will also provide a template for how unstructured data from social media can be systematically harnessed and analysed.

Secondly, the methodologies proposed for accurately annotating and categorising social media discussions related to mental health are poised to introduce a new benchmark in qualitative data analysis. By refining and validating these methodologies, our research enabled a more nuanced analysis of mental health narratives, contributing to more personalised and timely mental health support.

Thirdly, the investigation into the efficacy of advanced AI models, in identifying the root causes of mental health issues represents a leap forward in the application of ML. Our research provides empirical evidence on the extent to which these AI systems align with human expert judgements, thereby establishing a foundation for AI-assisted diagnostics that can operate alongside human practitioners.

Moreover, by comparing the interpretability of AI models with human reasoning, this study illuminates the specific areas where AI models are more or less effective at interpreting mental health content. This will enhance not only our understanding of AI's capabilities and limitations but also guide the development of AI tools that are more closely aligned with human cognitive processes.

Finally, the comprehensive performance analysis of AI models, focusing on error patterns and discrepancies with human raters, yields insights that are vital for the improvement of AI diagnostic tools. By identifying the key areas where AI models under perform, we can inform the direction of future AI development to ensure that these tools become more effective, nuanced, and ethically responsible. Such advancements will significantly contribute to the evolving landscape of mental health diagnostics and intervention strategies, ensuring that they keep pace with the changing social dynamics and technological capabilities.

Through these contributions, our research stands to substantially enrich the academic dialogue and practical applications in the realms of mental health and artificial intelligence, fostering a more empathetic and responsive approach to mental health care in the digital age.

1.5 Statement of Significance :The Why Behind the What

The implications of this thesis extend far beyond the academic sphere, holding profound significance for the broader landscape of mental health care, technology policy, and societal well-being. As mental health issues continue to surge globally, exacerbated by crises such as the COVID-19 pandemic, the need for innovative, expandable, and accessible diagnostic tools has never been more critical. This research, situated at the confluence of AI, social media analytics, and mental health diagnostics, offers timely and crucial insights with several key areas of significance:

Enhancing Mental Health Diagnostics: By leveraging a novel dataset derived from social media and the methodologies for manually annotating mental health narratives , this research provides unprecedented insights into the root causes and evolving nature of mental health issues. The ability to identify these causes accurately and in real-time represents a significant leap forward in mental health diagnostics, offering the potential for more proactive and preventative care strategies.

Innovating Public Health Strategies: The analysis of mental health trends across different periods, especially surrounding the COVID-19 pandemic, offers valuable information for public health officials and policymakers.This also equips mental health professionals with a nuanced understanding of how global events impact mental health. Our findings can inform targeted mental health interventions, support services, and communication strategies, ultimately contributing to more resilient public health systems in the face of future crises.

Augmenting Diagnostic Accuracy with AI: The exploration of advanced AI models in identifying mental health issues paves the way for AI-assisted diagnostic tools that complement human expertise. These tools promise to increase the accessibility and timeliness of mental health support, particularly in under served or remote areas where healthcare resources are limited.

Enhancing AI Ethics and Human-Centric Design: By identifying the strengths and limitations of AI models in interpreting human emotions and psychological states, this research offers guidelines for the responsible advancement of AI technologies, ensuring they serve to augment human expertise rather than replace it. It advocates for a human-centric approach in AI design, ensuring that these technologies are sensitive to the complexities of human emotions.

Empowering Mental Health Professionals: The insights gleaned from this study are invaluable for mental health professionals, educators, and researchers. They provide a deeper understanding of the complex interplay between societal events and mental health, aiding in the development of more effective, empathetic, and nuanced intervention and care strategies.

Guiding Future AI Development: The comprehensive analysis of AI model performance, including error patterns and discrepancies with human judgements, informs the future direction of AI development. Insights gleaned from this research will ensure that subsequent generations of AI diagnostic tools are more effective, nuanced, and aligned with ethical standards.

In essence, this research holds significant implications for the advancement of mental health diagnostics, the development of AI tools in healthcare, and the formulation of effective public health policies. Its contributions are poised to benefit a broad spectrum of stakeholders, from technologists and healthcare professionals to policymakers and mental health researchers, ultimately fostering a more informed and compassionate approach to mental health intervention and care in the digital age.

1.6 Scope of the study

This research is dedicated to advancing the understanding of mental health issues through the lens of social media narratives, employing a combination of NLP and AI technologies. The scope of this study is defined along several dimensions:

1.6.1 Data Source and Selection

Our study exclusively utilises data from Reddit, a platform chosen for its rich and diverse discussions on mental health. The focus is on English-language posts to maintain consistency in language analysis, spanning from January 2019 to August 2022. This period includes pre-pandemic, pandemic, and post-pandemic phases, allowing for a comprehensive analysis of how mental health discussions have evolved in response to the COVID-19 pandemic.

1.6.2 Temporal Scope

While the primary focus of this research is on the discourse related to mental health issues on social media, the temporal boundaries are deliberately chosen to capture the fluctuations in mental health narratives influenced by the COVID-19 pandemic. This stratification offers an opportunity to observe potential shifts in mental health concerns and the ways they are expressed on social media platforms. The data from the pre-pandemic period serves as a baseline, enabling the exploration of the general mental health landscape before the unprecedented impact of COVID-19. The "during-pandemic" period provides insight into the immediate reactions and adjustments to the global crisis, highlighting the psychological stresses and strains of the pandemic. The "post-pandemic" period (or ongoing recovery phase at the time of this study) allows an investigation into how the mental health narrative might have evolved, sustained, or rebounded in the wake of the pandemic.

1.6.3 Role of Domain Expert

A pivotal aspect of our study's scope is the significant involvement of a single domain expert in mental health. This expert contributed extensively at various critical junctures of our research. Initially, the expert was instrumental in developing the annotation guidelines to ensure that our dataset accurately captures the intricacies of mental health narratives. They also led the training of annotators, sharing crucial knowledge and insights to maintain the high quality and reliability of data annotation.

Throughout the annotation process, the domain expert played a vital role in conflict resolution and consensus building, thus ensuring the annotated dataset's integrity and validity. Moreover, their expertise was invaluable in validating the analyses and predictions made by our AI models. This not only served as a benchmark for assessing the models' performance against the backdrop of established mental health knowledge but also affirmed the scientific robustness and ethical integrity of our findings.

1.6.4 Methodological Scope

The methodology involves creating a semantically comprehensive corpus from social media data, annotated for key mental health-related factors and sentiments. The study further includes the development and evaluation of AI models for their effectiveness in analysing this data, comparing AI interpretations with human judgements, and conducting a performance analysis to identify areas for improvement in AI diagnostics.

1.6.5 Technological Tools

The research employs advanced AI models, alongside NLP techniques for data analysis. These tools are used to develop models capable of identifying, categorising, and interpreting the root causes of mental health issues from the unstructured data of social media posts.

1.6.6 Limitations

While the study aims to provide significant insights into mental health diagnostics through social media analysis, it acknowledges limitations such as the potential bias in self-reported data on social media, the exclusion of non-English posts, and the challenges in generalising findings across all social media platforms or demographics. Additionally, as detailed in Chapter 7, the dataset's potential to reflect transient emotional states, particularly in pandemic era, may not accurately represent long-term mental health conditions, necessitating caution in interpreting trends as indicators of lasting mental health changes. Furthermore, the use of AI models, introduces challenges due to the subjectivity of human-derived benchmarks and the rapid evolution of AI technology. The evaluation process, limited to a small number of raters, restricts the generalizability of our findings, highlighting the need for a broader and more diverse pool of raters to enhance the validity of our comparisons.

1.6.7 Ethical Considerations: Steering with Integrity

Ethical considerations are paramount, given the sensitivity of mental health data. The research adheres to strict ethical guidelines to ensure the privacy and anonymity of individuals' data used in the study, with a focus on minimising harm and maximising the benefits of the findings for mental health diagnostics and interventions. This study received approval from the Victoria University Human Research Ethics Committee on 29 May 2023 (ID: HRE23-005), ensuring compliance with all relevant ethical standards.

1.7 Thesis Outline

The rest of the thesis includes the following chapters:

Chapter 2: Reviews relevant literature on mental health analysis via social media, AI and NLP applications in healthcare, and the integration of domain expertise in research. Identifies gaps your study aims to address and lays the theoretical groundwork for your research approach.

Chapter 3: Describes the conceptual framework, research design, chosen social media platform (Reddit), and time frame (pre-pandemic to post-pandemic). Discusses the domain expert's role and ethical considerations in the study.

Chapter 4: Dedicated to delineating the methodologies employed to tackle Research Question 1 (RQ 1), explaining the procedures for data collection from Reddit, the development of annotation guidelines, and the annotation process led by a domain expert. This chapter goes into detail about the training of annotators and the consensus-building process essential for ensuring the high quality of data, thereby laying the foundation for the AI models' training and subsequent analysis of mental health narratives on social media.

Chapter 5: Addresses Research Question 2 (RQ 2) by detailing the experimental setup, which includes the specifications of the AI models, their training, and testing protocols. It describes the methodology employed for conducting an error pattern analysis and a qualitative evaluation of the discrepancies between AI and human judgement in the context of analysing mental health narratives.

Chapter 6: Covers Research Question 3 (RQ 3), presenting the research findings related to the crucial aspects of AI model interpretability for extracting meaningful insights from mental health discussions on social media. It highlights AI's performance in identifying mental health root causes, its correlation with human judgement, and delineates the key patterns of errors and discrepancies in AI model performance that can inform the enhancement of AI proficiency and the development of more effective mental health diagnostic tools. Additionally, the chapter discusses the implications of these findings within the broader context of mental health diagnostics and AI development, emphasising areas where AI complements existing methods or requires further refinement for optimal utility.

Chapter 7: Summarises the key research outcomes, contributions to knowledge, and the study's significance. Reflects on the research's broader impact, potential applications, and suggests directions for future research to enhance AI tools in mental health diagnostics.

Chapter 2

Literature Review: AI, Social Media, and Mental Health Detection

Historically, a significant portion of research has been devoted to the detection of mental health disorders using ML, with a conspicuous lack of focus on the identification of root causes. The criticality of pinpointing the root causes of mental health issues cannot be overstated, as early identification can prevent the escalation into severe mental disorders.

The emerging role of technology in health science, particularly in the domain of mental health, necessitates a thorough understanding of existing literature. Specifically, it is crucial to comprehend how digital resources such as social media have been leveraged for mental health detection and how computational methods like NLP have been applied to mental health issues.

The burgeoning recognition of social media data as a rich source of information for the detection and analysis of mental health disorders has underscored the crucial role of NLP and ML techniques [15]. Notably, the application of these technologies presents a valuable tool for decision-makers, assisting in the formulation of preventive measures to mitigate the risk to vulnerable individuals [16]. Mental health investigation is a systematic process that requires meticulous collection and analysis of patient data to accurately measure indicators and improve mental health outcomes [17]. Timely and reliable information about triggers of mental health issues forms the cornerstone of successful illness detection.

This chapter provides an exhaustive review of the current literature to contextualise the existing body of work, emphasising the necessity and novelty of our approach. We delve into prior studies, examining how researchers have harnessed social media as a tool for mental health detection and how NLP has contributed to enhancing our understanding of mental health issues. The successes and limitations of these studies will be identified and analysed, highlighting the gaps in current research and demonstrating how our study fits within this broader academic conversation.

2.1 Mental Health Detection on Social Media

In the context of mental health, social media platforms have gained tremendous relevance. They serve as a digital forum where people from all walks of life share their feelings and experiences, including those associated with mental health. The COVID-19 pandemic has underscored the critical role of social media as a platform for individuals seeking support [18], highlighting the potential of such platforms for supporting and assisting individuals during their times of need. The necessity to harness online tools for public health observation and assistance has become an undeniable reality in our increasingly digital world. Moreover, it's worth noting that not all individuals are comfortable with traditional face-to-face clinical consultations. Many prefer to express their experiences and emotions in the more anonymous and accessible online environment.

As of April 2023, the global internet user base stood at 5.18 billion, accounting for 64.6 percent of the world's population. Among these, 4.8 billion individuals, representing 59.9 percent of the global population, were active users of social media [19]. Platforms such as Reddit, Twitter, and Facebook have paved the way for innovative research methodologies by providing a rich and diverse source of data. These platforms, due to their ability to capture diverse data on human experiences, have sparked interest in utilising them for public health observation and research. Through the analysis of user expressions and behaviours on these platforms, researchers can discern patterns and trends that may otherwise remain undetected. This interest stems from the aim to find practical solutions for real-world problems [20].

2.1.1 Social Media Role in Mental Health Support

The surge in social media popularity has opened up opportunities for it to serve as a viable platform for supporting individuals with mental health issues [21]. Numerous studies have shown that people with mental disorders use social media at rates comparable to the general population [22, 23]. This is particularly significant given that individuals with severe mental disorders are more likely to experience social isolation [24] and may prefer online connections. Majority of the younger generation with mental health issues admitted that social media usage take them out of the feeling of being alone [25]. The ease of interaction on these networks is believed to be beneficial for people with mental health issues [26]. Their online communication provide them a chance of interacting anonymously with each other without a fear of being judged when they are having highly denouncing health conditions [27]. Social media medium is also extremely helpful when individual with mental health disorders face elevated level of loneliness due to have limited social interaction in real life [28]. A study suggested that individuals with schizophrenia experience more difficulties communicating in person and interact more easily on social media [29].

People are using social media networks to express their feelings without any hindrance [30]. There is also an observable shift towards individuals discussing their symptoms and treatment experiences online, as opposed to traditional clinical settings [31]. For Instance, Naslund, et al.[32] have reported that how online platforms have been used by individuals with mental health illnesses to share their treatment experiences and seek support. Study suggested a conceptual model which was informed by existing studies to exemplify how online connections help people to deal with stigma and access for online interventions for their well-being. Such sharing and support dynamics have also been observed amongst users with schizophrenia, further underlining the potential of social media as a tool for mental health support [33]. Another research also supported the preference of seeking help online rather face to face among individual with mental health disorder [34].

The importance of considering the relationship between language expression and mental health cannot be overstated. The advent of social media has made it possible to observe and analyse this relationship in a manner that was inconceivable in the pre-digital era. It has opened a unique window into the minds of individuals, particularly those suffering from mental health disorders. The information divulged through their posts and patterns of social media usage offers a profound understanding of their mental states [35]. Research indicates that individuals with mental health disorders are active on social media platforms at a rate commensurate with the general population [23, 36]. This easy access to social media platforms provides an opportunity for individuals with mental health

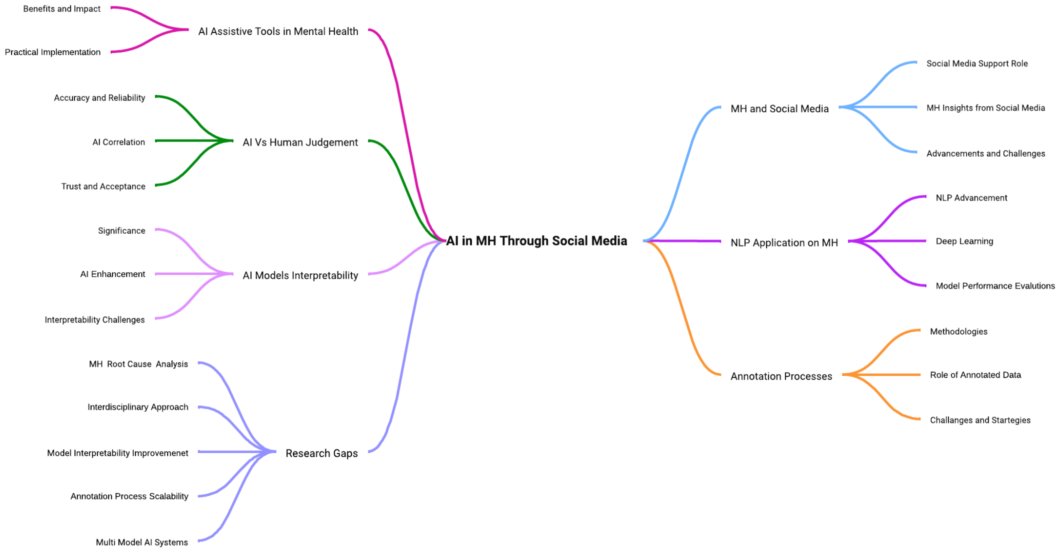

Benefits and Impact
Practical Implementation
AI Assistive Tools in Mental Health
Accuracy and Reliability
AI Correlation
Trust and Acceptance
AI Vs Human Judgement
Significance
AI Enhancement
Interpretability Challenges
AI Models Interpretability
MH Root Cause Analysis
Interdisciplinary Approach
Model Interpretability Improvemenet
Annotation Process Scalability
Multi Model AI Systems
Research Gaps
AI in MH Through Social Media
MH and Social Media
Social Media Support Role
MH Insights from Social Media
Advancements and Challenges
NLP Application on MH
NLP Advancement
Deep Learning
Model Performance Evalutions
Annotation Processes
Methodologies
Role of Annotated Data
Challanges and Startegies

Figure 2.1: Intersecting Paths of AI, Social Media, and Mental Health

disorders to express themselves freely and anonymously. This encourage researchers to avail this avenue to detect and analyse early risks of mental health disorders [37] that may otherwise go unattended and leads individuals to serious implications or suicidal ideation.

2.1.2 Social Media and Mental Health Research

Studies leveraging social media data have offered profound insights into mental health trends and individual experiences. For instances, Park et al. utilised Twitter data to compare self-reported depression levels with linguistic patterns in tweets, demonstrating the platform's validity in researching depression [38]. The research team observed that users on Twitter, a platform known for its succinct and immediate mode of communication, frequently discuss their mental health issues and even their treatment histories. This insight sparked a novel approach in their research design. They executed a cross-sectional analysis comparing self-reported depression levels from survey data with the linguistic patterns in tweets. For this, they employed Linguistic Inquiry and Word Count (LIWC), a widely used text analysis software that quantifies various linguistic dimensions in a text. The findings affirmed that Twitter data can indeed serve as a valid source for researching depression.

On the other hand, Coppersmith et al. examined the language used by social media users who had openly expressed their depression diagnostics in their tweets [39]. The study recognised the expressive and communicative nature of social media language as a powerful tool for mental health research. It underscored that the spontaneous and raw expression of emotions and experiences on social media provides a unique, unfiltered insight into individuals' mental health.

Furthermore, the applicability of this research is not limited to Twitter. Reddit, a popular discussion platform, has also been used extensively to understand and detect mental health issues. This platform is known for its numerous discussion forums or 'subreddits' dedicated to a multitude of topics, including mental health. De Choudhury et al. conducted a seminal study leveraging Reddit posts to detect depressive symptoms among users. By analysing language patterns and thematic content in posts from mental health-related subreddits, they developed a model that could predict depression with significant accuracy [40]. This study underscores the potential of Reddit as a rich data source for mental health research. Studies have examined posts discussing mental health issues that range from depression to self-harm. A multitude of text analysis methods, such as LIWC, N-gram, Latent Dirichlet Allocation (LDA), and emotional analysis, have been deployed to detect linguistic features and patterns in Reddit posts indicative of mental health issues [41–43].

2.1.3 Challenges and Advances in Social Media-Based Mental Health Research

In essence, studies leveraging social media data have collectively underscored the immense potential of these platforms as vital resources for mental health research. The digital age has ushered in innovative methodologies that utilise the rich tapestry of data available on platforms like Reddit, Twitter, and Facebook. These methodologies have illuminated the varied and complex ways individuals express their mental health struggles and seek support online.

However, the path to harnessing the full potential of social media for mental health research is fraught with challenges. The primary hurdle lies in the need for more sophisticated tools and techniques capable of effectively analysing and interpreting the massive volume of unstructured data generated on these platforms. Despite these challenges, the usage of social media data introduces a new dimension to mental health research, enabling a deeper understanding of the intricate relationship between language expression and mental health.

The collective findings from these studies not only highlight the opportunities presented by social media platforms as conduits for mental health research but also shed light on the necessity for advanced analytical tools and methodologies. The goal is to effectively glean significant insights from vast datasets, identifying patterns that reveal individuals' mental states. The importance of considering the relationship between language expression and mental health cannot be overstated. Social media has provided a unique window into the minds of individuals, particularly those suffering from mental health disorders, offering profound understanding and insights into their mental states.

Several studies affirm that individuals with mental health disorders are active on social media platforms at a rate commensurate with the general population. This accessibility allows individuals with mental health disorders to express themselves freely and anonymously, encouraging researchers to leverage this avenue to detect and analyse early risks of mental health disorders that may otherwise go unattended, potentially leading to serious implications or suicidal ideation.

As we navigate the future of mental health research in the digital era, it is clear that social media platforms will continue to play a pivotal role. The task ahead involves not only addressing the analytical challenges but also ensuring ethical considerations are at the forefront of our methodologies. By advancing our tools and approaches, we can unlock deeper insights into mental health, foster better support systems, and ultimately contribute to a more nuanced understanding of mental well-being in our increasingly connected world.

2.2 NLP Application on Mental Health Issues

With the prevalent use of social media among people with mental health disorder, presents an opportunity to leverage these platforms as tools to enhance public health research and improve the provision of mental health services [15]. The ability to detect mental health challenges early and intervene with appropriate treatment can significantly mitigate the suffering experienced by these individuals [44]. The emergence of ML predictive models and NLP techniques has now made it feasible to identify early indicators of mental health issues from social media posts [45, 46].

Expanding on the utilisation of social media, research has identified patterns in language use, post frequency, and even image content that correlate with various mental health conditions. For instance, analysis of Twitter data has revealed that changes in tweeting patterns and sentiment can indicate depressive states [47]. Similarly, Instagram posts have been analysed for visual indicators of depression, with findings suggesting that colour analysis, filter choice, and engagement metrics can serve as early warning signs [48].

2.2.1 ML and NLP Advances in Mental Health Detection

ML has emerged as a potent tool in the detection and diagnosis of mental health problems. It enables the learning of behavioural patterns, formulates predictions, and identifies symptoms of mental health disorders. Concurrently, advancements in NLP techniques have significantly contributed to proactive mental healthcare and diagnosis. The content available on social media has been found to be invaluable for detecting and analysing mental health issues using ML and NLP techniques [49]. Several studies have employed classification techniques to identify depression, stress, or suicidal ideation [50–53] and have examined the use of ML and NLP in social media datasets to gain deep insights through patterns [54]. Moreover, NLP and ML applications extend beyond detection and diagnosis into areas such as personalised therapy and support systems. Therapeutic chat bots, for instance, utilise NLP to provide real-time emotional support and cognitive behavioural therapy interventions, showing promising results in reducing symptoms of depression and anxiety [122]. This expansion into therapeutic applications highlights the potential for NLP and ML to not only identify mental health issues but also to offer scalable interventions.

2.2.2 Deep Learning in Mental Health Detection

The application of NLP in detecting mental health issues has experienced significant growth, particularly with the advent of deep learning techniques. From 2012 to 2021, the shift towards employing deep learning models has marked a pivotal change in research methodologies, making it a popular choice among scholars for its ability to handle complex linguistic patterns and vast datasets [55, 56].

The field has predominantly used supervised learning, with Support Vector Machines, Random Forest, Decision Trees, Naive Bayes, Adaptive Boosting (AdaBoost), Supervised Linear Discriminant Analysis (LDA), and Logistic Regression being among the most employed techniques [57–61]. These methods rely on labelled data to train models, which then learn to identify patterns indicative of various mental health conditions.

Conversely, unsupervised learning techniques have emerged as powerful tools for uncovering new insights from unlabelled data. These techniques are particularly valuable in mental health research for their ability to detect novel indicators of mental health issues without the need for pre-defined categories, allowing researchers to explore data more freely and discover unexpected patterns [62–64].

Deep learning, a subset of ML, has introduced models that excel in extracting meanings from complex, unstructured text data. Techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) networks, have been particularly effective in mental health applications. These models leverage various embedding techniques, including Global Vectors for Word Representation (GloVe), Word2Vec, and Embedding from Language Model (ELMo) to transform text into formats that deep learning models can process, leading to more accurate predictions of mental health states from social media content and other text sources [65–68]

The comparison of different ML methods has underscored the benefit of an ensemble approach, where combining multiple models can lead to improvements in prediction accuracy and reliability. This methodological innovation highlights the dynamic nature of NLP applications in mental health, where the integration of various models and techniques can address the limitations of individual methods and enhance overall performance[69–71].

2.2.3 Impact of Large Language Models (LLM)

In recent years, LLMs have demonstrated remarkable capabilities in various NLP tasks, as evidenced by numerous studies [72–78]. This success largely stems from the expansion of transformer-based models' training [79], where it's been observed that both model performance and efficiency in data usage improve with increases in model and dataset sizes [80]. Typically, LLMs undergo training through self-supervision on vast, showcasing notable achievements across an array of tasks, some of which demand specific scientific knowledge and analytical skills [81].

LLMs have proven to be invaluable as implicit knowledge bases in the medical field, despite the risks of generating misleading information, amplifying biases, and displaying reasoning limitations [82, 83]. Focusing on the medical and scientific domains, specialised models such as BioNLP [84], BioGPT [85] , PubMedBERT [86], and ScholarBERT [87] exemplify the successful adaptation of LLMs to the nuanced requirements of medical and scientific inquiry.

Though these models offer promising directions, their scale and scope are often narrower when compared to generalist LLMs like GPT-3 [72] and PaLM [73]. Despite their smaller scale, these models contribute uniquely to the medical industry by enabling applications that range from augmenting clinical assessments to summarising detailed medical communications [88, 89]. Echoing the foundational efforts of investigating the reasoning capabilities of LLMs in medical question-answering [90], our study integrates these advancements to further mental health research. We aim to improve analytical outcomes on mental health-focused datasets such as Med-PaLM and PubMedQA, showcasing the critical role of LLMs in advancing mental health diagnostics and interventions.

2.2.4 ML Models Performance in Mental Health Detection

The performance of ML models is evaluated by the accuracy level of their prediction ability of the correct class or label. Metrics such as recall, accuracy, precision, and F1 scores are commonly reported [91, 92]. Precision and recall indicate the usefulness and completeness of prediction respectively, while the F1 score combines precision and recall into a single metric by taking their harmonic mean. Accuracy indicates how many samples are accurately classified out of the total and how close the predictions are to the true value. In some studies, these metrics have also been used to score the severity of mental illness along with detection [93].

Ethical considerations are paramount when leveraging NLP and ML in mental health contexts, especially concerning social media data. Issues such as data privacy, consent, and the risk of stigmatisation must be meticulously managed. Developing transparent and ethical AI models, which incorporate patient and public involvement in their design and implementation, is critical for ensuring these technologies serve the best interests of individuals struggling with mental health issues [94].

In conclusion, NLP is aiding the mental healthcare system in various capacities using a range of techniques [95]. Applications based on NLP techniques have demonstrated a range of benefits from diagnosis, to support and treatment, and should be considered as supportive tools to reduce the spread of mental health illness.The ongoing evolution of these technologies, coupled with ethical safeguards, promises to further revolutionise mental health care, making it more accessible, personalised, and effective.

2.3 Annotation Processes for Model Training

In the realm of mental health research, with a special focus on the analysis of social media content, the procedure of data annotation emerges as a cornerstone in the training of AI models. This process is instrumental in enabling these models to accurately discern and analyse narratives related to mental health expressed across various social media platforms. The importance of data annotation extends across multiple dimensions, influencing not only the accuracy and effectiveness of AI models but also their ability to interpret complex mental health themes, thereby enhancing their utility in both clinical and research settings.

2.3.1 Importance of Annotated Datasets

The foundation upon which AI models are trained to identify mental health issues is largely constituted by high-quality annotated datasets. The precision with which these models can classify and interpret mental health-related data is directly linked to the quality of the annotations they are trained on. This training involves labelling social media posts with specific mental health themes, equipping the models with the capability to identify similar patterns in data they have not previously encountered. The studies by Gkotsis et al.[96] and the collaborative work of Ansari, Garg, and Saxena [97] underscore the critical role that meticulously annotated datasets play in enhancing model accuracy and reliability in identifying mental health issues.

2.3.2 Dataset Annotation Approaches

The strategies adopted to ensure the rigour and reliability of data annotations are varied, encompassing manual annotation by domain experts, automated initial labelling followed by human verification, and the iterative enhancement of annotations through active learning methods. This directly influence the effectiveness of AI models in identifying and interpreting mental health conditions from text data. These methods strive to balance accuracy, scalability, and the intricacies of mental health themes being studied. The works of Sarkar et al.[98] and Mohd, Jan, and Hakak [99] illustrate the diverse approaches to dataset annotation, highlighting the employment of advanced techniques to refine the process and improve the efficacy of AI models in mental health research. Researchers have developed various approaches to enhance the quality of annotations, ranging from expert annotation by clinical psychologists to crowdsourcing methods that leverage the wisdom of the crowd [100]. Moreover, the development of a user-centred, web-based crowdsourcing-integrated Semantic Text Annotation Tool (STAT) for building a mental health knowledge base presents an innovative method [101]. By leveraging crowdsourcing, STAT aims to efficiently gather labelled data from non-expert individuals, highlighting the potential for scalable and accessible data annotation in mental health research. Each approach has its merits and limitations. Expert annotations are typically more reliable but can be resource-intensive and subject to limited scalability. In contrast, crowdsourced annotations offer scalability and diversity of perspectives but may vary in quality [102].

2.3.3 Challenges in Annotating Social Media Posts

The annotation of social media posts for mental health research is fraught with challenges, including the informal and diverse nature of language, the varied expressions of mental health issues, and the need for contextual interpretation [103]. Additionally, the sensitive nature of mental health data demands careful consideration of privacy and ethical guidelines [104]. To navigate these challenges, researchers develop comprehensive annotation guidelines and train annotators to recognise a broad spectrum of mental health-related expressions. One approach involves the development of comprehensive annotation guidelines that account for the linguistic characteristics of social media and the specificities of mental health discourse [105]. Another strategy is the use of iterative annotation processes, where initial annotations are continuously refined through rounds of review and adjustment [106]. Moreover, hybrid models that combine ML with human annotation have shown promise in improving the efficiency and accuracy of the annotation process [107]. These models leverage ML to pre-annotate data, which is then refined by human annotators, thus combining the scalability of automated methods with the nuanced understanding of human experts.

Additionally, ML techniques are employed to assist in the identification of relevant posts for annotation. The studies by Muzafar et al. [108], and Tadesse et al. [109], demonstrate the strategic approaches taken to overcome the complexities inherent in social media data annotation, enhancing the ability of AI models to perform nuanced analysis of mental health narratives. A systematic review further illuminates the intricate steps from recognising mental health issues to accessing specialised care, emphasising the importance of effectively communicating and recognising mental health symptoms within primary care settings. This study reinforces the necessity for accurate and comprehensive data annotation to bridge the gap in presenting and recognising mental health symptoms [110].

The collective insights from these studies reveal the indispensable role of data annotation in the advancement of AI models for mental health analysis on social media. By diligently annotating social media datasets, researchers empower AI models to bridge the gap between technological capabilities and the sophisticated understanding required by human raters. This foundational effort not only aids in the development of effective AI tools for mental health analysis but also enriches our comprehension of the intricate relationship between mental health narratives and social media discourse, as highlighted by Lim et al.[111], and Pandey[112]. This body of work not only underscores the complexity of data annotation but also its critical importance in the intersecting fields of AI and mental health research. By adopting rigorous and reliable annotation practices and addressing the unique challenges posed by social media data, researchers can significantly enhance the performance of AI models. This, in turn, paves the way for advancements in the use of AI to understand and address mental health issues, contributing valuable insights to both the fields of artificial intelligence and mental health research.

2.4 Comparative Studies between AI and Human Judgement

The integration of AI in various domains, particularly in mental health settings, has sparked significant interest in understanding how AI-driven decisions correlate with human judgement. This section of the literature review explores studies that compare the outcomes of AI models with human judgements, focusing on their application in mental health, and examines the impact of expertise variability on annotation and interpretation accuracy.

2.4.1 AI Model Outcomes vs. Human Judgements

Recent literature indicates a growing examination of AI models' performance against human judgement across multiple domains, including healthcare, finance, and criminal justice [113, 114]. In healthcare, AI models have demonstrated capabilities that sometimes surpass human experts in diagnostic accuracy [115], though the consensus underscores the complementary role of AI to human expertise rather than replacement.

2.4.2 AI Classifications and Expert Decisions Correlation in Mental Health

In mental health settings, studies have specifically focused on the correlation between AI classifications and expert decisions. Miner et al. investigated the use of NLP and ML in assessing patient sentiments and emotions, finding a strong correlation between AI classifications and those made by mental health professionals [116]. Another study by Gkotsis et al. employed AI to classify social media posts according to mental health conditions, revealing that AI could match, and in some aspects, predict mental health conditions with an accuracy comparable to human clinicians [96].

2.4.3 Expertise Variability in Annotation Accuracy

The variability of expertise among annotators from clinical experts to laypersons, significantly affects the annotation and interpretation accuracy of both AI models and human judgements. Torous et al. highlighted that while expert annotations provide high-quality data for training AI models, layperson annotations offer scalability and insights into public perceptions of mental health [117]. Rezaii et al. further demonstrated that AI models trained on datasets annotated by experts in psychiatry did not only achieve high accuracy in identifying psychotic disorders but also revealed that models could discern nuances in language that were not evident to lay annotators [118].

Furthermore, studies such as Conway and O'Connor argue for the integration of AI in mental health to leverage the strengths of both expert and non-expert annotations, suggesting that a hybrid approach may enhance the overall accuracy and reliability of mental health assessments [100].

The comparative analysis between AI models and human judgement in mental health underscores the potential of AI as a valuable tool in augmenting professional healthcare services. The literature reveals that while AI models can approximate or even exceed human judgement in specific tasks, the variability in expertise among annotators plays a crucial role in shaping the effectiveness of AI applications. This variability not only influences the accuracy of AI model training but also impacts the interpretation and applicability of AI-driven insights in real-world mental health settings. As AI continues to evolve, further research is needed to explore how these technologies can be optimised to complement professional expertise, thereby enhancing mental health care delivery and outcomes.

2.5 AI Assistive Tools in Mental Health

The integration of AI in mental health care represents a significant advancement, offering potential to augment clinical decision-making, enhance patient care, and address ethical and privacy concerns. This section of the literature review delves into how AI models function as assistive tools in mental health settings, supported by case studies, ethical considerations, and future research directions.

2.5.1 AI as Assistive Tools for Mental Health Professionals

AI models, particularly those based on NLP and ML, have increasingly become valuable assets for mental health professionals. These models assist in various capacities, from early detection of mental health issues to personalised treatment recommendations [119]. Notably, AI-driven algorithms have shown promise in identifying early signs of mental health issues, such as depression, by analysing speech patterns and social media activity [95]. Their application ranges from enhancing diagnostic accuracy to providing scalable interventions, such as digital therapeutics and personalised patient monitoring systems.

2.5.2 Practical Implementations in Mental Health Care

Several case studies illustrate the impactful role of AI in mental health care. For instance, the use of predictive analytics in identifying patients at risk of depression or suicide has enabled early interventions, potentially saving lives [120]. Similarly, conversational AI agents or "mental health chat bots" have been introduced to provide instant psychological support, aiding in the management of anxiety and depression outside the conventional therapy sessions[121, 122]. There is also AI-enhanced monitoring tools have been used for patients with PTSD, leveraging ML to analyse physiological signals and predict stress levels [123].

Looking ahead, the development of more sophisticated AI models that better understand and interpret the complexities of human emotions and behaviours is a primary research focus. Enhancing the interpretability of these models and their alignment with human judgement remains a challenge but is critical for their effective integration into clinical practice [124, 125]. Moreover, advancing multi modal AI systems that integrate textual, auditory, and physiological data can offer a more comprehensive understanding of mental health conditions [126].

AI models stand as transformative tools in mental health care, with the capacity to augment clinical practices, support patient care, and navigate complex ethical considerations. As the field advances, ongoing research will be vital in refining these technologies to align closely with clinical needs and human judgement, ensuring their ethical application and maximising their benefit in mental health settings.

2.6 Interpretability and Transparency in AI Models

The deployment of AI in healthcare, and specifically in mental health applications, underscores a growing need for interpretability and transparency in AI models. These attributes are crucial for clinicians to trust and effectively utilise AI tools, ensuring that decisions made by AI systems are understandable and justifiable. This literature review section explores the importance of model interpretability, methods for enhancing transparency, and the inherent challenges in elucidating complex AI models.

2.6.1 Model Interpretability Significance in Healthcare

In healthcare, the interpretability of AI models is paramount for several reasons. Firstly, it facilitates the validation of model decisions by healthcare professionals, ensuring that AI recommendations align with clinical knowledge and practice [127]. Secondly, interpretability is essential for patient trust and acceptance, as patients are more likely to trust AI-assisted decisions if the rationale behind these decisions can be explained [128]. Lastly, regulatory compliance often necessitates transparency in decision-making processes, making interpretability a legal requirement in many healthcare applications [129].

2.6.2 Enhancing AI Model Transparency and Interpretability

A variety of methods have been developed to improve the transparency and interpretability of AI models. Feature importance techniques, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), provide insights into how individual features influence model predictions [130, 131]. Visualisation tools, like saliency maps for neural networks, help visualise the parts of input data most responsible for model decisions, enhancing understanding of model behaviour [132]. Moreover, model-agnostic methods offer the flexibility to apply interpretability techniques across different types of models, broadening the applicability of these approaches [133].

2.6.3 Challenges in AI Interpretability for Mental Health Applications

Despite advancements in interpretability techniques, making complex AI models, such as those based on deep learning, interpretable to humans remains a significant challenge. Deep learning models, often described as "black boxes," have intricate architectures that make their decision-making processes opaque [134]. The high dimensional and non-linearity of these models exacerbate the difficulty of translating model decisions into understandable explanations. Efforts to demystify these models involve both technical solutions, aimed at simplifying the model's architecture without sacrificing performance, and educational initiatives to improve human understanding of AI processes [135]. Furthermore, there is an ongoing debate about the trade-off between model complexity and interpretability, with concerns that increasing interpretability might come at the cost of model performance [136].

Interpretability and transparency in AI models are critical for their ethical and effective application in mental health care. While significant strides have been made in developing methods to enhance model interpretability, challenges remain, particularly with complex models. Future research must continue to address these challenges, seeking innovative solutions that do not compromise model accuracy for interpretability. Achieving a balance between these aspects will be paramount for the successful integration of AI in mental health settings.

2.7 Key Research Gaps

Root Cause Analysis: Our exploration of existing literature has illuminated significant gaps in harnessing the full potential of social media narratives for understanding mental health disorders. While previous studies have made strides in detecting mental health conditions through social media analysis, there remains an under explored area in identifying and understanding the root causes of these conditions. This gap underscores the need for a nuanced approach that goes beyond detection to explore the underlying reasons contributing to mental health issues.

Interdisciplinary Approach: There's a noted need for interdisciplinary research combining insights from mental health professionals, AI technologists, and linguists to enhance the interpretability and effectiveness of AI models in mental health applications. This challenge is significant in mental health applications, where understanding the "why" behind a model's decision is as critical as the decision itself. Collaborative efforts can lead to the development of more sophisticated models that align closely with clinical needs.

Improving Model Interpretability: Despite advancements in making AI models more interpretable, there remains a challenge in elucidating complex models like deep learning. Future research should focus on developing innovative methods to enhance the transparency and interpretability of these models without sacrificing performance.

Scalability of Annotation Processes: The literature review suggests that while expert annotation provides high-quality data, it lacks scalability. Future research should investigate scalable annotation methods that maintain high data quality, possibly through semi-automated annotation systems.

Multi modal AI Systems Enhancement: There's a potential for future research to develop multi modal AI systems that can analyse data from various sources, such as text, voice, and physiological signals, for a holistic view of mental health conditions. Future research could explore the development of multi modal AI models that can analyse this integrated data, offering a more holistic view of an individual's mental health state.

Through addressing these identified gaps, Future research should aim to significantly advance the field of mental health diagnostics. in our study, we aspire to uncover insights that are both profound and actionable by leveraging advanced AI and NLP technologies in conjunction with the rich narrative data available on social media platforms. This endeavour will not only enhance our understanding of mental health disorders and their root causes but also contribute to the development of more empathetic, responsive, and tailored intervention strategies, ultimately fostering a more informed, nuanced, and compassionate approach to mental health care.

Chapter 3

Research Design and Framework

This chapter delineates the comprehensive architecture of the research methodology employed in this study, establishing a structured framework for investigating the intricate dynamics of mental health narratives as discussed on social media platforms. In this chapter, we detail the research design to address our research questions, establishing a robust framework that integrates advanced AI techniques with a detailed exploration of mental health narratives shared on social media. The design and execution of the research are critical in ensuring that the findings are both valid and reliable, providing meaningful insights into the root causes of mental health issues and the potential of NLP in interpreting these complex phenomena.

3.1 Overview of the Research Design

The research design of this study is orchestrated around a mixed-methods approach, intertwining qualitative and quantitative analyses for a holistic examination of the research problem. Our primary objective is to evaluate how accurately AI models can identify and interpret the root causes within mental health narratives, distinguishing these foundational elements from everyday language across various temporal contexts. This involves a critical examination of AI's capacity to deeply understand and extract the underlying causes of mental health issues from social media discourse, leveraging AI models and in-depth analytical techniques. This strategy ensures we capture the true essence of mental health conversations in the digital era. Below, we outline the specific methodologies tailored to each of our research questions:

3.1.1 Creating a Semantically Comprehensive Corpus and Developing Annotation Methodologies (RQ-1)

To address the challenge of constructing a semantically rich corpus that encapsulates mental health factors and sentiments across pre-pandemic, pandemic, and post-pandemic periods, we embarked on an extensive data collection initiative. This initiative targeted narratives from Reddit, aiming to capture a broad spectrum of mental health discussions reflective of the changing dynamics due to the COVID-19 pandemic. The process involved selecting subreddits known for their active mental health communities and employing data mining techniques to gather posts from specified time frames. This approach ensures the development of a dataset that is both comprehensive and temporally nuanced, offering a fertile ground for subsequent analysis.

In response to the need for accurately annotating and categorising mental health discussions on social media, we crafted an annotation framework. This framework was informed by both existing literature and input from domain experts in mental health. Annotators, selected for their blend of technical skills and domain expertise, underwent rigorous training to ensure the reliability and consistency of the annotation process. This process aimed to categorise discussions into predefined root causes of mental health issues, thereby setting the stage for an in-depth analysis of the data.

3.1.2 Evaluating AI Models and Human Judgement and Interpreting AI Models Compared to Human Reasoning (RQ-2)

To determine the efficacy of advanced AI models, in identifying mental health root causes from social media posts, our study designed a comparative analysis between these models and human raters. This analysis focused on the models' ability to classify discussions accurately and how these classifications align with the nuanced judgements of human raters with varying levels of expertise. By training and testing these models on our annotated dataset, we seek to uncover the strengths and limitations of AI in mirroring human cognitive processes in the assessment of mental health narratives. To account for potential bias and enhance the validity of findings, the study employed two models and compare their outcomes.

Exploring the interpretability of AI models in the context of mental health narratives from social media involves a critical examination of how AI's understanding compares to human reasoning. This facet of our research investigates which aspects of mental health content are more or less effectively interpreted by AI models, shedding light on the potential capacity of AI in grasping the subtleties of mental health discourse.

3.1.3 Identifying Patterns of Errors in AI Models (RQ-3)

To enhance AI model proficiency and contribute to the development of more effective mental health diagnostic tools, our study closely examines the patterns of errors and discrepancies in AI performance. This analysis delves into the root causes of mental health issues as identified by AI models, aiming to pinpoint specific areas where models under perform or misinterpret data. Insights gained from this investigation are pivotal for informing targeted improvements in AI methodologies, guiding future research toward refining AI's capability to understand and interpret complex mental health narratives accurately.

Our research design, tailored to address key research questions, aims to advance mental health research and the application of AI and ML technologies. By ethically applying these technologies, we seek to enhance our understanding and approach to mental health issues as depicted in social media narratives. We employed method triangulation, combining quantitative model classifications with qualitative evaluations from human raters [144]. This approach bridges ML and social sciences, emphasising the need to understand context and complexity in mental health discourse. Incorporating triangulation strengthens our findings' reliability and aligns our study with best practices for a holistic and multidimensional inquiry. This methodology is vital in exploring the nuances of mental health through the lens of technology and human cognition.

In summary, our study contributes to mental health diagnostics and intervention discussions by using a comprehensive approach to research and validation. Our goal is to foster innovative therapeutic strategies and inform mental health care policy, underscoring the significance of integrating AI, mental health, and social sciences.

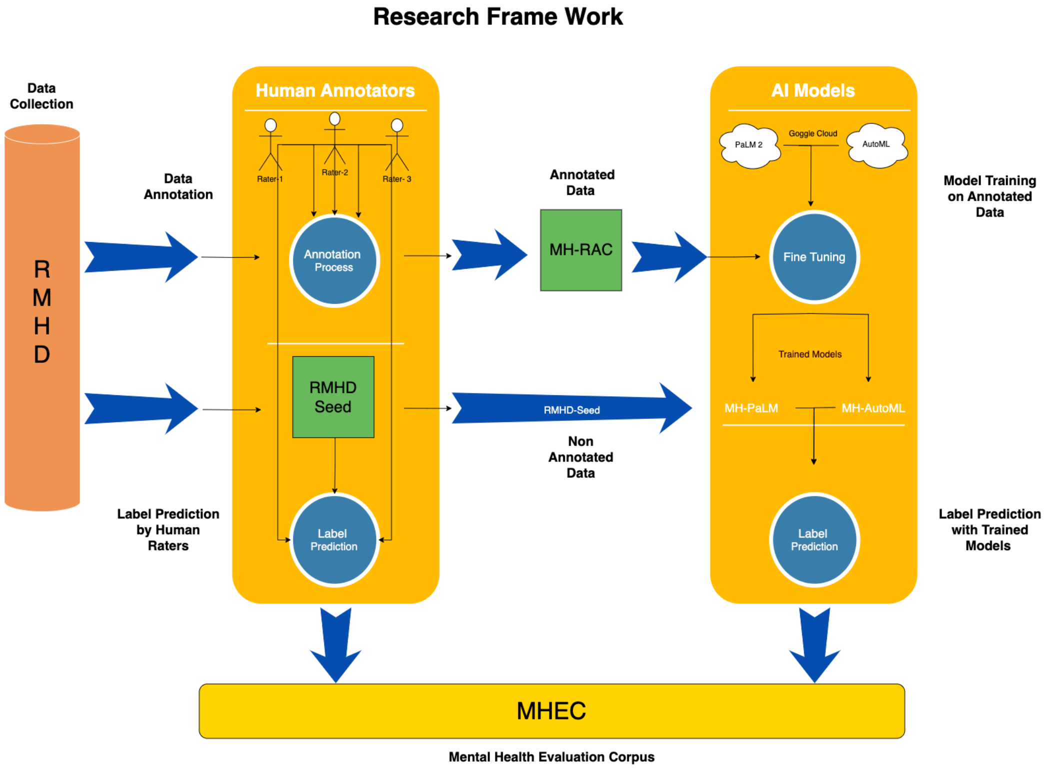


Figure 3.1: Research Framework

3.2 Conceptual Framework

The methodology adopted in this thesis is designed to navigate the intricate landscape of mental health narratives within social media through a comprehensive and methodologically sound approach. At the core of this research is a systematic and structured methodology divided into six pivotal stages, each contributing to the overarching goal of deciphering the root causes of mental health issues as expressed in natural language and emotional states on social media platforms. The sequential stages as shown in Figure 3.1, outlined below constitute the conceptual framework of this research, ensuring a rigorous and holistic analysis of mental health discussions online:

3.2.1 Data Collection

This initial stage involves meticulously gathering social media posts related to mental health from platforms such as Reddit. The collection process strategically focuses on acquiring a diverse and representative sample of narratives spanning pre-pandemic, pandemic, and post-pandemic periods. Inkster et al. highlight the importance of varied data sources in mental health analytics to ensure comprehensive understanding [145].

3.2.2 Data Annotation

Following data collection, the annotation stage involves labelling the raw data with specific tags that identify the root causes of mental health issues. This process is facilitated by a team of trained annotators who can recognise various mental health factors and sentiments. Their expertise ensures that the labelling is both accurate and reflective of the data's complexity as accurate annotation enhances model training and reliability in psychiatric research [175].

3.2.3 Pre-Processing

Prior to analysis, the collected data undergoes pre-processing to prepare it for model training. This stage involves cleaning the data, normalising text, and extracting relevant features that will aid in the accurate analysis of mental health narratives. The pre-processing ensures that the data is in a suitable format for the subsequent ML stages. Clean data and feature engineering are essential for robust and reliable outcomes in scientific studies involving data analysis, particularly in machine learning applications within health informatics [146].

3.2.4 Model Training

At this stage, AI models are trained on the pre-processed and annotated dataset. The training process is designed to equip the models with the ability to recognise the root causes of mental health issues as depicted in the social media narratives. Sophisticated modelling techniques are essential for extracting meaningful insights from complex datasets, ensuring that the models can accurately interpret the nuanced information present in social media data. This approach is crucial for developing reliable models that can predict mental health states effectively [147].

3.2.5 Performance Evaluation

This is the final stage where the performance of the trained models is evaluated using various metrics such as precision, recall, and F1-score. These evaluations help in understanding the strengths and weaknesses of each model and serve as a preliminary check to ensure that the models are well-trained and capable of identifying mental health root causes with a high degree of precision. Thorough validation practices, including the use of performance metrics, are crucial for deploying effective AI tools in clinical settings. For example, Muetunda et al. highlight the use of precision, recall, and F1-score in evaluating AI models for predicting mental health issues, emphasising their importance in understanding model effectiveness [148]. This approach provides initial insights into the models' readiness for deployment in real-world scenarios.

3.2.6 Evaluation Framework for Core Analysis

The culmination of the methodology is the core analysis stage, where various analyses are conducted to shed light on the models' capabilities and limitations.This includes a detailed exploration of how models interpret natural language and emotions, their alignment with human raters, and their overall efficacy in uncovering the underlying causes of mental health issues from social media content.

3.2.6.1 Human Raters Composition: The Human Element

The evaluation team is composed of a domain expert in mental health (Rater 1), a rigorously trained annotator (Rater 2), and a layperson (Rater 3), collectively encompassing a comprehensive range of interpretative perspectives [149].

Domain Expert: Rater 1 (R-1), the domain expert, is integral to the evaluation process [150]. With extensive experience in mental health, R-1 leverages deep expertise to authoritatively classify mental health narratives, identifying subtle diagnostic cues and therapeutic communication nuances.

Trained Annotator: Rater 2 (R-2), a member of the research team and trained annotator, has developed expertise through extensive training in our annotation protocols

[151]. Though lacking the clinical background of the domain expert, R-2 brings a methodically informed viewpoint to the data annotation process, ensuring a consistent and systematic classification approach.

Layperson: Rater 3 (R-3), the layperson, represents the general public's perception and identified by their lack of formal training in mental health. R-3's interpretations are particularly valuable for reflecting the broader audience's viewpoint and are crucial for assessing public mental health awareness, a key factor in designing effective interventions [152].

This diverse assembly ensures a comprehensive evaluation framework that integrates clinical accuracy, methodological reliability, and public understandability. The expert's deep clinical insight guarantees that nuanced health conditions are precisely identified, while the trained annotator ensures interpretations are consistent and aligned with established guidelines, yet adaptable to broader contexts. By considering the lay perspective, AI models can be better designed to serve a preventative role in mental health care, recognising and flagging narratives that signal a need for intervention, even when articulated in non-clinical terms. This inclusivity in model training aligns with the diverse range of individuals who may benefit from or interact with AI-assisted mental health tools, making these models more comprehensive and universally applicable. This comprehensive approach ensures that the AI models are well-rounded, sensitive to a variety of interpretations, and thus, more effective in real-world applications where individuals from all walks of life may rely on their analysis for support and guidance in mental health matters.

Together, these stages form the backbone of the research, providing a solid conceptual and methodological framework that guides the study towards achieving its objectives. By systematically following this framework, the research contribute valuable insights into the potential of artificial intelligence in enhancing our understanding of mental health dynamics as reflected in the digital realm of social media.

3.3 Ethical Considerations and Data Privacy

Maintaining the highest ethical standards and integrity is essential in all research fields, as these elements critically influence the study's reliability and validity. This study sourced data from Reddit posts, eliminating the need for direct participation from human subjects. Social media research ethics emphasise the principle of 'autonomy', advocating for informed consent from individuals whose posts are utilised [153]. However, the National Statement on Ethical Conduct in Human Research acknowledges the impracticality of obtaining consent from every social media user due to the vast number of posts [154].

Consequently, this research secured a waiver from the Victoria University Ethics Committee after a rigorous process, aligning with the National Health and Medical Research Council (NHMRC) National Statement on Ethical Conduct in Human Research (2007, updated in 2018). This study received approval from the Victoria University Human Research Ethics Committee on 29 May 2023 (ID: HRE23-005).

Reddit's pseudonymous system protects user privacy, ensuring the data collected excludes personally identifiable information and preserves user anonymity. Special care was taken to exclude any data mentioning ages within the 1-18 range, aligning with ethical guidelines protecting minors' privacy.

The research methodology was meticulously designed to respect these ethical considerations. The ethics committee's approval of our research plan underscores our dedication to the highest ethical research standards and integrity. This endorsement, following a comprehensive review of our ethical approach, data privacy measures, and efforts to ensure participant anonymity and confidentiality, underscores our commitment to responsible research practices. This approval not only marks a significant achievement but also reinforces our commitment to prioritising the welfare of individuals represented in the data we analyse.

Chapter 4

The Blueprint of Data: Harvesting Insights from Reddit

In the chapter, we explore the strategic use of Reddit as a potent data source for mental health research. This platform uniquely offers access to a diverse, voluminous, and anonymous dataset that provides profound insights into the nuances of mental health discourse. This chapter outlines the methodologies involved in selecting Reddit, harnessing its API for data extraction, and meticulously organising the resulting datasets, the Reddit Mental Health Dataset (RMHD), the Mental Health-Reddit Annotated Corpus (MH-RAC), and the Mental Health Evaluation Corpus (MHEC). A significant focus is placed on the comprehensive data annotation processes, where we discuss the compilation and refinement of data subsets, the implementation of stratified sampling, and the rigorous training of annotators. The chapter elaborates on the annotation guidelines, techniques for manual labelling, and strategies to ensure data quality and reliability through consensus-based validation and inter-annotator agreement measures. Our methodical approach aims to transform raw data into a structured and reliable resource that enhances the understanding of mental health trends on social media.

4.1 Reddit Mental Health Dataset (RMHD): The Digital Tapestry

The ascendance of social media data in recent years is indisputable, offering an unprecedented volume of real-time insights into human behaviour. This data, when effectively harnessed, empowers researchers to discern patterns, trends, and sentiments, leading to pioneering discoveries across various fields, including mental health. Recognising the potential of different social media platforms, A cross-platform analysis was performed to ascertain the most suitable social media platform for our research objectives. This step was instrumental in ensuring the robustness and relevance of data with our study.

4.1.1 Platform Selection: Choosing the Canvas

Selecting an appropriate social media platform is a critical initial step in conducting research in the digital domain. The chosen platform can significantly influence the type of data collected, the insights gained, and ultimately, the conclusions drawn. In the context of our research, we aimed to understand mental health patterns by analysing user-generated content. Therefore, it was crucial for us to select a platform that aligned with our research objectives.

We evaluated multiple platforms, each with its own set of advantages and potential limitations. Our selection process involved an in-depth analysis of various social media platforms based on several criteria, such as the nature and volume of the content, Mental health relevancy, privacy and ethical considerations, and the ease of data accessibility.

Facebook, a globally recognised social media platform, boasts a staggering user base of 2.9 billion people [155]. Many studies suggest that Facebook posts offer profound insights into human behaviour due to the platform's ability to facilitate ongoing user engagement [156]. Despite this potential, Facebook's inherent structure creates obstacles for researchers. A significant amount of content on Facebook is private and accessible only to the company [157], thereby limiting the amount of data available for research. This lack of access can hinder the development of comprehensive analyses.

Twitter, with approximately 556 million users [155], is another widely used platform that is often regarded as a rich resource for mental health research. Yet, the platform shows an unequal distribution of user engagement, with a small fraction of users contributing most of the content. For instance, one study reported that the top 10 percent of active users are responsible for 80 percent of all tweets [158]. This suggests that data derived from Twitter originates from a limited group of users and may not accurately represent the broader user population.

Contrastingly, Reddit, an anonymous community-oriented platform, presents a more promising avenue for the in-depth exploration of sensitive topics such as mental health.

The platform promotes open, candid, and extensive discussions, largely due to the anonymity it affords its users. Furthermore, the structure of Reddit, with its community-centric 'subreddits', enables efficient filtering and collection of pertinent data. The platform's up vote and down vote system aids in preserving the quality of contributions. Consequently, when gathering data on posts related to mental health, Reddit, owing to its inclusive and comprehensive nature, emerges as the logical choice.

Subreddits	**Members**
r/depression	998,608
r/anxiety	647,120
r/suicide watch	452,111
r/mentalhealth	428,742
r/lonely	361,023

Table 4.1: Mental Health Subreddits with the Largest Memberships

4.1.2 Reddit as Data Source: The Wellspring of Narratives

There is a growing body of evidence pointing to the increasing use of Reddit as a data source across a variety of research disciplines, signifying an upward trend in academic publications utilising this platform [159]. In the context of our proposed study, we aim to create our dataset using Reddit, employing NLP techniques to analyse the data. This approach, aimed at detecting and assessing the impact of mental health on users, has been validated as effective in prior studies [160, 161].

Our rationale for choosing Reddit as our data collection medium is anchored on the following considerations:

4.1.2.1 Access to High-Volume Mental Health Data

- Reddit, as a widely utilised social media site with approximately 430 million users, hosts numerous dedicated, topic-specific forums known as 'subreddits.' These subreddits provide focused information on a plethora of subjects, including a significant number related to mental health. Prominent examples include r/mentalhealth, r/-suicidewatch, r/lonely, r/depression, and r/anxiety, all of which offer deep insights into users' mental health issues. These forums have gained recognition due to their size, consistency, and active user participation, and have been deemed beneficial and prevalent by domain experts [60, 70].

- Reddit serves as a platform where advice on a variety of mental health issues is sought and given. One of the most popular subreddits, r/IAMA, boasting 22.4 million members, addresses questions on anxiety, trauma, and general mental health under its affiliate r/Iama Health (Mental Health AMA), run by professional psychotherapists [162].
- Mental health-related subreddits host a considerable number of active members, as illustrated in Table 4.1. These users candidly post about their mental health issues, and these subreddits have already been used for research purposes [163, 164]. As demonstrated in Table 4.2, a significant number of mental health-related posts were made within just three sample months.
- Reddit allows for extensive textual submissions, with a generous character limit of 40,000 characters per post. This enables users to express their sentiments in detail, seek advice, or provide support. The potential for such detailed posts can provide a richer and more nuanced dataset for mental health studies.

4.1.2.2 Impacts of Anonymity on Data Quality and Relevance

A distinctive advantage of Reddit is the pseudonymity to its users. This feature encourages individuals to express themselves freely without fear of social stigma. Subreddit moderators play a crucial role in maintaining the anonymity of users by enforcing site rules that prohibit disclosure of personal identities. Consequently, this results in high-quality, subject-specific content that is likely less biased than data collected through questionnaires and surveys [165]. The authenticity and relevance of the subreddits are also maintained through the moderators' actions, who ensure the deletion of off-topic posts or 'submissions.' Reddit, unlike other social media platforms, allows for lengthy user posts or 'submissions.' This aspect permits users to express their sentiments in detail and either seek or provide support, making Reddit an ideal source for researchers seeking profound insights into sensitive aspects of the human psyche.

4.1.2.3 Reddit's Role in Mental Health Research

Reddit has emerged as a key resource for mental health research, thanks to its promotion of candid and in-depth user conversations. This platform has proven invaluable for exploring various mental health issues, including depression and anxiety, by offering a conducive space for such discussions. Additionally, its ability to provide timely mental health data was particularly beneficial during the COVID-19 pandemic [166]. The platform's credibility as a repository of genuine mental health stories, offering both detail and privacy, further underscores our rationale for choosing Reddit for in-depth mental health research. By leveraging these attributes, our investigation seeks to decode the patterns and emotions prevalent in mental health conversations, with the aim of gaining deeper insights.

4.1.3 Time Frame Selection

The selection of time frame, from January 2019 to August 2022, was carefully selected to cover periods before, during, and transitioning out of the pandemic. This decision was based on evidence indicating significant psychological impacts and increased reliance on social media during the pandemic period [167]. Such studies shed light on the evolving dynamics and emotional undertones within social media conversations throughout the pandemic, including notable increases in anxiety and depression [168, 169]. Our intention is to offer a thorough examination of the mental health dialogue on Reddit across these varied stages, aiming to encapsulate the shifting public mood and conversation trends during these pivotal times.

- Pre-Pandemic Phase (January 2019–December 2019): This segment acts as a reference point, providing insights into the state of mental health discussions prior to the worldwide awareness and effects of COVID-19.

- Mid Pandemic Phase (January 2020–December 2021): This interval is pivotal for observing the peak effects of the pandemic on mental well-being. It allows for an in-depth look at immediate responses, adaptation strategies, and the transformation in mental health conversations as influenced by the unfolding of the pandemic, including lockdowns and social isolation.

- Post-Pandemic Phase (January 2022–August 2022): Selected to investigate the gradual shift towards a post-pandemic reality, this period aims to reveal how mental health discussions are adjusting to the emergence of a 'new normal'.

Understanding trends during different phases informs targeted strategies for mental health interventions and policy decisions, enhancing the effectiveness and relevance of support measures. Ultimately, our dataset's holistic approach will contribute significantly to mental health research and empower stakeholders with the necessary information for effective and timely interventions.

	Jan-20	**Jan-21**	**Jan-22**
r/depression	17516	18706	14953
r/anxiety	5750	7931	7605
r/suicidewatch	8161	15135	14309
r/mentalhealth	4700	9098	9644
r/lonely	3061	4382	5215

Table 4.2: Number of Sample Posts from January 2020, 2021, 2022

4.1.4 The keys: Application Programming Interface (API) for Data Extraction

In line with our decision to adopt Reddit as our platform of choice for study, our attention turned towards leveraging the most efficient APIs available for data extraction [170], with a particular emphasis on Pushshift and PRAW. Pushshift stands out for its superior functionality compared to Reddit's native API, PRAW, by providing access to a broader scope of historical data and more generous query limits. The creation of Pushshift by Jason Baumgartner has been instrumental in simplifying the process of collecting and archiving Reddit data, thereby enhancing its utility for extensive research endeavours [171].

Pushshift's role as a pivotal resource for scholarly research is confirmed by its ability to amass comprehensive and impartial datasets, thus facilitating a more exhaustive data collection effort than what is achievable with other tools [172]. Its effectiveness is further underscored in the realm of mental health research, especially in projects focused on identifying instances of suicidal ideation within Reddit discussions, where Pushshift has proven its precision and efficiency in handling sophisticated data needs [173]. These attributes make Pushshift a critical element in our research approach, enabling us to efficiently harness the vast amount of data available on Reddit while ensuring the maintenance of user confidentiality and the integrity of the data collected.

4.1.5 Data Collection: Sourcing the Seeds

We used the Pushshift API (https://github.com/pushshift/api) to collect posts and corresponding metadata from chosen mental health subreddits. In order to select subreddits that focus on mental health issues, we utilized Reddit's search feature and identified the top five mental health disorder subreddits with the largest memberships, as detailed in Table 4.1.

Our data extraction was targeted towards Reddit posts from these subreddits across three distinct periods:

- Pre-Pandemic
- Mid-Pandemic
- Post-Pandemic

For an initial assessment of content relevancy and volume, we extracted sample posts from five subreddits for the months of January in 2020, 2021, and 2022. Our preliminary analysis showed that 99 percent of the sampled subreddit content directly addressed mental health concerns. A more detailed breakdown of these posts can be found in Table 4.2 and Figure 4.1.

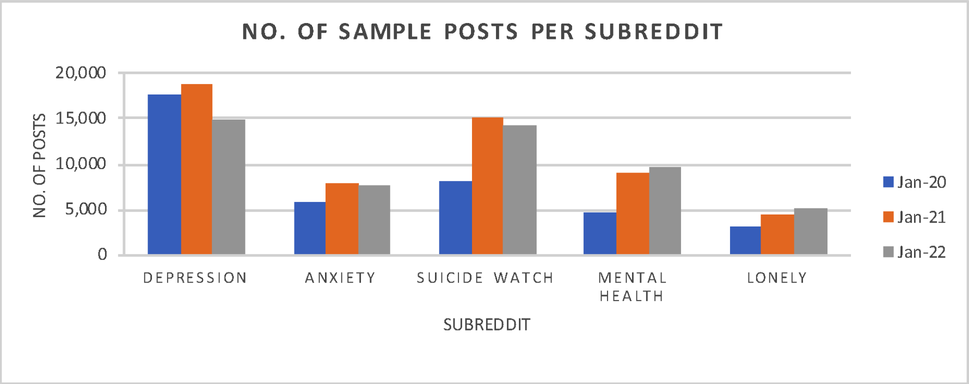


Figure 4.1: Number of Sample Posts

Table 4.3 presents an average post length of 103 words related to mental health illness. This metric denotes the quantity of processed tokens subsequent to the pre-processing phase.

Title	Post Text
lonely as hell	*"I feel like as I am a pathetic loser. I don't feel like if any of my friend care about me. I'm always the one sitting alone, and no one invite me to sit with them or take part in anything, and I know they all prefer each other over me. I don't know what wrong am doing. I want to live my life without worries. There must be something wrong with me I feel like a freak. It's making my depression worse. I am always on the verge of tears. I don't know what to do with myself I have no one."*

TABLE 4.3: Average Word Length of Sample Reddit Post

Data was downloaded from a total of five subreddits, r/depression, r/suicidewatch, r/-mentalhealth, r/anxiety, r/lonely as indicated in the Table1. The scope of the study was limited to original posts, thereby excluding comments and images. The posts were then cleaned to be exclusively in English. Data were obtained from 01-01-2019 to 31-08-2022 with a total of 1,494,019 posts as shown in Figure 4.2.

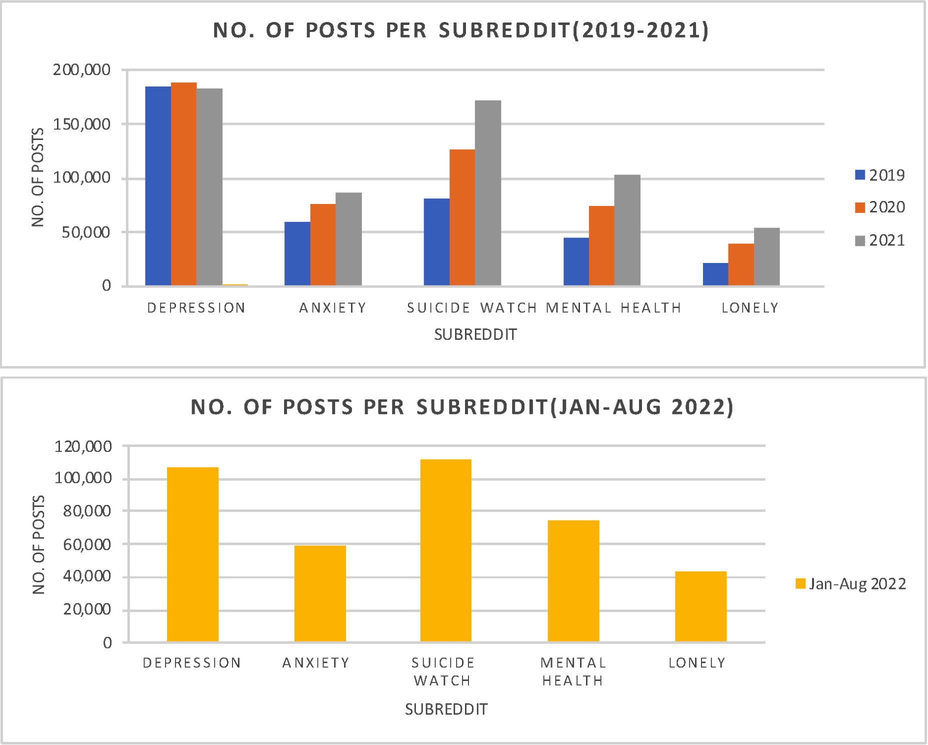


FIGURE 4.2: Total Number of Posts per Subreddit

4.1.6 Dataset Organisation

Our dataset comprises of an orderly collection of Reddit posts harvested from five principal mental health-focused subreddits: r/anxiety, r/depression, r/mentalhealth, r/suicidewatch, and r/lonely. These specific subreddits were chosen due to their concentrated discussions on mental health issues, offering a rich source of data for pertinent research studies. Our complete RHMD is divided into two major sections as illustrated in Figure 4.3. The first section encompasses the raw Reddit posts in their totality. However, the second section MH-RAC serves as the cornerstone of our study, containing a carefully annotated subset of RMHD posts.

4.1.6.1 RMHD:The Core Structure

The RMHD is chronologically arranged, divided into yearly folders from January 2019 through August 2022, with each year further broken down by month. Within these monthly divisions, subreddit-specific CSV files are stored. This systematic structure not only facilitates efficient temporal analysis but also enables researchers to access data specific to particular time frames and subreddits with ease. Each CSV file in RMHD includes the following columns, providing a detailed view of the Reddit posts along with essential metadata:

- **Author:** The Reddit username of the post author.
- **Created utc:** The post's creation time in UTC.
- **Score:** The post's net score, calculated as upvotes minus downvotes.
- **Selftext:** The body text of the post.
- **Subreddit**: The source subreddit of the post.
- **Title:** The post's title.
- **Timestamp:** The local creation date and time of the post, converted from UTC.

This structured organisation supports detailed temporal analyses and facilitates straightforward retrieval of data from specified subreddits.

4.1.6.2 Mental Health-Reddit Annotated Corpus(MH-RAC): The Annotated Layer

Central to our study is the MH-RAC, a subset sourced from RHMD. MH-RAC (n=800) encapsulates a diverse range of mental health narratives,representation of user-generated content. Each post meticulously annotated following robust annotation process to facilitate Model training. This process was carefully designed to annotate the broad range of mental health conditions discussed online. The annotation protocol, detailed in our section 4.2, involved rigorous rounds of review to ensure the utmost precision in the classification of each post. The integrity of this process was reinforced by stringent Inter-Annotator Agreement (IAA) metrics and a consensus method among a team of annotators, including both domain experts and trained annotators. This collaborative effort guaranteed the annotations' high reliability and validity.

The MH-RAC is structured with an equitable distribution of posts across four principal mental health root cause categories, Drug and Alcohol, Early Life, Personality, and Trauma and Stress with 200 posts dedicated to each. This equal representation was a deliberate choice, aimed at providing a balanced representation for the subsequent ML tasks. These categories pinpoint crucial factors affecting mental health, enabling a deeper exploration into the origins of mental health challenges and offering enriched understanding.

The columns in MH-RAC include:

- **Score**
- **Selftext**
- **Subreddit**
- **Title**
- **Label:** The assigned root cause category determined through our annotation.

In adherence to ethical and privacy guidelines, any references to individuals aged 1–18 within the posts have been anonymised, removing specific age mentions while keeping the posts intact. This measure ensures our dataset meets the highest ethical standards and respects the privacy of individuals' shared experiences on social media.

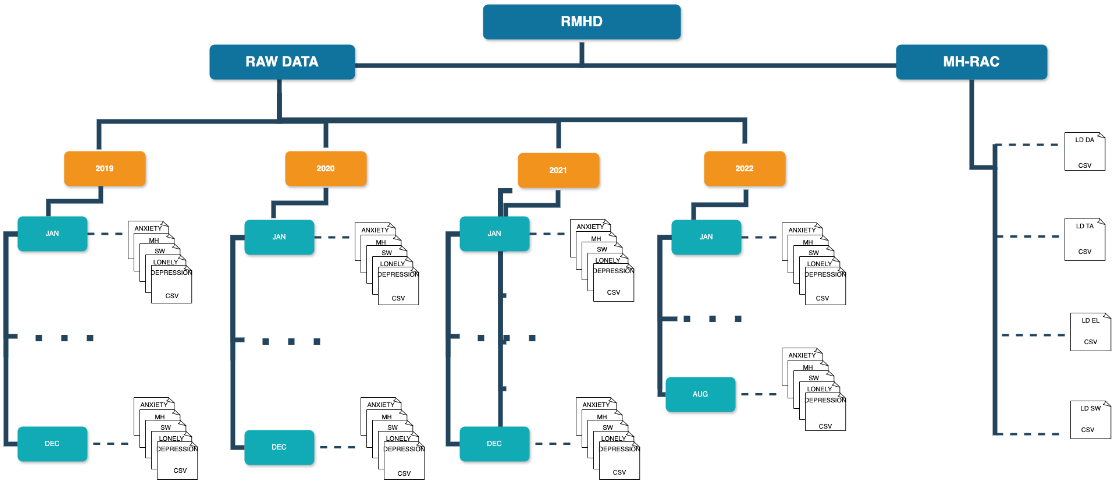


FIGURE 4.3: Data Organisation

4.1.6.3 Mental Health Evaluation Corpus (MHEC):The Testbed

To evaluate the trained models in comparison with human raters, we introduced the MHEC, which was derived from 50 new, unseen posts. These posts were specifically selected as a representative seed subset from the larger RMHD, herein referred to as RMHD-Seed.

MHEC plays a pivotal role in the second phase of the study, where both finetuned AutoML and PaLM 2 models, alongside human raters, engage with this distilled essence of RMHD-Seed to predict the root cause of mental health issues as depicted in user posts. The RMHD-Seed was meticulously annotated by three human raters to establish a consensus-based gold standard. This process transformed RMHD-Seed into MHEC, thereby enabling a comprehensive comparative analysis between the predictions made by the models and the judgements of human raters. Such a comparison offers valuable insights into the models' ability to generalise and emulate human-level discernment in classifying mental health narratives.

4.1.7 Dataset Availability Statement

The RHMD along MH-RAC, with a zipped size of approximately 1.68GB, is publicly available and serves as a rich resource for researchers interested in exploring the root causes of mental health issues as represented in social media discussions, particularly within the diverse conversations found on Reddit.

Dataset can be accessed on Kaggle at https://rb.gy/ewtjy.

4.2 Data Annotation: Crafting Clarity in the Narrative Weave

Embarking on the intricate task of employing ML for the nuanced classification of text, data annotation stands as the cornerstone of our methodology. This essential step transforms the raw, unstructured text data from social media into a meticulously organised dataset, ready for in-depth analysis. Our project draws from the forefront of mental health discourse analysis, incorporating proven strategies to meticulously label complex mental health concepts within social media content. Emulating and extending the categorisation framework detailed in seminal works [175], our study sets a new benchmark for identifying and interpreting the root causes behind mental health discussions online.

4.2.1 Compilation and Stratification of Initial Data Subset

The first phase in our data structuring process involved the careful curation of an initial data subset, pulling together a month's worth of discussions from five strategically selected subreddits: r/anxiety, r/lonely, r/suicidewatch, r/depression, and r/mentalhealth. This selection process was deliberately designed to capture a broad spectrum of mental health dialogues, aiming for a dataset that is as inclusive and representative as possible of the myriad mental health experiences shared online.

4.2.1.1 Implementing Stratified Sampling for Dataset Refinement

To navigate the intensive demands of manual labelling, we adopted a stratified sampling strategy to refine our extensive dataset into a focused subset of 800 posts. This method, inspired by successful approaches in analysing online communities [176], segments the larger pool of data into distinct groups aligned with each subreddit. Utilising keyword analysis for random selection within these groups ensures a balanced and equitable representation across our annotated sample, laying a solid foundation for targeted and meaningful analysis.

4.2.1.2 Ensuring Equitable Label Distribution

In the final step of our data annotation process, we emphasised the importance of distributing the annotated posts evenly across predetermined labels, each representing a key root cause of mental health issues. Our objective was to allocate 200 posts to each of the four labels, aiming for a dataset that is not only balanced but also rich in diversity. This balanced approach is crucial for reducing biases, covering a wide range of mental health topics fairly, and maintaining the analytical integrity of our study. Through this meticulous distribution, we enable ML models to engage more accurately and ethically with the complex narratives of mental health on social media, enhancing our collective understanding and approaches to mental health care.

4.2.2 Selection and Training of Annotators

Integral to the integrity of our data annotation process was the meticulous selection and comprehensive training of our annotators, a strategy that mirrors the best practices within mental health research [150, 151]. Our approach to assembling the annotation team was guided by the imperative for a balanced mix of domain knowledge and technical acumen, essential for the nuanced task at hand.

4.2.2.1 Selection Process and Inclusion Criteria

- **Internal Selection:** Due to resource limitations and the specialised nature of our research, we selected annotators from within our existing research team. This internal selection process guarantees a deeper initial understanding of the project's objectives and methodologies.
- **Educational Background:** Candidates were required to possess a minimum of a Master's degree in fields relevant to the project, such as Computer Science, Psychology, or Data Science, ensuring they had the necessary academic foundation.
- **Professional Experience:** We prioritised team members who had demonstrated capabilities in data analysis and had previous exposure to mental health topics through academic or project work. This experience was essential for understanding the complex nature of the annotation tasks.

- **Interdisciplinary Understanding:** Preference was given to those who demonstrated a blend of technical skills and an understanding of psychological principles, as this combination is vital for accurately interpreting mental health narratives.

The initial member of our annotation team, with a Master's degree in Computer Science and a background as a business analyst and resolution specialist, brought an invaluable analytical and problem-solving skill set to the project. Complementing this, the second member of the team, endowed with a Ph.D. in ML, contributed a wealth of experience from previous research endeavours, underscoring the critical role of meticulously annotated datasets in the realm of ML applications.

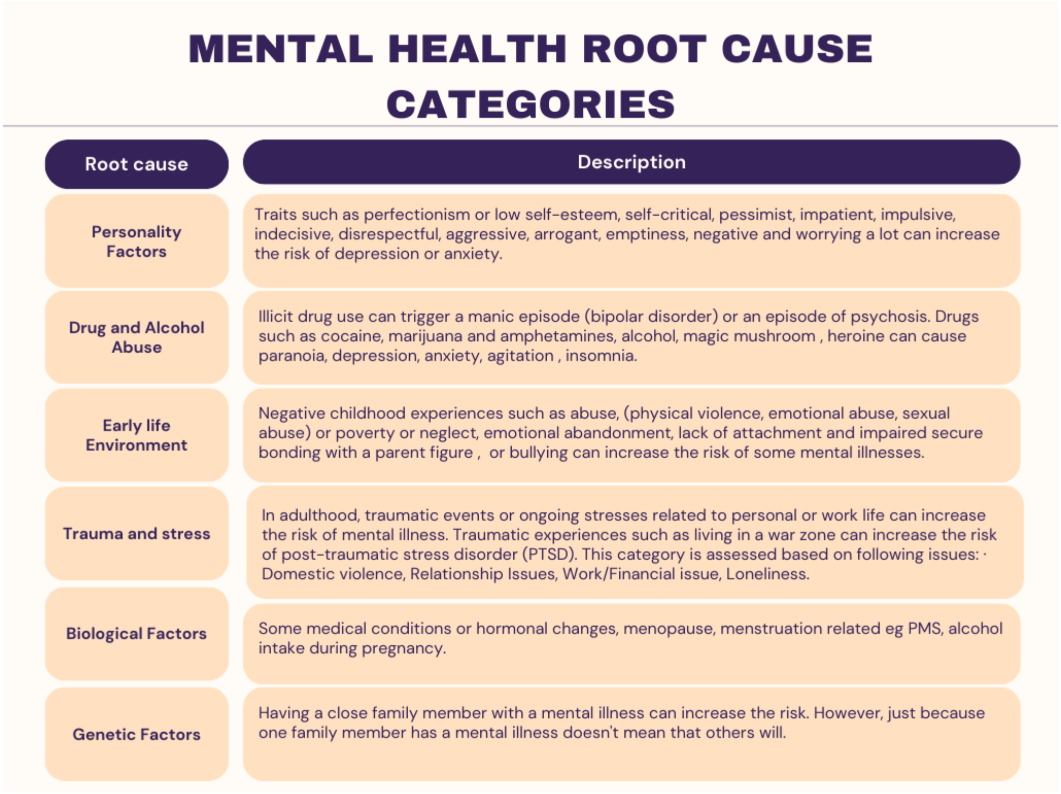
To ensure the highest standards of domain specificity and contextual understanding, both annotators underwent an intensive training program led by an expert in psychiatry. This model of collaboration, wherein domain experts such as clinicians or psychologists impart essential insights to the research team, is a well-established practice in the field [177, 178]. The domain expert's profound knowledge, particularly in mental health across various cultural landscapes, played a pivotal role in refining our research methodology. Such expertise enabled our annotators to grasp the complexities of mental health discourse, echoing the successful synergies between domain experts and computer science researchers documented in the literature [179].

The training regimen was designed to be iterative, fostering a progressive deepening of the annotators' understanding and application of the annotation guidelines. Central to the training were exercises in keyword identification and context interpretation within the posts, ensuring that the assignment of root cause categories was deeply informed by the essence of the discussions. Illustrative examples of posts, their corresponding root cause labels, and the rationale behind each annotation decision were provided (see Table 4.5). Through a series of practice annotations and subsequent feedback sessions led by the domain expert, the annotators honed their skills, aligning closely with the study's goals. The aim of this thorough and iterative training was to mitigate subjective biases and bolster the consistency of annotations, establishing a robust and reliable foundation for subsequent ML analysis. This deliberate and rigorous approach to annotator selection and training underscores our commitment to producing a dataset of unparalleled quality and relevance for advancing mental health research.

4.2.3 Annotation Guidelines

Given the intricate and multifaceted nature of mental health issues, the development of our annotation guidelines necessitated a meticulous multi-class label approach. Our objective was to establish a robust framework capable of encapsulating the nuanced expressions of these issues as they manifest in social media posts. The annotation guidelines were predicated on primary categories of root causes identified by esteemed national mental health resources, such as Health Direct Australia and Beyond Blue [174].

Drawing on these sources, we identified six main root causes, outlined in Figure 4.4, which aid in pinpointing the origins of mental health concerns. For our research, we chose to focus on four specific root causes: Personality Factors, Drug and Alcohol Abuse, Early Life Environment, and Trauma and Stress. These categories were selected based on their wide coverage of significant mental health concerns, their prevalence in both scholarly research and societal conversations, and their ability to collectively encompass a range of potential factors contributing to mental health issues. [180–186].



MENTAL HEALTH ROOT CAUSE CATEGORIES

Root cause	Description
Personality Factors	Traits such as perfectionism or low self-esteem, self-critical, pessimist, impatient, impulsive, indecisive, disrespectful, aggressive, arrogant, emptiness, negative and worrying a lot can increase the risk of depression or anxiety.
Drug and Alcohol Abuse	Illicit drug use can trigger a manic episode (bipolar disorder) or an episode of psychosis. Drugs such as cocaine, marijuana and amphetamines, alcohol, magic mushroom , heroine can cause paranoia, depression, anxiety, agitation , insomnia.
Early life Environment	Negative childhood experiences such as abuse, (physical violence, emotional abuse, sexual abuse) or poverty or neglect, emotional abandonment, lack of attachment and impaired secure bonding with a parent figure , or bullying can increase the risk of some mental illnesses.
Trauma and stress	In adulthood, traumatic events or ongoing stresses related to personal or work life can increase the risk of mental illness. Traumatic experiences such as living in a war zone can increase the risk of post-traumatic stress disorder (PTSD). This category is assessed based on following issues: · Domestic violence, Relationship Issues, Work/Financial issue, Loneliness.
Biological Factors	Some medical conditions or hormonal changes, menopause, menstruation related eg PMS, alcohol intake during pregnancy.
Genetic Factors	Having a close family member with a mental illness can increase the risk. However, just because one family member has a mental illness doesn't mean that others will.

FIGURE 4.4: Mental health Root Causes By Health Direct Australia [8]

To systematically categorise the complex factors contributing to mental health issues, we have classified these key root causes into distinct classes. This classification not only facilitates a clearer analysis and interpretation of the data but also adheres to model training constraints by restricting them to two tokens each. Table 4.4 presents four selected root causes and their corresponding classes.

The annotators received guidance on identifying key terms within the posts and, importantly, on interpreting the context surrounding these terms. This method ensured that the assignment of root cause categories was precise, capturing the true nature of the discussions by the users. For a select group of these posts, Table 4.5 showcases a comprehensive explanation that connects specific posts to their designated root cause labels, demonstrating how our annotation framework was applied in actual scenarios.

To ensure uniformity and accuracy in our process, our instructions also outlined steps for handling unclear cases, either by reaching a consensus among annotators or consulting a domain expert when consensus was not achievable. Quality assurance was maintained through practices such as checks for agreement between annotators and regular reviews of the guidelines by a domain expert.

Root Cause	**Class Name**
Personality Factors	personality
Drug and Alcohol Abuse	drugalcohol
Early Life Environment	earlylife
Trauma and Stress	trauma

Table 4.4: Root Causes and Corresponding Classes

The formulation of our annotation guidelines marked a pivotal phase in our research, requiring deep knowledge of the topic, a thorough examination of existing literature, and careful consideration of the annotation process's logistical elements. The result was a comprehensive set of guidelines that offered a definitive and solid structure for annotating our dataset.

4.2.4 Annotation Process Strategy

The annotation process in our study was designed to accurately identify the underlying factors contributing to mental health issues as reflected in social media discussions, utilising a comprehensive methodology to ensure the consistency and accuracy of our data annotations. This involved a three-part strategy: semi-automated analysis of keywords, thorough review of entire posts, and meticulous manual annotation, as illustrated in Figure 4.5.

4.2.4.1 Keyword-Based Analysis

Initially, we employed a semi-automated process for identifying keywords. This method involved using a search function to locate predefined keywords within the dataset, which were chosen based on their relevance to each root cause category and listed in Table 4.7. Keywords such as 'perfectionism', 'low self-esteem', and 'pessimistic' were selected for the category of "personality factors", with the selection process guided by the expertise of a consultant psychiatrist to ensure an empirical and grounded methodology. The search function executed precise string-matching to flag relevant mentions within the social media posts, facilitating early categorisation and paving the way for further in-depth analysis.

4.2.4.2 Whole Post Analysis

Following keyword identification, we moved to an exhaustive examination of the full content of each post. This step ensured that the root causes identified were in alignment with the overall context and narrative of the post. It was crucial for verifying that the identified root cause was the main focus of the post, in accordance with our preset annotation guidelines. This detailed review was vital not just for confirming the findings from the keyword-based phase but also for gaining a deeper insight into the complex experiences and expressions of the authors, providing a more comprehensive understanding of the content beyond mere keyword occurrence.

4.2.4.3 Manual Labelling

The final of our annotation process involved manually assigning a single category label to each post, based on the defined root causes. To preserve the clarity and precision of our dataset, posts that spanned multiple categories were excluded. This selective labelling approach resulted in a dataset where each post distinctly represented a specific root cause, thereby offering clear and focused insights into the particular factors influencing mental health as depicted in social media discussions.

By employing these strategies as a unified approach, we captured the nuanced expressions of root causes in the posts, enhancing the reliability and effectiveness of our dataset. This comprehensive strategy will enable subsequent ML models to learn and classify the root causes of mental health issues more accurately in new, unseen data.

4.2.4.4 Annotation Time

On average, annotators spent approximately 20 minutes per post. This duration varied depending on the complexity and length of the post. Posts that were straightforward typically took about 10 minutes, whereas posts with complex emotional content or ambiguous contexts could take up to 30 minutes to annotate accurately. This variability reflects the careful attention to detail required to ensure high-quality data annotation, particularly when dealing with nuanced subjects such as mental health.

4.2.5 Validation and Quality Assurance: The Seal of Trust

Following the annotation phase, our study advanced to the pivotal stage of validation and quality assurance, underscoring the integrity and utility of our annotated dataset. The bedrock of our approach was a consensus-driven method, bolstered by rigorous IAA metrics [187]. These components were paramount in affirming the accuracy, objectivity, and reliability of our annotations, playing a crucial role not only in validating initial annotations but also in the continual refinement of our annotation guidelines. This process significantly enhanced the consistency and reliability of our dataset. We detail below the integration of IAA metrics into our annotation process, reinforcing the methodological soundness of our study.

4.2.5.1 Consensus-Based Approach for Annotation

In our process, two annotators, following detailed guidelines, independently categorised each post into one of the pre-defined mental health root causes: personality factors, trauma, drug and alcohol abuse, and early life experiences. This independent classification set the stage for a vital step in our methodology, achieving a unanimous consensus on each post in cases of initial divergence. This dialogue between annotators was critical for thoroughly understanding the nuanced personal experiences and the specific contexts related to the individuals' mental health depicted in the posts.

During their discussions, annotators evaluated if a post's content explicitly or implicitly reflected any of the designated root causes, and whether the described scenarios could be accurately categorised under these headings, adhering strictly to our annotation guidelines. This balanced approach aimed to capture the complexity of individual experiences and the factual bases of their mental health conditions, offering a richer understanding of the underlying causes of mental health issues.

4.2.5.2 Implementing Inter-Annotator Agreement Measures

A cornerstone of our quality assurance was the implementation of IAA measures. After annotators independently reviewed the posts, we applied IAA calculations, utilising a Contingency Table 4.6 prepared from the initial annotation phase. This quantitative analysis, which included Cohen's Kappa statistic, went beyond simple percentage agreements to consider the likelihood of chance agreement. This was particularly useful for resolving disagreements between annotators [188].

The Cohen's Kappa calculation was conducted as follows:

- Total number of posts: 800 (722 agreements + 78 disagreements)
- Observed Agreement P_o: The proportion of times both annotators agree (including the 722 agreed posts).
- Expected Agreement by Chance P_e: Remains the same, calculated based on the proportion of each category chosen by each annotator.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{4.1}$$

Our analysis yielded a Cohen's Kappa score of approximately 0.869, indicating a very high level of agreement beyond mere chance. This high score attests to the reliability of our annotation process and the robustness of our methodological framework.

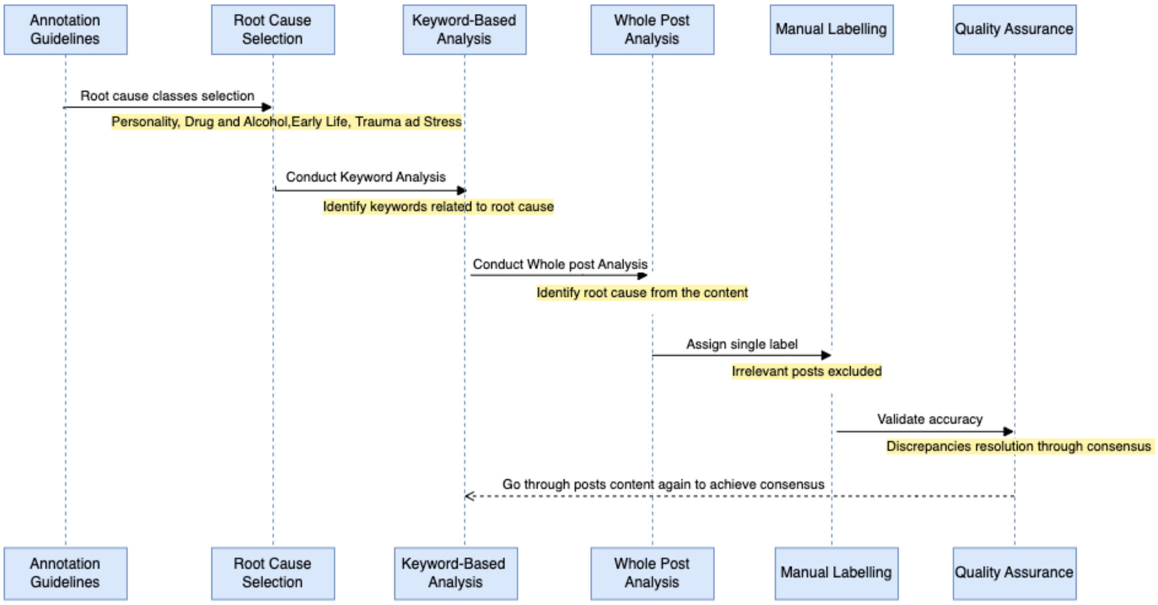

Annotation Guidelines
Root Cause Selection
Keyword-Based Analysis
Whole Post Analysis
Manual Labelling
Quality Assurance
Root cause classes selection
Personality, Drug and Alcohol,Early Life, Trauma ad Stress
Conduct Keyword Analysis
Identify keywords related to root cause
Conduct Whole post Analysis
Identify root cause from the content
Assign single label
Irrelevant posts excluded
Validate accuracy
Discrepancies resolution through consensus
Go through posts content again to achieve consensus
Annotation Guidelines
Root Cause Selection
Keyword-Based Analysis
Whole Post Analysis
Manual Labelling
Quality Assurance

Figure 4.5: Data Annotation Sequence

4.2.5.3 Domain Expert Consultation

To address the complexities inherent in mental health discussions and deadlock situations between annotators, we enlisted the expertise of a domain expert. This crucial step provided additional clarity and direction, enhancing the precision and depth of our annotations and, by extension, the overall quality of our dataset. The expert's involvement was structured at several critical stages of the annotation process:

1. **Initial Training:** Before the annotation process began, the domain expert conducted a series of training sessions to familiarise annotators with necessary mental health concepts and specific context of the data. These sessions included detailed discussions on various mental health conditions, annotation criteria, and the nuances of interpreting emotional content in social media posts.

2. **Ongoing Support**: During the annotation process, the domain expert was available for weekly consultation meetings where annotators could discuss challenging posts and seek clarification on complex cases. This ongoing support helped maintain the consistency and accuracy of the annotations.

3. **Post-Annotation Review:** After the completion of annotations, the domain expert reviewed a sample of the annotated posts to ensure quality and consistency. Feedback from this review was used to refine the annotations further if necessary.

The incorporation of IAA measures and expert consultation into our workflow was instrumental in ensuring the validity and reliability of our dataset. This meticulous approach guaranteed that our dataset accurately reflected the mental health discussions from the chosen subreddits, setting a high standard for data quality and supporting the generation of trustworthy research findings.

	Root Cause	**Post**	**Rationale**
P1	Personality Factors	*Essence of human life is love isn't it?unfortunately i am unable to feel it. I feel like a stranger. I don't belong anywhere. Everything is same i want diversity.I don't have any moral values or guilt. I strongly desire self actualisation[Some text omitted for brevity]*	Low Self Esteem
P2	Drug and Alcohol	*I use cannabis every day. I look forward to it because I feel so much better/content when I use. However some say it's actually making my depression and anxiety worse so it's like a vicious cycle[Some text omitted for brevity]*	Drug Consumption
P3	Early Life	*When iwas a kid i used to be really skinny and too short for my age. I got bullieda lot because of that.I'm 20 y.o now, and i think i have distorted view of my own appearance. I keep forgetting that i'm the size of an adult nowadays, and i feel like i'm skinnier and weaker than i actually am. I still feel like a little boy.I keep feeling like i'm still not big enough.[Some text omitted for brevity]*	Bullying Poor Body Image since Early Life
P4	Trauma and Stress	*No one LITERALLY .... NO ONE has talked or even invited me to hang out for New Years. I just looked at my mates story and it's my 3 “closest friends” hanging out and having fun. Why does everyone seem to hate me? I'm not “too nice” or a cunt to people, I don't make a fuss and I just chill. All my life I've seem to ignored for some reason and it's driving me fucking insane! F... everything right now.[Some text omitted for brevity]*	Lonelines Social Rejection
P5	Trauma and Stress	*I'm miserable. Things went really bad when I was 16 and my parents split up and my cousin killed himselfwithin a month. my parents splitting up was and is still extremely bitter, their hate for each other consumes them and they both want each other dead. When they say it out loud it's normal for me now but it feels so wrong.My dad says, he's gonna make her homeless. So I'm going to bepaying £600 a month when I will only be earning about £700 or less. I had a weekend job which I left because it was too stressful working.it is all so wrong[Some text omitted for brevity]*	Relationship Financial Disorganised Attachment, Bitterness Sadness
P6	Trauma and Stress	*I don't understand what's wrong with me. My family makes me feel wrong. According to them everything i do is selfish. I don't need them though. They don't understand me, My family hates me. I'm their least favourite. Always have been. I wish there was a way for them to share how my head feels so they could understand why i am how i am. I hate living. There's nothing i want to do more than die. Someone who wants to care. I need someone please [Some text omitted for brevity]*	Relationship Issues Anger Sadness

Table 4.5: Mental Health Root Cause Rationale

		Annotator 1				
Annotator 2	**Label**	**Personality**	**Trauma**	**Early Life**	**Drug alcohol**	**Total**
	Personality	157	23	16	4	200
	Trauma	15	185	0	0	200
	Early Life	0	15	185	0	200
	Drugalcohol	0	5	0	195	200
	Total	172	228	201	199	800

Table 4.6: Inter-Annotator Agreement (IAA) Contingency Table

Detailing concordance and discordance between two annotators across four key mental health root cause categories. This table serves to quantify inter-annotator agreement as part of the data validation process.

Class	**Potential Keywords**
Personality	Perfectionism, Low self-esteem, Self-critical, Negative, Worry, Pessimistic, Impatient, Impulsive, Indecisive, Disrespectful, Aggressive, Arrogant, Emptiness
Drug and Alcohol	Alcohol, Drugs, Substance abuse, Addiction, Medication, Dependence, Overdose, Intoxication, Withdrawal, Rehab, Detox, Relapse, Sobriety
Early Life	Childhood, Upbringing, Parenting, Neglect, Abuse, Trauma, Environment, Family, Poverty, Parent's Divorce, Bullying, School, Early age mention
Trauma and Stress	Trauma, Stress, PTSD, Anxiety, Depression, Violence, Abuse, Accident, Environment, Disaster, Loss, Grief, Financial Crisis, Work Stress, Breakup

Table 4.7: Potential Keywords of Mental Health Root Causes

Chapter 5

Deciphering Minds: AI and Human Analysis of Mental Health Narratives

In this chapter, we delve into the technical intricacies of our model training strategy, from the preparatory stages of data cleaning and normalisation to the elaborate fine-tuning processes tailored to enhance model performance. Our goal was to meticulously adjust these models to capture the unique linguistic features of mental health discourse on social media, thus ensuring the accuracy and reliability of our analysis. Through this methodical approach, facilitated by the cutting-edge capabilities of Vertex AI, we are equipped to unearth meaningful insights from social media narratives, contributing significantly to the fields of mental health research and artificial intelligence.

5.1 Data Pre-processing: Breaking Down Barriers

Data pre-processing is a critical step in the analysis of textual data, particularly in the context of NLP and ML applications. This process involves preparing the raw data for analysis by removing or modifying data that could interfere with the results. In the study of mental health issues using social media data, data pre-processing ensures the clarity, relevance, and quality of the text data, enabling more accurate and insightful analyses. The pre-processing steps applied in this thesis include text cleaning, tokenization, and stop word removal. Python, a versatile programming language with robust libraries for data manipulation and NLP, was used to execute these processes.

5.1.1 Text cleaning

Text cleaning is the first step in the data pre-processing phase, aimed at removing unnecessary and irrelevant elements from the data. This step is crucial for reducing noise in the dataset and improving the performance of NLP models. In this study, the text cleaning process involved several key operations:

- **Removing URLs:** Social media posts often contain URLs that are irrelevant to the analysis of mental health issues. Python's re library was utilised to identify and remove these URLs from the text.
- **Eliminating Emojis:** While emojis can convey emotions, their wide variety and the complexity of interpreting them consistently across different contexts make them challenging to analyse. As such, emojis were removed from the dataset.
- **Age Filtering:** A nuanced approach was taken to filter out explicit mentions of ages ranging from 1 to 18 using Python code. This was accomplished without altering the context of the text, thereby maintaining its original intent and meaning. The filtering specifically targeted references to age that could imply the data pertained to minors, aligning the study with ethical considerations and focusing the analysis on the adult population.

The Python code for these text cleaning operations leveraged libraries such as **re** for regular expressions and **pandas** for data manipulation, streamlining the cleaning process.

5.1.2 Tokenization

Following the initial cleaning, the text data was tokenized. Tokenization is the process of splitting the text into individual words or tokens, which serves as the basis for further analysis and processing. Python's Natural Language Toolkit (nltk) library was used for tokenization, providing a straightforward method to divide the cleaned text into manageable pieces. This step is fundamental in preparing the data for tasks such as frequency analysis, sentiment analysis, and the training of ML models.

5.1.3 Stopword removal

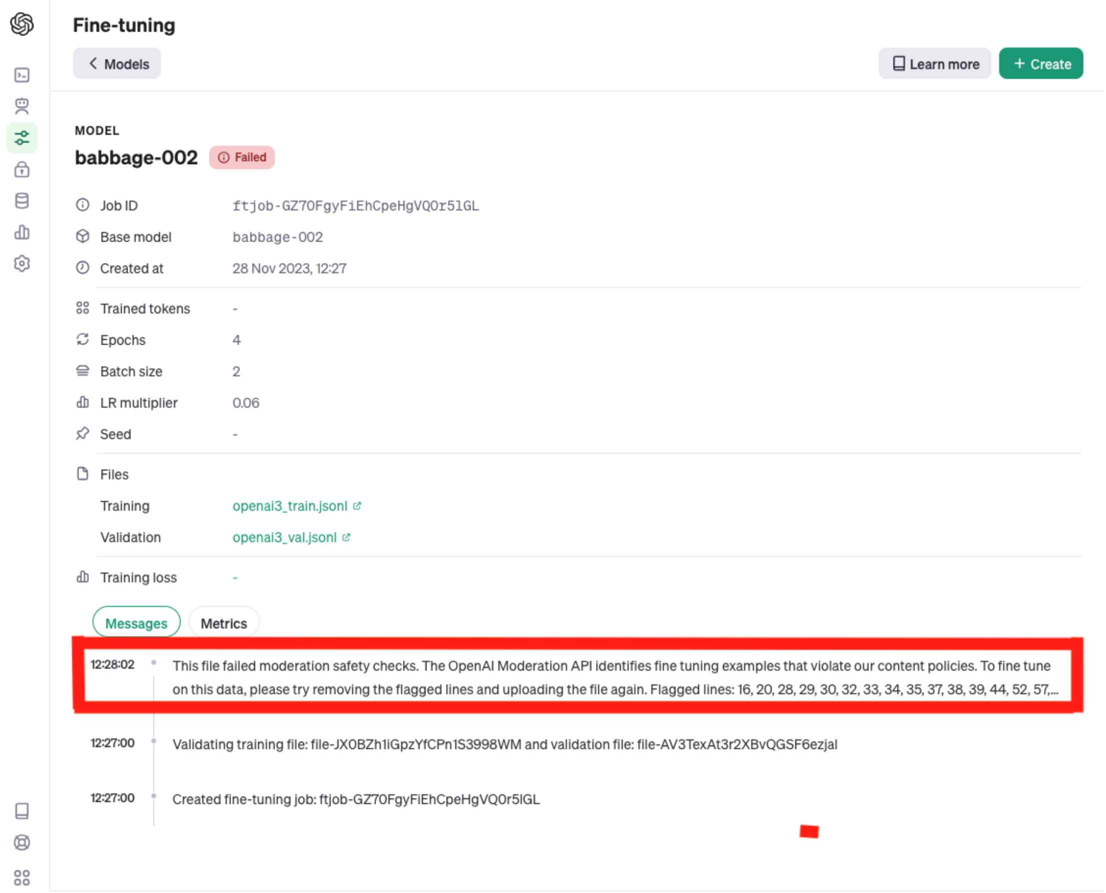
The final step in the data pre-processing phase was the removal of stop words. Stop words are common words that carry little to no semantic value in the context of the analysis, such as "the", "is", and "in". Their removal is essential for focusing the analysis on the most meaningful words in the text. The nltk library's comprehensive list of English stop words was employed to filter out these terms from the tokenized text. Additionally, custom stop words specific to the context of social media and mental health discussions were identified and removed to further refine the dataset.

By executing these data pre-processing steps, the dataset was effectively prepared for subsequent analysis, ensuring that the text was clean, relevant, and structured in a way that maximises the potential for insightful findings in the study of mental health issues on social media. This meticulous preparation forms the foundation for accurate and meaningful NLP and ML applications, enhancing the study's contributions to understanding mental health in the digital age.

5.2 Model Selection and Training: Crafting Intelligence

The selection of the most suitable models for our research was a journey marked by careful consideration and strategic adaptation, aimed at navigating the complexities of analysing mental health narratives within the vast, unstructured datasets of social media. Initially, our ambition was to harness the combined strengths of OpenAI's GPT-3.4 Turbo and Google AI's PaLM 2, both of which are at the forefront of NLP technologies. Our rationale was grounded in their proven excellence in handling complex NLP tasks, making them prime candidates for the nuanced analysis required to discern mental health discussions from social media content. This approach was driven by the anticipation that such sophisticated tools could offer unparalleled insights into the subtleties of mental health narratives, given their advanced capabilities in text analysis and interpretation.

However, our plan encountered a significant challenge when we faced OpenAI's content policy restrictions as shown in Figure 5.1, which posed a roadblock to fine-tuning GPT-3.4 Turbo with our dataset, enriched with sensitive mental health content. This unforeseen constraint necessitated a pivotal shift in our methodology, leading us to refine



babbage-002

Job ID	ftjob-GZ70FgyFiEhCpeHgVQ0r51GL
Base model	babbage-002
Created at	28 Nov 2023, 12:27
Trained tokens	-
Epochs	4
Batch size	2
LR multiplier	0.06
Seed	-
Files	
Training	openai3_train.jsonl
Validation	openai3_val.jsonl
Training loss	-

```
12:28:02  This file failed moderation safety checks. The OpenAI Moderation API identifies fine tuning examples that violate our content policies. To fine tune
          on this data, please try removing the flagged lines and uploading the file again. Flagged lines: 16, 20, 28, 29, 30, 32, 33, 34, 35, 37, 38, 39, 44, 52, 57,...

12:27:00  Validating training file: file-JX0BZh1iGpzYfCPn1S3998WM and validation file: file-AV3TexAt3r2XBvQGSF6ezjal

12:27:00  Created fine-tuning job: ftjob-GZ70FgyFiEhCpeHgVQ0r5lGL
```

FIGURE 5.1: OPEN AI Content Restriction

our focus exclusively on Google AI's PaLM 2. Unlike GPT-3.4 Turbo, PaLM 2 offered a more adaptable framework capable of processing our dataset without compromising the integrity and depth of our analysis. The decision to proceed with PaLM 2 was not merely a matter of convenience but a strategic choice to employ a tool that matched our research requirements with its exceptional NLP performance and flexibility in handling sensitive data.

To complement PaLM 2's foundational capabilities, we also integrated an Automated Machine Learning (AutoML) system into our research design. The inclusion of AutoML was strategic, aiming to exploit its adaptive capabilities and automated optimisation processes in analysing the intricacies of mental health narratives. This dual-model approach, leveraging both PaLM 2 and AutoML, was conceived to capitalise on their distinct but complementary strengths in processing complex linguistic data.

Our deployment of these advanced NLP models was facilitated by Vertex AI, Google

Cloud's premier platform, which offered a comprehensive ecosystem for training, deployment, and customisation of ML models and AI-powered applications, including LLMs. Vertex AI's robust infrastructure and integrated workflows provided an ideal environment for our project, enabling seamless collaboration and efficient management of the extensive data processing demands inherent in our study. This platform not only allowed us to customise and fine-tune PaLM 2 and AutoML to our specific research needs but also ensured the scalability and computing power necessary to handle our dataset's complexity.

This facilitates the customisation of LLM like PaLM 2 for nuanced text analysis and leverages the adaptive capabilities of AutoML for our specific research needs. By hosting our project on Vertex AI, we not only streamline the training and fine-tuning of these sophisticated models with our curated dataset but also harness the platform's powerful computing resources to efficiently manage the extensive data processing requirements. Through the Vertex AI platform, we are poised to achieve a high degree of model accuracy and reliability, essential for the sensitive task of analysing mental health narratives and ensuring the applicability of our findings in real-world scenarios.

5.2.1 PaLM 2 Approach

Developed by Google AI, PaLM 2 is a part of Vertex AI's Generative AI suite, a pivotal model within the domain of foundation decoder-only architectures akin to those within the GPT lineage, is adeptly engineered to excel across a broad spectrum of NLP tasks. This includes, but is not limited to, complex reasoning challenges spanning code interpretation and mathematical problem-solving, nuanced classification and question answering, adept translation and multilingual capabilities, as well as sophisticated natural language generation. PaLM 2 utilises the Transformer architecture, renowned for its self-attention mechanisms that assess the importance of different words in a sentence, irrespective of their distance from each other. This architecture processes all parts of the input data simultaneously rather than sequentially, enhancing both the efficiency and effectiveness of understanding language context. Google has made significant advancements in the core architecture of PaLM 2, enhancing its ability to learn complex relationships within language and code. These enhancements include an increased number of layers, attention heads, and overall parameter count, which enable a deeper and more nuanced understanding and generation of language.

In terms of scale and training data, PaLM 2 is trained on a vast corpus comprising a wide array of internet texts, which underpins its robust performance across various tasks and languages. Additionally, PaLM 2 utilises Google's Pathways system, designed to optimise the use of computational resources during training. This system enhances the efficiency and flexibility of how tasks are managed and executed across different hardware units, significantly improving the model's training efficiency and performance capabilities.

Moreover, PaLM 2 utilises a highly efficient architecture that allows it to scale effectively on powerful computer systems, enabling it to process massive amounts of data. This compute-optimal scaling ensures that PaLM 2 can handle extensive datasets with ease, further bolstering its applicability across diverse and demanding AI tasks. Its architecture leverages billions of parameters that enable deep understanding and generation of human-like text, making it ideal for interpreting complex mental health narratives. The model's ability to integrate context significantly enhances its performance in text classification and sentiment analysis, pivotal in extracting meaningful insights from diverse mental health discussions.

we incorporated PaLM 2 for Text (text-bison@001) model, an advanced derivative of the PaLM 2 framework. The text-bison@001 model underwent a specialised instruction based fine-tuning process, meticulously tailored to interpret the nuanced lexicon endemic to mental health discourse from our MH-RAC. Such instruction-based fine-tuning of LLM significantly enhances their contextual understanding and responsiveness within specific domains of application. By refining their capacity to interpret domain-specific content, these models demonstrate marked improvements in aligning with the intended purpose, thereby mitigating inherent limitations observed in their foundational configurations. Consequently, instruction-fine-tuned LLMs emerge as the preferred choice for applications in sensitive fields such as healthcare and medicine, offering a tailored approach that underscores both efficacy and ethical considerations [189].

5.2.1.1 Dataset Formatting for PaLM 2

Our venture began with establishing a Google Cloud account, a crucial step to access the Vertex AI platform. With access secured, we navigated to the Language section within Vertex AI, setting our sights on the generative AI capabilities for language-related tasks. This exploration was foundational, allowing us to gauge the potential of generative AI in enhancing our research.

For fine-tuning, we opted for a label-supervised approach, leveraging the MH-RAC, which we annotated with mental health root causes to align with our research aims. We formatted MH-RAC in a JSON Lines (JSONL) file, adhering to a specific schema to facilitate the fine-tuning process of text-bison@001. The schema is structured as follows:

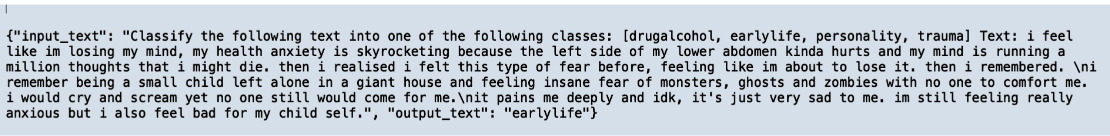
Each line in the JSONL file contains a pair of "input_text" and "output_text".

Input_Text : The "input_text" field is designed to include the prompt or the textual content that requires analysis or response.

Output_Text : Correspondingly, the "output_text" field contains the appropriate label or classification for the given input, aligning with our project's focus on mental health themes such as trauma, stress, personality factors, and substance use.

```
{
  "input_text": "Prompt or Textual content for analysis"
  "output_text": "Corresponsding label Or completion"
}
```

An example of original JSNOL file for training data is shown in Figure 5.2.



```
{"input_text": "Classify the following text into one of the following classes: [drugalcohol, earlylife, personality, trauma] Text: i feel
like im losing my mind, my health anxiety is skyrocketing because the left side of my lower abdomen kinda hurts and my mind is running a
million thoughts that i might die. then i realised i felt this type of fear before, feeling like im about to lose it. then i remembered. \ni
remember being a small child left alone in a giant house and feeling insane fear of monsters, ghosts and zombies with no one to comfort me.
i would cry and scream yet no one still would come for me.\nit pains me deeply and idk, it's just very sad to me. im still feeling really
anxious but i also feel bad for my child self.", "output_text": "earlylife"}
```

Figure 5.2: Text-bison Training DataSet Format

Each "input_text" was confined to a maximum of 8,192 tokens, while the "output_text" was capped at 1,024 tokens, adhering to the model's tokenization parameters to balance detailed representation with computational efficiency. The training set consisted of 560 data points, with the remaining 240 allocated for validation, providing the model with an extensive foundation for learning and subsequent performance assessment. This strategic division allowed for a robust training environment and a substantial validation set to verify the model's generalizability and precision.

5.2.1.2 PaLM 2 Fine Tuning Process

Vertex AI employs automated mechanisms to handle feature extraction and model tuning, which streamline the training of machine learning models for complex datasets. The feature extraction process within the platform involves analysing the text to automatically derive features that are most relevant for training the model based on the embedded algorithms in the AI systems.

With our dataset prepared, we embarked on the fine-tuning process, focusing on several key areas:

Working Directory Setup: A designated Google Cloud storage bucket was established as our working directory, organising and storing the model alongside the associated data. This strategic setup ensured an efficient workflow and easy access to necessary resources throughout the fine-tuning process.

Training Region Selection: The US central region was our choice for the training process, utilising the GPU 8*A100 80 GP hardware resources. This decision was pivotal in leveraging high computational power to ensure the efficiency and speed of our model's training, crucial for handling our extensive dataset and aligning with the project's objectives.

Hyperparameters Tuning: The fine-tuning of the supervised model was conducted using Vertex AI's PaLM 2 Text Bison model, specifically the variant text-bison@001.The fine-tuning process involved setting key hyperparameters, which were crucial in optimising the model's performance while preventing over fitting. This was done within the managed environment of Vertex AI, which automates certain aspects of the machine learning workflow, allowing for efficient and reproducible results.

The process included setting the following hyperparameters:

- **Learning Rate Multiplier:** Set at 1. This was chosen to maintain the balance between model convergence and stability during training.
- **Number of Training Steps:** 100 steps were determined to be optimal for the dataset size and complexity, ensuring sufficient learning without over fitting.
- **Model Variant:** The text-bison@001 variant was selected based on its architecture's suitability for the task at hand.

The Vertex AI platform facilitated the fine-tuning process by providing automated tools to streamline the workflow, ensuring that these hyperparameters were consistently applied across experiments. This method ensured that the model was fine-tuned effectively, with a careful balance between model complexity and generalisation capacity.

Model Pipeline Execution and Development: The Vertex AI pipeline's visualisation, an embodiment of the fine-tuning journey, began with the validate-pipeline step. This initial phase confirmed the precision of our pipeline setup, preempting potential configuration mishaps.

Progressing to the **tuning-graph**, the platform engaged in hyper parameter tuning, an intensive struggle aimed at pinpointing the model's optimal performance parameters. The **export-managed-dataset** followed, where Vertex AI seamlessly exported our dataset prepared as JSNOL file, rendering it accessible for the upcoming training phase.

The **dataset-encoder** stage was where the platform converted our dataset into a model-friendly format, a prerequisite for the subsequent model training. Ambiguously labelled as **vertex-pipelines-prompt-**, the next step likely involved specifying the training job's parameters, an action shrouded by the truncated label yet crucial for the process.

The orchestration of the model's parameters was adeptly handled by Vertex AI during the **compose-params-for-mo-** step. Here, various training options were defined, including the training steps and learning rate. The heart of the operation was the **large-language-model-tu-** phase, where the model's intellectual capacity was honed, absorbing knowledge from the data. Post-training, the **saved_model** served as the repository of my now-trained language model, earmarked for evaluation and deployment. Vertex

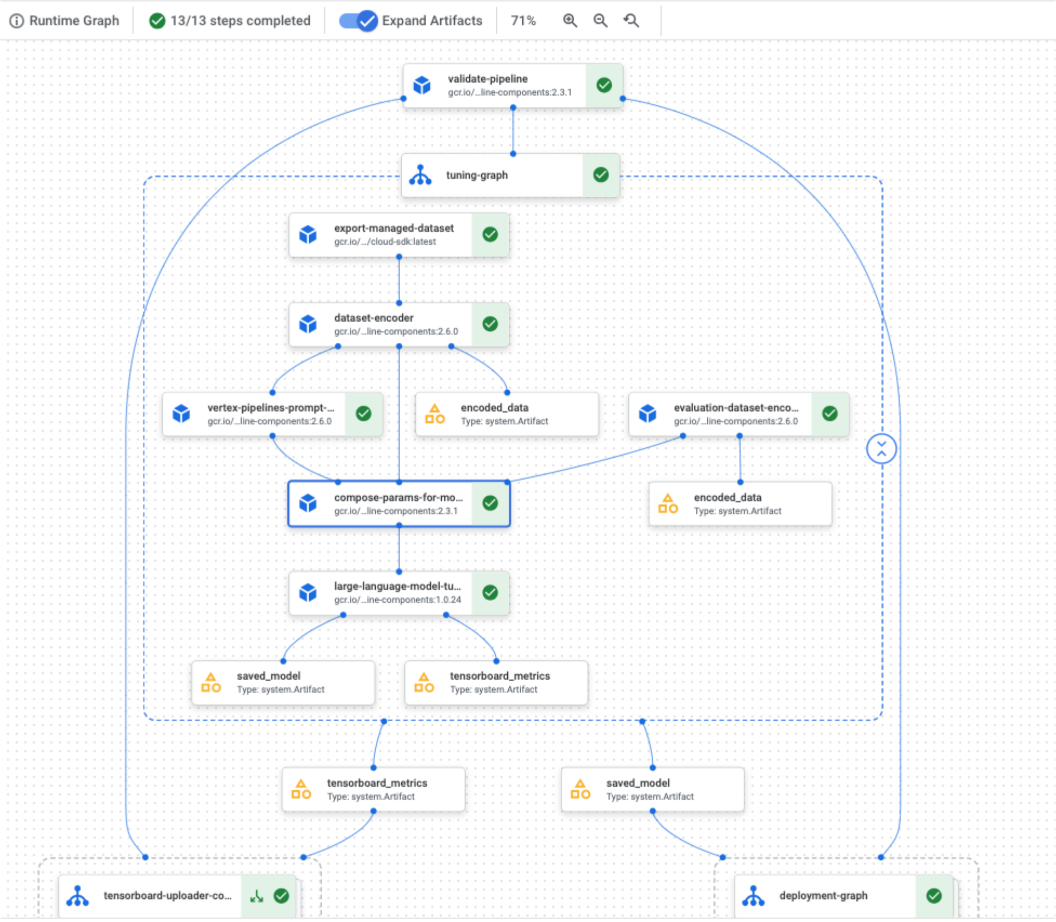


Figure 5.3: MH-PaLM Pipeline Run Analysis

AI's automated logging of **tensorboard_metrics** facilitated an uncomplicated monitoring experience via TensorBoard's insightful visualisations.

The **evaluation-dataset-enco-** suggested that an evaluation dataset was also crafted and encoded by the platform, staged and ready for the crucial performance assessment. Vertex AI's **tensorboard-uploader-co-** component seamlessly uploaded the training metrics to TensorBoard, streamlining my review and analytical tasks. Culminating the fine-tuning narrative was the **deployment-graph**, signalling the steps Vertex AI enacted to prime MH-PaLM model for deployment, a testament to the platform's automation in serving models for prediction.

The interface confirmed that 13/13 steps were completed, tallying a perfect 100 percent, signalled the successful culmination of the training. The array of green check marks across the board symbolised the triumph of each step. The pipeline, as depicted in the Figure 5.3, is a visual testament to the automated and orchestrated dance of Vertex AI's cloud-based ML platform. It allowed me to immerse myself in the code, while the platform adeptly handled the intricate ballet of the training process, from inception to deployment.

5.2.1.3 MH-PaLM Performance Evaluation

Performance evaluation of the MH-PaLM model was an integral component of our research, providing insights into the model's capability to interpret and analyse mental health narratives accurately. Our approach involved configuring a model tuning job in Vertex AI to systematically collect tuning and evaluation metrics. To facilitate this, we connected our model tuning job to Vertex AI TensorBoard, utilising a specific instance ID along with an appropriately prepared evaluation dataset.

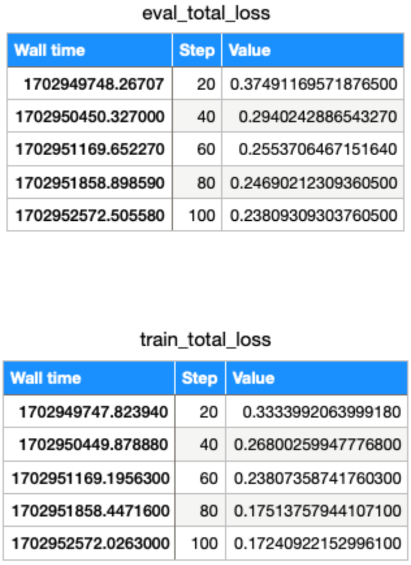
Preparation of the Evaluation Dataset

The evaluation dataset was constructed with careful attention to relevance and precision, consisting of prompt and ground truth pairs that closely mirror the tasks we aimed to evaluate. Comprising 800 pairs, our dataset was formatted to provide meaningful metrics that could substantiate the effectiveness of the MH-PaLM model. This JSON Lines (JSONL) formatted evaluation dataset was hosted on Google Cloud storage, conveniently located in the same US-central1 region as our resources, adhering to the maximum token lengths for both prompts and ideal responses.

Evaluation and Training Loss Metrics

The evaluation loss, measured as '/eval_total_loss', provided us with a clear indicator of the model's performance on the evaluation dataset illustrated in Figure 5.4. Observing the reduction in loss from 0.3749 at step 20 to 0.2381 at step 100, we could infer a significant improvement in the model's ability to predict the ground truth over time. The consistent downward trend in evaluation loss, particularly the decline from 0.2940 at step 40 to 0.2469 at step 80, attests to the model's increasing accuracy and fine-tuning efficacy.

In parallel, the training loss, noted as '/train_loss', was equally indicative of the model's learning trajectory. Starting at 0.3334 at step 20, we witnessed a substantial decrease to 0.1724 by step 100. This decrease, especially notable between step 40's 0.2680 and step 80's 0.1751, emphasises the model's growing adeptness at understanding and assimilating the training dataset.



eval_total_loss

Wall time	Step	Value
1702949748.26707	20	0.37491169571876500
1702950450.327000	40	0.2940242886543270
1702951169.652270	60	0.2553706467151640
1702951858.898590	80	0.24690212309360500
1702952572.505580	100	0.23809309303760500

train_total_loss

Wall time	Step	Value
1702949747.823940	20	0.3333992063999180
1702950449.878880	40	0.26800259947776800
1702951169.1956300	60	0.23807358741760300
1702951858.4471600	80	0.17513757944107100
1702952572.0263000	100	0.17240922152996100

Figure 5.4: MH-PaLM Eval loss And Training Loss

Interpretation of Loss Reduction

The reduction in both evaluation and training loss is significant. It demonstrates the model's enhanced capability to not only grasp the complexities of the MH-RAC but also to generalise its predictions effectively when confronted with new data. The fact that the training loss is slightly lower than the evaluation loss at corresponding steps suggests that our model is learning the training data well without over fitting. This is crucial for maintaining the model's robustness when applied to real-world scenarios outside of the training and evaluation datasets.

MH-PaLM Performance Metrics

Following the promising trajectory indicated by the loss metrics, we extended our evaluation to encompass additional performance indicators critical to our classification task. Precision, recall, and the Micro-F1 score as indicated in Table 5.1, which collectively form the cornerstone of classification evaluation, were meticulously measured to provide a holistic view of the MH-PaLM model's performance.

Precision, in this context, stood at 79 percent, demonstrating the model's robustness in identifying relevant instances of mental health narratives accurately. This high precision

Label	Precision	Recall	F1 score
All labels	79.38%	79.00%	0.79
Drug/Alcohol	86.36%	86.36%	0.86
Early Life	87.50%	87.50%	0.88
Personality	85.71%	85.71%	0.86
Trauma	63.64%	63.64%	0.64

Table 5.1: MH-PaLM Model Evaluation Metrics

rate reflects the model's ability to minimise false positives, ensuring that the majority of narratives it identifies as pertinent to mental health are indeed correctly classified.

Recall, measured at 79 percent, is equally crucial, as it assesses the model's success in capturing all relevant instances. The balanced precision and recall indicate that the MH-PaLM model is not only precise but also comprehensive in detecting pertinent narratives from the vast dataset..

The F1 score, which harmonies precision and recall, was determined to be 0.79. This score is pivotal for our task due to its sensitivity to the distribution of classes within our dataset. An F1 score synthesized from the balanced precision and recall values suggests that MH-PaLM not only identifies relevant instances with accuracy but does so consistently across the varied spectrum of mental health discourse presented in the dataset.

Class Specific Evaluation Moreover, class-specific evaluations further illustrate the model's capabilities. For instance, the Drug/Alcohol class exhibited a precision and recall of 86.36 percent, yielding an F1 score of 0.86. Similarly, the Early Life and Personality classes achieved precision and recall values of 87.50 percent and 85.71 percent, with corresponding F1 scores of 0.88 and 0.86, respectively. The Trauma class, while slightly lower, still managed a precision and recall of 63.64 percent, resulting in an F1 score of 0.64.

Combining the insights from the declining trends in both training and evaluation loss with the reported precision, our analysis presents a well-rounded understanding of the model's capabilities. The integration of these metrics endows us with confidence in the MH-PaLM model's refined tuning. It has demonstrated a marked improvement in learning from the MH-RAC and has substantiated its potential to generalise its understanding when presented with novel, real-world data.

In conclusion, the ensemble of loss metrics and classification performance measures precision, recall, and the Micro-F1 score paints a comprehensive picture of the MH-PaLM model's capabilities. The model exhibits not just a deepened understanding of the training data but also an adeptness in accurately classifying and recalling instances within the evaluation set. These results underscore the success of our fine-tuning efforts, indicating that the MH-PaLM model is well primed for deployment in applications involving the nuanced analysis of mental health narratives.

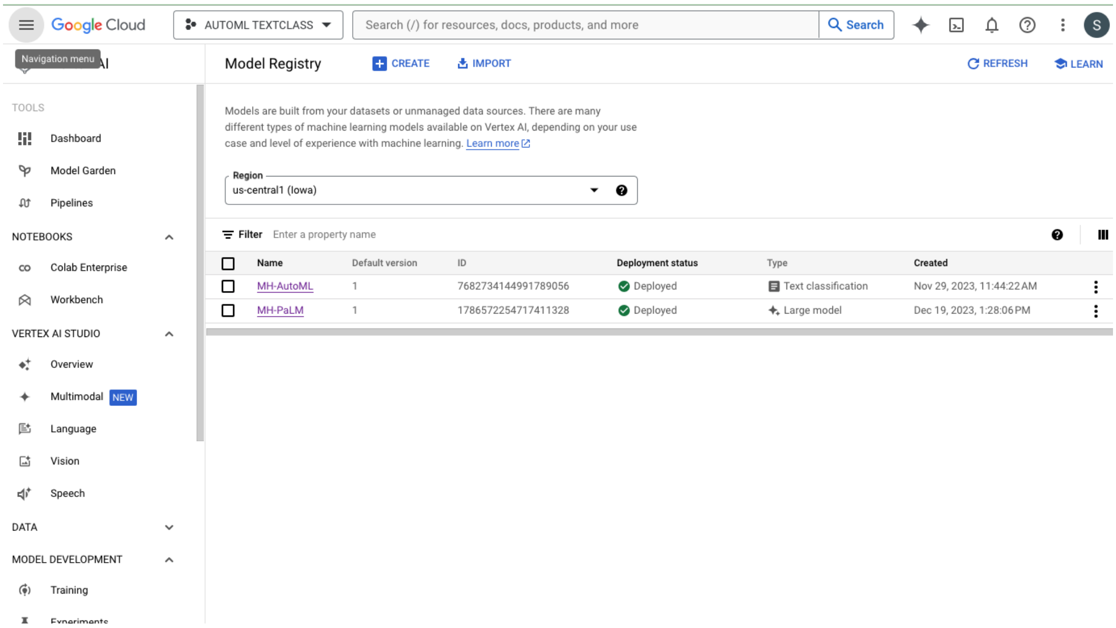
Our label-supervised fine-tuning aimed to teach the model to accurately interpret the nuanced lexicon inherent in mental health discourse. This approach not only enhances the model's contextual understanding but also significantly improves its responsiveness within this specific domain. The versatility of the text-bison model for classification, contributes to the broader objective of our study, demonstrating the model's capability to process and understand complex mental health narratives.

5.2.1.4 MH-PaLM Deployment

Deploying MH-PaLM to an endpoint was a crucial step for our research as it enabled the model to serve online predictions. This deployment process involved the allocation of the necessary physical resources to ensure low-latency responses for real-time inference. Utilising the Google Cloud console, we deployed our model seamlessly as shown in Figure 5.5.

Our approach was methodical: we created an endpoint that balanced resource availability and geographic considerations. The model was deployed to this endpoint with the intent to harness its computational resources efficiently. Within the deployment configuration, we made deliberate choices regarding the allocation of computational resources, setting a minimum number of compute nodes to maintain model readiness, and defining an upper limit to scale up to, should the demand for predictions increase. This auto scaling capability was instrumental in managing operational costs while ensuring responsiveness during peak load times.

A custom service account was used for the deployment, allowing for refined control over permissions and ensuring secure access to the resources. Upon finalising the configurations, the deployment to the selected region was initiated, chosen for its proximity to our



Models are built from your datasets or unmanaged data sources. There are many different types of machine learning models available on Vertex AI, depending on your use case and level of experience with machine learning. Learn more

Region: us-central1 (Iowa)

Filter Enter a property name

□	Name	Default version	ID	Deployment status	Type	Created
□	MH-AutoML	1	7682734144991789056	✓ Deployed	Text classification	Nov 29, 2023, 11:44:22 AM
□	MH-PaLM	1	1786572254717411328	✓ Deployed	Large model	Dec 19, 2023, 1:28:06 PM

FIGURE 5.5: Models Deployment Status

data sources, which would further reduce latency. In conclusion, our deployment process was executed with precision and foresight, setting the stage for robust online predictions and providing a foundation for comprehensive evaluation of MH-PaLM.

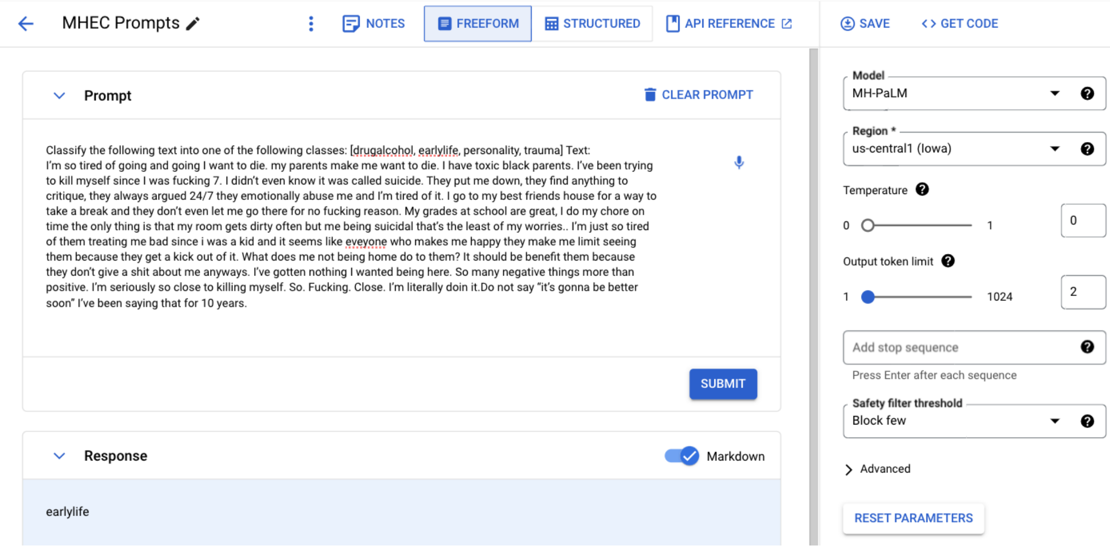
5.2.1.5 Prompt Design and Testing

Once the model was deployed, In the subsequent phase of our research, we dedicated our efforts to the nuanced discipline of prompt design, essential for eliciting precise responses from MH-PaLM post-deployment. Our approach was both methodical and empirical, crafted to probe the model's proficiency in content classification with a specificity that aligns with the capabilities developed during training.

We formulated a prompt that served a dual purpose as illustrated in Figure 5.6. It instructed MH-PaLM to perform classification while simultaneously providing the context necessary for it to understand the task at hand. This construction was deliberate; it served as a direct query to classify a given text into predefined categories, reflecting common root causes identified in mental health narratives.

Key Parameters for Prompt Tuning

In refining our prompt design to test MH-PaLM after deployment, we paid special attention to three critical parameters: Temperature, Top-K, and Top-P. These parameters are



Prompt

Classify the following text into one of the following classes: [drugalcohol, earlylife, personality, trauma] Text:
I'm so tired of going and going I want to die. my parents make me want to die. I have toxic black parents. I've been trying to kill myself since I was fucking 7. I didn't even know it was called suicide. They put me down, they find anything to critique, they always argued 24/7 they emotionally abuse me and I'm tired of it. I go to my best friends house for a way to take a break and they don't even let me go there for no fucking reason. My grades at school are great, I do my chore on time the only thing is that my room gets dirty often but me being suicidal that's the least of my worries.. I'm just so tired of them treating me bad since i was a kid and it seems like eveyone who makes me happy they make me limit seeing them because they get a kick out of it. What does me not being home do to them? It should be benefit them because they don't give a shit about me anyways. I've gotten nothing I wanted being here. So many negative things more than positive. I'm seriously so close to killing myself. So. Fucking. Close. I'm literally doin it.Do not say "it's gonna be better soon" I've been saying that for 10 years.

Response ☑ Markdown

earlylife

Model: MH-PaLM

Region *: us-central1 (Iowa)

Temperature 0 1 0

Output token limit 1 1024 2

Safety filter threshold: Block few

FIGURE 5.6: MH-PaLM Root Cause Prediction

instrumental in guiding the model's response behaviour, tailoring it to produce outputs that align closely with our experimental goals.

- **Temperature** is a pivotal parameter that influences the randomness of the predictions. A low temperature, set close to 0 in our experiments, steers MH-PaLM towards the most likely responses, reducing the variability in its output. This setting was crucial for our specific use case, where the aim was to classify text into discrete categories [drugalcohol, earlylife, personality, trauma] with high confidence. The deterministic nature of the output, achieved by minimising temperature, ensured that MH-PaLM's classifications were both repeatable and reliable, foundational qualities for validating the model's accuracy and utility in practical applications.
- **max_output_tokens** we strictly set to 2, as it is essential in understanding the precision of our prompt design and the consequent tuning of MH-PaLM's output for our classification task. This decision to limit the model's response to exactly two tokens was strategic and played a pivotal role in achieving the desired specificity and conciseness in MH-PaLM's responses. This parameter serves as a direct control over the length of the generated output, ensuring that MH-PaLM delivers its classification within the predefined constraint of two tokens. This limitation is particularly important in the context of our classification task, where each response is expected to be a concise label from among the specified categories (drugalcohol,

earlylife, personality, trauma). By setting this parameter to 2, we enforce a disciplined output structure that aligns perfectly with the categorical nature of our inquiry, allowing for no deviation or unnecessary elaboration that could detract from the model's precision.

- **Top-K** controls the diversity of the model's responses by limiting the selection pool to the K most probable next tokens. For our experiment, we set Top-K to 1, a choice that reflects our requirement for precision and determinism in the model's output. By restricting MH-PaLM to consider only the single most probable token for each decision point, we effectively guided it to produce the most relevant and contextually appropriate classification labels. This strict constraint was integral to our goal of achieving clear, unequivocal categorisations from the model, eliminating ambiguity and enhancing the interpretability of its responses.
- **Top-P**, another critical parameter, further refines the model's token selection process by choosing tokens cumulatively up to a specified probability threshold. Setting Top-P to a value close to 0 aligns with our stringent requirement for deterministic outputs. This configuration ensures that only the tokens with the highest confidence levels are considered, further reinforcing the model's focus on generating highly probable and consistent responses.

The combination of these parameters low temperature, Top-K set to 1, and Top-P close to 0 was chosen to underpin our classification task. These settings coalesce to form a prompt design strategy that emphasises precision, repeatability, and clarity in MH-PaLM's outputs. Such an approach is indispensable when the objective is to distill complex natural language inputs into categorical classifications that bear significant implications for understanding and interpreting mental health-related text.

Our exploration and application of these parameters underscore the nuanced balance required to fine-tune large language model outputs for specific tasks. By recommending these values and elucidating their impact on MH-PaLM's performance, we aim to provide a blueprint for future research and application development, guiding practitioners in harnessing the full potential of advanced language models in nuanced and domain-specific tasks.

5.2.2 AutoML Approach

AutoML is revolutionising the way we approach ML challenges. Its key objective is to simplify the complex process of model development by automating crucial steps such as selecting the right algorithms and fine-tuning their hyperparameters. In recent years, AutoML has gained significant traction for its ability to streamline various ML tasks, including data pre-processing, model selection, and hyperparameter optimisation. This level of automation enables the system to recommend the most effective model pipelines with minimal input from the user, making advanced ML techniques more accessible to a broader audience.

The operational model of AutoML solutions generally involves creating and evaluating multiple pipelines using sophisticated optimisation techniques like Evolutionary Algorithms or Bayesian Optimisation. These pipelines are rigorously assessed using performance metrics such as accuracy or F1-Measure, often utilising cross-validation methods to ensure comprehensive evaluation.

In our exploration of mental health narratives on social media, we employed the Google Cloud AutoML, a state-of-the-art AutoML system [190]. It has demonstrated its benefits in hyperparameter tuning and algorithm selection in several ML scenarios [191]. This framework facilitated the development of a sophisticated text classification model, adept at parsing and categorising complex mental health discourse from MH-RAC. Central to our methodology was AutoML's capacity for automated pipeline generation, encompassing data pre-processing, model architecture selection, and hyperparameter tuning, which collectively expedited the model development life cycle. The model's efficacy was predicated on the AutoML system's algorithmic ability to navigate this diversity and discern latent thematic structures within the data.

While the ease of use offered by Google Cloud AutoML is a significant advantage, it comes with certain limitations. Users might not have complete visibility into the intricate processes of text processing, model training, and hyperparameter optimisation. However, these trade-offs are often outweighed by the platform's ability to automate model development and deployment, making it a valuable asset in our research. The platform's capabilities are particularly vital in analysing our mental health-focused dataset, where it processes and interprets large volumes of nuanced text data, thereby playing a crucial role in our understanding of mental health narratives on social media.

Model Relevance

The selection of AutoML, particularly for text classification, as a key component in our project stems from its unique capabilities and alignment with our research objectives. Here's why AutoML was particularly suited for analysing our labelled training data.

Suitability for Complex Text Classification

- Efficiency in Handling Labelled Data: MH-RAC, comprising 800 labelled data points across four distinct mental health themes, required an efficient and accurate classification system. AutoML excels in managing labelled datasets, making it ideal for our project's needs.
- Adaptability to Varied Classes: The diversity of MH-RAC, with classes including trauma, earlylife, personality, and drugalcohol necessitates a model that can adeptly handle multiple, distinct categories. AutoML's ability to adapt to various classes and train accordingly is a significant advantage.

Streamlining the Model Development Process

- AutoML simplifies the model development process by automatically selecting the most appropriate models and optimising their hyper parameters. This feature is particularly beneficial for our project, as it removes the need for extensive trial-and-error in model selection and tuning.

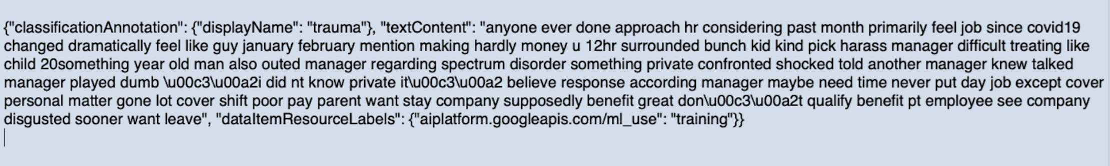
Enhancing Analytical Accuracy

- Cross-Validation for Reliable Performance: AutoML's use of cross-validation techniques ensures a more reliable estimate of the model's performance. This is crucial in a project like ours, where accurate classification is key to deriving meaningful insights.
- Optimised for Text Classification Tasks: Given that our project is centred around text classification, AutoML's optimisation for such tasks ensures that the models it generates are particularly well-suited for our specific research needs.

The selection of Google Cloud AutoML, for our project was driven by its efficiency in handling labelled datasets, adaptability to diverse classes, automation in model selection and tuning, and its user-friendly nature. These features align perfectly with our project's requirements, making AutoML an indispensable tool in our research on mental health-related text classification.

5.2.2.1 Data Formatting for AutoML

The foundation of any ML model is its data. We began by thoroughly extracting and cleaning MH-RAC. This was followed by exploratory data analysis (EDA) to understand the intricacies of the data, which is vital for effective model training. We conducted data transformations and feature engineering to make the data conducive to the ML model's learning process.



{"classificationAnnotation": {"displayName": "trauma"}, "textContent": "anyone ever done approach hr considering past month primarily feel job since covid19 changed dramatically feel like guy january february mention making hardly money u 12hr surrounded bunch kid kind pick harass manager difficult treating like child 20something year old man also outed manager regarding spectrum disorder something private confronted shocked told another manager knew talked manager played dumb \u00c3\u00a2i did nt know private it\u00c3\u00a2 believe response according manager maybe need time never put day job except cover personal matter gone lot cover shift poor pay parent want stay company supposedly benefit great don\u00c3\u00a2t qualify benefit pt employee see company disgusted sooner want leave", "dataItemResourceLabels": {"aiplatform.googleapis.com/ml_use": "training"}}

Figure 5.7: AutoML Training DataSet Format

The MH-RAC was formatted as JSON Lines (JSONL) files, adhering to a schema as shown in Figure 5.7 optimised for the AutoML environment. This precise formatting ensures the data is accurately parsed by the system, enhancing the efficiency of the training and evaluation process. The structure of the JSONL files was strictly designed as follows to align with the AutoML specifications. The structure of the JSONL files adhered to the following format:

```
{
  "classificationAnnotation": {
    "displayName": "label"
  },
  "textContent": "inline_text",
  "dataItemResourceLabels": {
    "aiplatform.googleapis.com/ml_use": "training/test/validation"
  }
}
```

In this schema, the ‘displayName’ attribute corresponds to the classified label from the MH-RAC, and the ‘textContent’ field encapsulates the actual textual content drawn from Reddit discussions. The ‘dataItemResourceLabels’ tag delineates the usage category of each entry, facilitating an effective stratification of the dataset into training, validation, and test sets. A randomised split was implemented to distribute the data points to ensure equitable class representation across training, validation, and test sets following an 80/10/10 distribution. It is essential to understand how Vertex AI uses the dataset to create a custom model tailored to our specific needs. The automated data splits played a vital role in this process: The training set was the primary source of learning, representing the scenarios and patterns the model needs to learn. The validation set served as a gauge for model performance during training, allowing for hyper parameter tuning and structural adjustments without compromising the model’s ability to generalise. The test set was completely untouched during training and acted as a stand-in for real-world data, offering us an unbiased evaluation of the model’s predictive prowess. Choosing the default split was a strategic decision to leverage Vertex AI’s built-in mechanisms for data management, freeing us from the complexities of manual data splitting and allowing us to focus on higher-level tasks and research questions.

DataSet Creation

With the MH-RAC ready, we used the following steps to create and import data into Vertex AI:

- Accessed the Datasets page in the Vertex AI section of the Google Cloud console.
- Created an empty dataset and specified the data type as Text for single-label classification, reflecting our need for a definitive categorisation of text data.
- Selected the appropriate region, us-central1, for consistency and compliance with our resource locations.

5.2.2.2 AutoML Fine-Tuning Process

- We proceeded to the Model Registry page within the Vertex AI section of the Google Cloud console.

- Ensuring the region was set to 'us-central1', we began the process to train a new model. We navigated through the 'Train new model' interface, which included specifying model-related details and confirming the start of the training process.
- The training process was expected to take several hours, Our Model took 14hours and 59 seconds for training, a duration justified by the extensive nature of our dataset and the complexity of the classification task.

Throughout the training, Vertex AI employed its robust AutoML capabilities to tune the model's parameters, optimising for performance based on the patterns learned from the MH-RAC. By automating feature selection, neural architecture search, and hyper parameter tuning, Vertex AI aimed to produce a model tailored to our specific classification needs. This automated approach not only streamlined the model training but also aimed to enhance the model's predictive performance significantly.

5.2.2.3 MH-AutoML Performance Evaluation

Following the training phase, the evaluation of the MH-AutoML model was imperative to assess its performance and validate its efficacy for our research. This enabled us to understand the effectiveness of the model in classifying mental health narratives. The summary we received post-training provided an in-depth look at how the model performed on the test set, using standard ML metrics to quantify its predictive capabilities.

MH-AutoML Evaluation Metrics Quantitative measurements provided by model evaluation metrics were instrumental in assessing how the MH-AutoML model performed on the test set. The interpretation of these metrics was closely tied to our specific requirements and the particular nuances of the problem we aimed to solve with our model. The balance between precision and recall was especially important in our context. Precision measures the accuracy of the model in assigning the correct label out of all labels it has assigned, while recall indicates the model's ability to correctly identify all relevant instances that should have been labelled.

Upon setting a classification threshold of 0.5, we reviewed the precision-recall curve for individual labels, which provided a granular view of the model's performance across different categories.

Label	Average Precision	Precision	Recall	F1 score
All labels	0.927	86%	86%	0.86
Drugalcohol	0.986	96%	96%	0.96
earlylife	0.943	87%	80%	0.833
Personality	0.902	84.6%	88%	0.863
Trauma	0.863	76.9%	80%	0.784

TABLE 5.2: MH-AutoML Model Evaluation Metrics

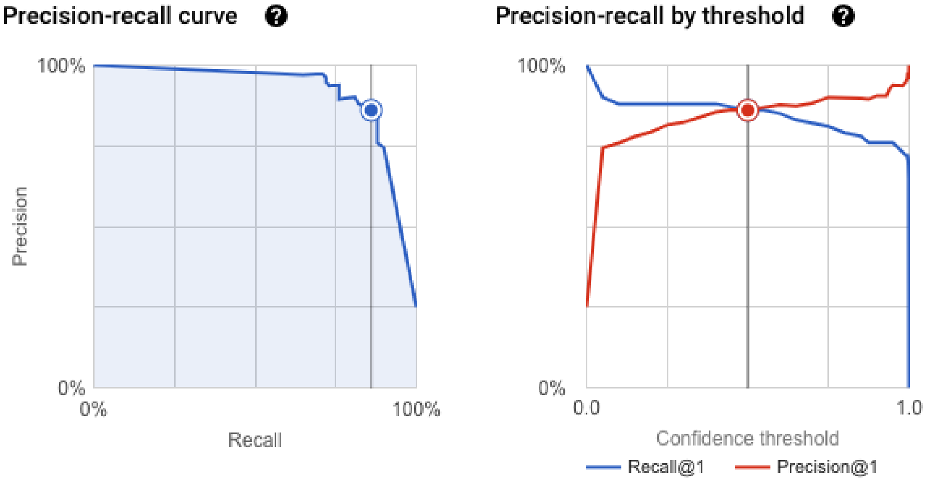


FIGURE 5.8: MH-AutoML Precision-Recall curve and Threshold

For all labels combined, the model demonstrated remarkable accuracy, as evidenced by the following metrics:

- **Average Precision:** At 0.927, this metric underscored the model's strong predictive power, significantly surpassing the baseline average precision of 0.5, which would indicate random guessing.
- **Precision:** The model achieved an 86 percent precision rate, signifying a high level of correctness in the labels it applied to the test data as reflected in Figure 5.8.
- **Recall:** With a recall rate also at 86 percent, the model proved its effectiveness in capturing the majority of relevant instances.
- **F1 Score:** The harmonic mean of precision and recall, the F1 score, was 0.86, reflecting the model's balanced classification capability.

Class specific Evaluation

When dissecting the results by specific labels, the MH-AutoML model's performance varied, indicative of its ability to adapt to the complexity and distinctiveness of each label as illustrated in Table 5.2 :

drugalcohol

- Average Precision: A high score of 0.986 showed exceptional model accuracy in this category.
- Precision: At 96 percent, the precision for this label was notably high.
- Recall: Also at 96 percent, indicating nearly all relevant instances were correctly identified.
- F1 Score: A near-perfect score of 0.96, suggesting outstanding model performance.

earlylife

- Average Precision: Scored 0.943, suggesting the model performed very well.
- Precision: Achieved a commendable 87 percent precision.
- Recall: At 80 percent, indicating the model's strong but slightly less effective capture rate in this category.
- F1 Score: Stood at 0.833, reflecting the model's solid performance for early life events.

personality

- Average Precision: At 0.902, indicating the model's robustness.
- Precision: With a precision rate of 84.6 percent, showcasing its reliable predictive ability.
- Recall: The recall of 88 percent pointed to the model's ability to identify the majority of relevant instances.

- F1 Score: An F1 score of 0.8627, indicating a balanced performance in identifying personality-related factors.

trauma

- Average Precision: Registered at 0.863, suggesting the model was quite effective in this category.
- Precision: With a precision rate of 76.9 percent, indicating room for improvement.
- Recall: The model had a recall rate of 80 percent, signifying a solid detection rate of relevant cases.
- F1 Score: The F1 score was 0.7843, reflecting the model's capability in trauma classification, albeit with some potential for enhancement.

The comprehensive evaluation of the MH-AutoML model presents a robust portrayal of its capabilities. With high scores across the board and particularly impressive results in certain labels such as drugalcohol, the model stands as a highly proficient tool for classifying mental health narratives. The precision-recall balance and the F1 scores for each label not only affirm the model's overall effectiveness but also reveal areas where further refinement could be beneficial. This evaluation solidifies the foundation for subsequent deployment and testing phases, setting a precedent for the applicability of the MH-AutoML model in real-world scenarios within the mental health domain.

5.2.2.4 MH-AutoML Deployment

Having successfully trained the MH-AutoML model as shown in Figure 5.5, the next critical steps were deployment to an endpoint and testing the model using unseen data from social media posts. The deployment process was as follows:

- We began by navigating to the Model Registry page in the Google Cloud console, where all trained models and their versions are listed.
- We ensured that the region was set to us-central1 (Iowa) to align with our data storage and processing region.

- By clicking on the name and version number of our trained MH-AutoML model, we accessed a detailed view of the model, including its performance metrics as observed on the Evaluate tab.

Creating an Endpoint:

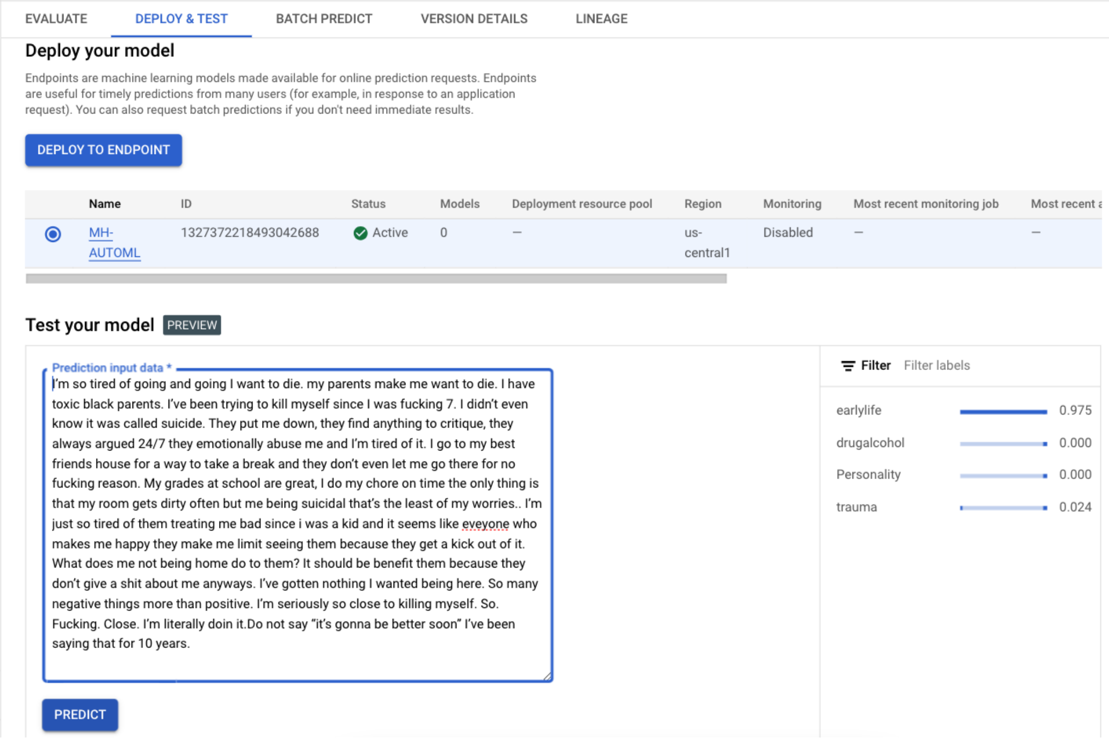
- We proceeded to the Deploy and test tab within the model details page to create an endpoint for our model.
- We selected 'Deploy to endpoint' and opted to create a new endpoint.
- We assigned 100 percent of the traffic to this new endpoint to ensure that all incoming inference requests were directed to our MH-AutoML model.
- After confirming our settings, we initiated the deployment process, which took several minutes to complete. During this time, Vertex AI provisioned the necessary resources and deployed the AutoML model to the newly created endpoint.

5.2.2.5 MH-AutoML Testing

With the MH-AutoML model deployed to an endpoint, we proceeded to the critical phase of testing, which involved evaluating the model's performance with new, unseen data from RMHD-Seed. This testing aimed to mirror the model's future application in real-world scenarios, providing insights into its predictive accuracy and reliability.

We utilised the Google's Vertex AI interface for testing the deployed MH-AutoML. This user-friendly interface facilitated the submission of individual posts for label prediction. A test set of 50 unique posts was selected, referred as RMHD-Seed, ensuring a variety of topics and sentiment that represented the diversity within the community on mental health issues. Each post was prepared according to the model's input requirements. We submitted them one by one through Vertex AI Test interface, carefully monitoring the model's response and recording each predicted label into MHEC. Model prediction testing can be seen in Figure 5.9. The real-time nature of this testing provided immediate feedback on the model's performance, essential for dynamic analysis.

Recording and Analysis:



Deploy your model

Endpoints are machine learning models made available for online prediction requests. Endpoints are useful for timely predictions from many users (for example, in response to an application request). You can also request batch predictions if you don't need immediate results.

	Name	ID	Status	Models	Deployment resource pool	Region	Monitoring	Most recent monitoring job	Most recent a
●	MH-AUTOML	1327372218493042688	✓ Active	0	–	us-central1	Disabled	–	–

Test your model PREVIEW

Prediction input data *

I'm so tired of going and going I want to die. my parents make me want to die. I have toxic black parents. I've been trying to kill myself since I was fucking 7. I didn't even know it was called suicide. They put me down, they find anything to critique, they always argued 24/7 they emotionally abuse me and I'm tired of it. I go to my best friends house for a way to take a break and they don't even let me go there for no fucking reason. My grades at school are great, I do my chore on time the only thing is that my room gets dirty often but me being suicidal that's the least of my worries.. I'm just so tired of them treating me bad since i was a kid and it seems like eveyone who makes me happy they make me limit seeing them because they get a kick out of it. What does me not being home do to them? It should be benefit them because they don't give a shit about me anyways. I've gotten nothing I wanted being here. So many negative things more than positive. I'm seriously so close to killing myself. So. Fucking. Close. I'm literally doin it.Do not say "it's gonna be better soon" I've been saying that for 10 years.

earlylife	0.975
drugalcohol	0.000
Personality	0.000
trauma	0.024

Figure 5.9: MH-AutoMl Root Cause Prediction

The testing of the MH-AutoML model against RMHD-Seed was an integral step in bringing our research from theoretical development into practical application. The MH-AutoMl demonstrated a commendable level of performance, with the live interface proving an effective tool for real-time predictions. The insights gained from this phase are pivotal for refining our understanding of the model's capabilities and limitations. As we continue to analyse the results and feed them back into the MHEC for deeper analysis, this testing phase marks a significant milestone towards achieving our thesis's objectives and contributing valuable knowledge to the field of mental health narrative analysis.

5.3 Comparative Evaluation Methodology: Human and AI-Driven Label Assignment

The evaluation process is fundamental to this study, providing a baseline for comparing the performance of our automated classification models (MH-PaLM and MH-AutoML) with human raters. By integrating assessments from a triad of evaluators, each offering a distinct perspective on interpreting mental health discourse in social media texts, this evaluation ensures that the models are not only clinically accurate and methodologically sound but also practically applicable in everyday situations. The subsequent comparison between human and AI-driven label assignments was focus on assessing the consistency, accuracy, and interpretative depth provided by each method, offering a comprehensive evaluation of the models' alignment with human judgement.

5.3.1 Human Rater Label Assignment

The process of label assignment was meticulously orchestrated to maintain the highest standards of methodological rigour. The RMHD-Seed, meticulously anonymised to preserve confidentiality, was uniformly formatted to ensure consistency in presentation. This corpus, henceforth referred to as RMHD-Seed, served as the foundation for the subsequent transformation into the MHEC upon annotation. It was distributed among human raters, each of whom was tasked with the independent classification of the content into one of four pre-determined categories:

1. drugalchol as Drug and Alcohol,
2. trauma as Trauma and Stress,
3. personality as Personality,
4. earlylife as Early Life.

The raters were furnished with foundational instructions that outlined the objective of their task and the framework of categories. In order to safeguard the integrity of individual interpretative processes, these guidelines were deliberately broad. This approach was designed to minimise any undue influence on evaluative judgement. Each rater undertook their task in isolation, a measure taken to preclude the possibility of inter-rater bias and, in turn, enhance the reliability of the analysis of interpretative variability. Moreover, they were obliged to furnish evaluative rationales; such qualitative insights are anticipated to augment the depth of the subsequent analysis, transitioning RMHD-Seed into a richly annotated MHEC.

5.3.2 AI-Driven Label Assignment

After fine-tuning on MH-RAC, both models were deployed to classify RMHD-Seed in a sequential and controlled manner using Google Cloud's advanced infrastructure. The output mechanism was tightly regulated, with each model constrained to assign one of the four stipulated labels. This process employed the fine-tuned capability of each model to parse and interpret the semantic and syntactic properties of the content, resulting in the most relevant category.

The deployment leveraged Google Cloud's high-performance infrastructure, which was essential for handling the computational demands of the AI models. This setup allowed the models to efficiently generate predictions while maintaining high accuracy, enabling a robust comparison with the human-generated labels.

5.3.3 Analysis Based Evaluation

With predictions obtained from both human raters and AI models, we will conduct a comprehensive analysis to evaluate the performance of the models. This analysis will focus on several key metrics:

- **Agreement Metrics:** We will use metrics such as Cohen's kappa to measure the agreement between the human raters and AI models, providing insights into how closely the models replicate human judgement.
- **Accuracy and Consistency:** We will assess the accuracy of the AI models in comparison to human raters, examining how consistently each model correctly identifies the relevant categories across different cases.
- **Interpretative Depth:** By analysing the evaluative rationales provided by human raters alongside the AI predictions, we will explore the interpretative depth of the models, assessing their ability to capture the nuances in mental health discourse.

The detailed analysis is to highlight the strengths and limitations of the AI models but also provides critical feedback for further refinement. By systematically comparing human and AI-driven predictions, we aim to ensure that the models are not only accurate but also applicable in real-world mental health scenarios.

5.4 Analysis Approach:Bridging Human and AI Understanding

Our analysis commenced by establishing a consensus benchmark, drawing on the collective judgements of three distinct human raters: a domain specialist, a trained annotator, and a layperson. This robust and representative baseline allowed for a consistent comparison against AI model interpretations. We employed Cohen's Kappa to measure inter-rater reliability, thus ensuring consistency in classifications across a spectrum of expertise, from domain specialists to the public. Our methodology synthesises quantitative element such as analysing agreement rates between AI predictions and the human consensus using confusion matrices with qualitative aspects, like examining narrative and linguistic contributions to AI discrepancies through content analysis. This dual-focused approach offers a comprehensive understanding of AI performance, pinpointing areas for enhancement.

5.4.1 Establishment of Consensus Benchmark

The consensus benchmark was meticulously developed from the collective judgement of the three human raters for each item. We aimed to identify a primary label agreed upon by at least two raters, documenting their independent reasoning. These consensus labels were assembled in a new dataset column, forming the foundation for subsequent analysis and AI model performance comparison.

5.4.2 Inter-Rater Reliability Analysis Approach

Consistency among human raters and between human raters and AI models (PaLM2 and AutoML) was assessed using Cohen's Kappa, a measure valued for considering chance agreement. In this study, we interpret Kappa scores above 0.60 as indicative of substantial agreement, reflecting a significant level of concordance. By including raters with diverse levels of expertise and treating AI models as quasi-raters, we constructed an agreement matrix to investigate the reliability and dynamics of human agreement.

5.4.3 Quantitative Alignment and Misclassification Patterns

Quantitative analysis focused on the alignment rate between AI model classifications and human consensus. We specifically examined how often AI model classifications corresponded with the consensus and scrutinised instances of divergence to comprehend the frequency and nature of discordance. Confusion matrices played a critical role, illuminating areas of potential confusion for both humans and AI, thereby informing areas where AI models may require additional training or adjustment.

5.4.4 Label-Specific Misclassification Analysis

To deepen our understanding of where and how the AI models PaLM2 and AutoML diverge from human consensus, we devised a focused analysis strategy on label-specific misclassification frequencies. This involved counting the number of times each model incorrectly classified a label compared to the human consensus. The frequency data serves as a quantitative measure of each model's alignment with human judgement on a label-by-label basis. This also provide a window to comparative evaluation between PaLM2 and AutoML models for each label, highlighting where each model exhibits strengths or weaknesses in specific areas of classification.

5.4.5 Contextual Analysis and Comparative Evaluation

Contextual Analysis Based on Consensus Rationale: Focusing on labels identified as particularly challenging, we examine how models' predictions align with the thematic and narrative rationale provided by human raters. This analysis reveals each model's ability to understand thematic versus semantic nuances.

Identification of Factors Leading to Confusion: We identify narrative factors that contribute to classification confusion, assessing whether discrepancies are due to semantic reliance or thematic misinterpretation. This step scrutinises the models' alignment with human thematic understanding and highlights instances where semantic alignment leads to misclassification.

Comparative Contextual Evaluation:The Ultimate Synthesis: By comparing the models' predictive decisions against the raters' thematic rationales, we elucidate the cognitive processes AI employs in complex narrative contexts. This comparison helps us understand the extent to which AI models grasp nuanced human judgement, shedding light on potential areas for enhancing AI interpretive capabilities.

This chapter has covered AI model training and human analysis to interpret mental health narratives from social media. Starting from data pre-processing to sophisticated model training using Vertex AI and AutoMl, enhancing the interpretative capabilities of our models to capture the unique linguistic features of mental health discourse. Through precise model training with annotated data, we validated AI classifications against human judgements, reinforcing the reliability and accuracy of our analytical framework. The integration of human insights with AI's analytical power unveiled the strengths and limitations of current technologies, highlighting areas for further improvement to achieve human-level understanding. This synthesis not only highlights the technical advancements but also emphasises the ongoing need for interdisciplinary collaboration to refine AI tools for enhanced mental health research. Reflecting on our findings, the chapter sets a path forward for future enhancements in AI applications within mental health, aiming to develop more empathetic and effective tools that can better serve public health initiatives globally.

Chapter 6

Insights and Implications: AI and Human Analysis of Mental Health

In this chapter, we explore the intricate dynamics of mental health conversations on Reddit during the pandemic, highlighting both temporal trends and the complexities of data interpretation. Utilising advanced AI models alongside traditional human analysis, we dissect the precision of annotations and the reliability of these interpretations. This dual approach allows us to rigorously evaluate the effectiveness of AI in mental health classification and understand the nuanced interactions between human and ML perspectives.

6.1 Findings: Navigating the Narrative Seas

This section presents a detailed analysis of our data collection, annotation, model training and human vs AI interpretability analysis efforts. We begin by outlining the temporal trends observed in mental health discussions on Reddit since the onset of the pandemic. This is followed by an in-depth look at the precision of annotation efforts, discrepancies identified, and the challenges encountered in AI classification by label. The section also explores the quantitative alignment and misclassification patterns that emerge from the AI models, providing a granular view of the data's interpretative layers.

6.1.1 Temporal Trends of Mental Health Conversations on Reddit in Pandemic Era

Our analysis meticulously charted the fluctuations in mental health discourse across different stages on Reddit, employing a heatmap to visualise the shift in post volume through the pre-pandemic, pandemic, and post-pandemic periods, as illustrated in Figure 6.1.

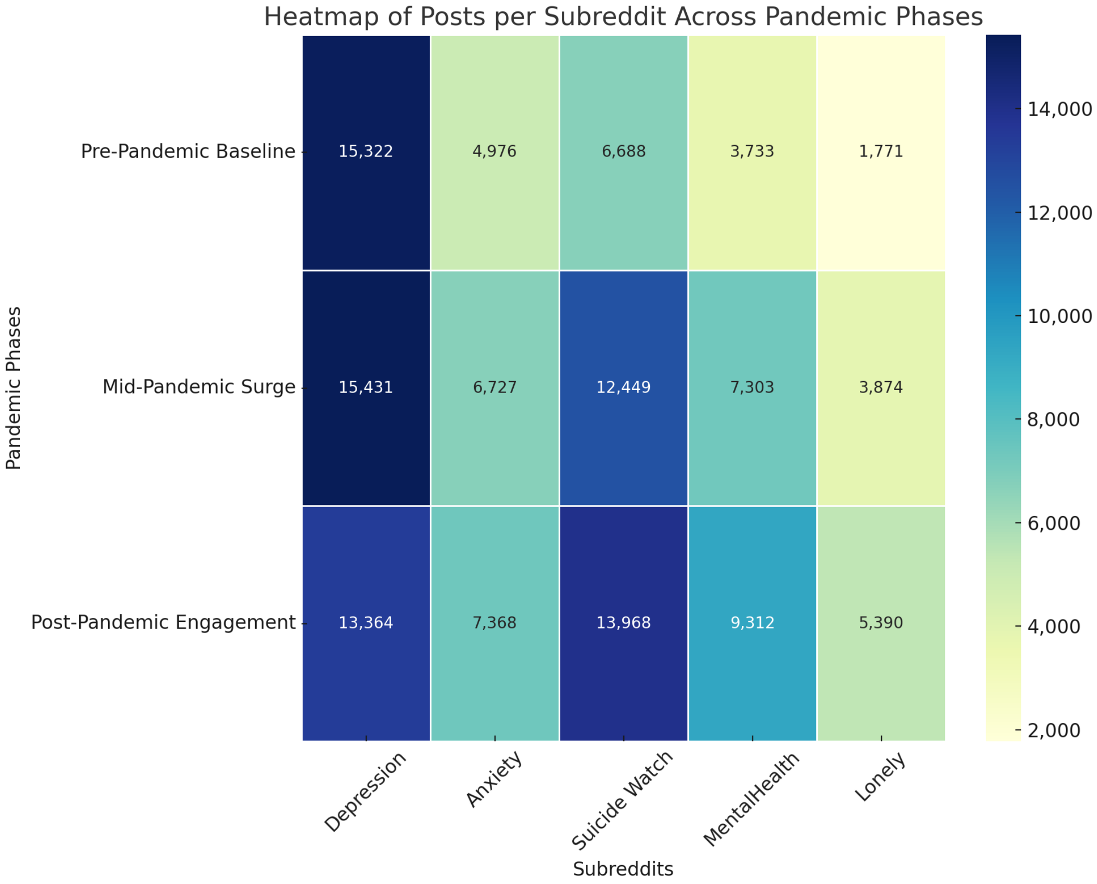


Figure 6.1: Heatmap Analysis of Average Posts per subreddit

6.1.1.1 Pre-Pandemic Context

Prior to the pandemic, subreddit activities formed a baseline for engagement, with r/depression notably having the highest frequency of posts per month. Activity within r/anxiety and r/suicidewatch was also significant, indicating an existing base of conversation about mental health issues.

6.1.1.2 Mid Pandemic Peak

The onset of the pandemic led to a sharp increase in posting activity, especially within r/anxiety, reflecting the escalating public anxiety and stress related to the pandemic's challenges. This increase is aligned with broader research findings that suggest a link between heightened social media use and an increase in reported anxiety symptoms globally [85]. Similarly, r/suicidewatch experienced a rise in activity, reflecting the exacerbation of mental health crises during this period. These trends emphasise the essential role of online platforms in providing support and enabling discussions amid increased feelings of isolation and anxiety due to the pandemic.

6.1.1.3 Post-Pandemic Participation

After the pandemic, post frequency declined but remained higher than pre-pandemic levels, particularly in r/anxiety and r/suicidewatch. This suggests a persistent impact of the pandemic on mental health discourse and a shift towards new norms in public conversation about mental health issues.

6.1.1.4 Heatmap Analysis:Visualising the Emotional Landscape

The heatmap in Figure 6.1 effectively demonstrates the dynamic shifts in engagement over time. The visual progression from darker to lighter hues from the mid-pandemic to the post-pandemic phases illustrates a persistent change in the landscape of mental health discourse, without a return to pre-pandemic engagement levels. The intensities of the colors, corresponding to post volumes, reveal that darker tones, indicative of heightened activity, are especially prominent in the r/depression subreddit across all phases, highlighting its constant significance as a hub for discussions. The r/anxiety subreddit demonstrated a significant peak in activity during the pandemic, mirroring the surge in public concern and discourse on anxiety. Notably, the level of discussion in the post-pandemic phase for both r/anxiety and r/suicidewatch subreddits remained elevated compared to the period before the pandemic, signifying a durable change in engagement patterns. The sustained activity in the r/suicidewatch subreddit post-pandemic underscores the ongoing need for support and dialogue platforms for crisis intervention.

These insights suggest that the pandemic may have irrevocably altered the manner in which individuals engage with and perceive mental health in public forums. This lays the groundwork for further investigation into the role of online communities in offering support, the evolving perceptions of mental health stigma, and the increased societal recognition of mental health issues.

6.1.2 Annotation Precision and Discrepancies

The process of annotating our dataset unveiled key patterns in the agreement levels between annotators, with significant findings outlined in Contingency Table 4.6. This analysis revealed a mix of agreement and disagreement among the annotators concerning the four main mental health root cause categories. A notable observation was the high degree of agreement in the "trauma and stress" category, where a remarkable 92.5 percent concurrence was achieved, underscoring the clarity and cohesion in identifying factors such as domestic violence, work-related stress, and the impact of relationship breakdowns. Illustratively, entries P4–P6 in Table 4.5 exemplify how financial difficulties, social isolation, and the end of significant relationships contribute to categorising posts under "trauma and stress".

Challenges arose, however, in clearly separating posts related to "personality" factors from those associated with "trauma and stress", highlighting areas of ambiguity. Posts related to "early life" experiences similarly presented some inconsistencies, whereas the "drug and alcohol" category experienced minimal discord, suggesting clearer delineation for annotators in this domain.

A specific case noted in Table 6.1 underscores the complexities inherent in mental health narratives, which can lead to divergent interpretations among annotators. For instance, one post was seen by Annotator 2 as embodying "trauma and stress" due to references to bereavement, a direct manifestation of emotionally taxing events. Conversely, Annotator 1 linked the sustained impact of the grief to adverse childhood experiences, thus aligning it with "early life" factors.

This difference underscores the nuanced distinction between enduring stressors originating in one's formative years and those associated with more immediate traumatic events. The "early life" category encapsulates influences from one's developmental stages that exert a long-term effect on mental health, as opposed to “trauma and stress”, which typically pertains to recent incidents and their direct emotional repercussions. Discrepancies such as these were methodically addressed through consensus discussions or, where necessary, expert consultation. This approach allowed for the harmonisation of diverse perspectives, bolstering the dataset’s annotation accuracy and consistency.

The Cohen’s Kappa statistic of 0.869 highlights an exemplary level of annotator agreement, reflecting the success of our rigorous training and comprehensive guidelines. Despite the overarching high concordance, instances of disagreement shed light on the intricate layers within mental health discussions. These moments of discord are vital, offering opportunities to refine our annotation methods and highlight the importance of continual enhancement to accurately capture the broad and nuanced spectrum of mental health narratives.

6.1.3 Effectiveness of AI Models in Mental Health Classification

In our assessment of artificial intelligence models for classifying mental health-related posts, the Google Cloud AutoML model emerged as highly effective following targeted fine-tuning. It demonstrated impressive results, achieving both Precision and Recall rates of 86 percent, and a matching F1 Score of 86 percent. As depicted in the confusion matrix (Figure 6.2), the AutoML model showed a high level of accuracy in categorising posts across various mental health conditions, especially notable in its performance within the ’drug and alcohol’ category, where it achieved an exceptional 96 percent accuracy. However, it was not without its limitations; particularly, it recorded an 80 percent accuracy in classifying ’trauma’ related posts, encountering some difficulties as evidenced by misclassification rates of 12 percent with ’early life’ and 8 percent with ’personality’ categories.

Conversely, the application of Google’s PaLM 2 model, although yielding slightly lower Precision, Recall, and F1 Scores at 79 percent, boasted a superior overall accuracy of 80 percent. The confusion matrix for PaLM 2 highlighted its robust classification capabilities across the board, with notably improved performance in the ’trauma’ category compared to the AutoML model. This balanced predictive efficiency across various categories indicates PaLM 2’s enhanced adaptability and generalizability, especially when evaluated against human benchmarks. In contrast, the AutoML model, despite exhibiting higher precision and recall rates during the training phase, exhibited tendencies that could be indicative of over fitting, suggesting it might perform less consistently in real-world applications.

This comparative analysis demonstrate the strengths and weaknesses inherent in deploying AI models for the nuanced task of mental health post classification. While AutoML showed unparalleled accuracy in certain categories, its potential for over fitting raises concerns for its applicability across diverse datasets. PaLM 2, with its more uniform performance and stronger generalizability, presents as a promising alternative, particularly for applications requiring consistent accuracy across a range of mental health topics. These findings highlight the critical importance of model selection based on the specific requirements and challenges of mental health discourse analysis, paving the way for further research into optimising AI models for nuanced content classification.

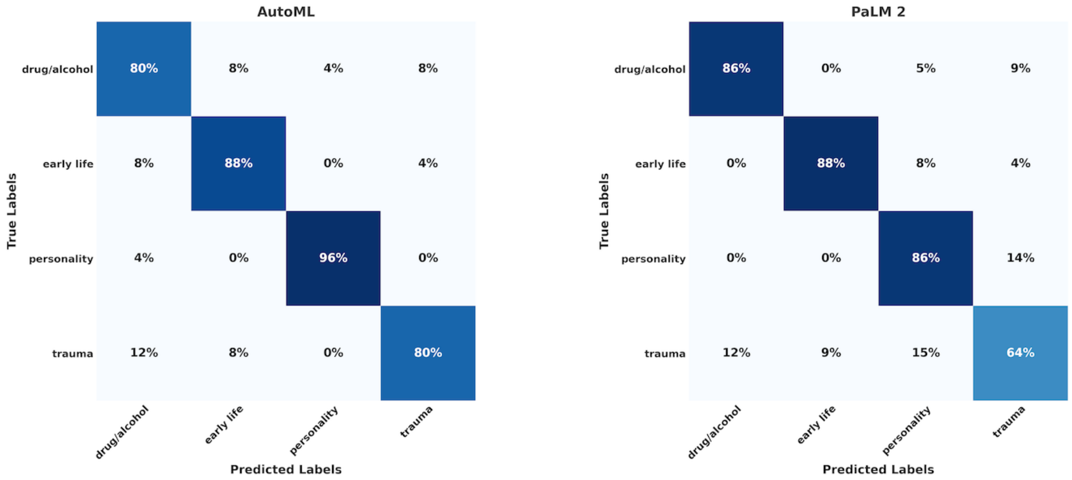


FIGURE 6.2: Comparative Performance of AutoML and PaLM 2 in Classifying Mental Health Root Cause

Table 6.1: Annotation Disagreement.

Annotator-1	Annotator-2	Post Text
Early life	Trauma and stress	*"It's been over a decade since my sister passed away and the next year prior to my sister's passing, my mom died. Two of them because of cancer. I was just 11 years old at that time. and now why am I still grieving? I am an adult now. Every time I watch shows or even read something about hospitals, diagnosis, griefs, I start cryinggg. Everything that reminds me of how they suffer, triggers this feeling. I just need to get out of this already. I have been grieving for so long, i am so tired. [Some text omitted for brevity]"*

Presenting a Reddit post with differing labels assigned by two annotators to illustrate an instance of annotation disagreement. This table is part of the analysis to understand and address the variability in annotator perspectives.

6.1.4 Analysis of Inter-Rater Reliability (IRR) Among AI Models and Human Raters

In our study, we embarked on an assessment of the concordance levels among human evaluators and between these evaluators and AI models, based on posts labelled for mental health evaluation categories of MHEC. The human raters displayed commendable agreement among themselves, with Kappa scores fluctuating between 0.60 and 0.78, suggesting a robust base for human consensus in the assessment of mental health-related content.

The rater agreement matrix, as visualised in Figure 6.3, shows that the PaLM 2 model maintained a consistently high degree of agreement with the human raters, with Kappa scores being 0.70, 0.76, and 0.70, accordingly. This suggests that PaLM 2's assessment of mental health categories aligns well with human judgement. The AutoML model, while still showing a respectable agreement with human raters, had slightly lower Kappa scores of 0.62, 0.68, and 0.72, respectively. Notably, the agreement between AutoML and PaLM 2 was the least at 0.60, indicating a disparity in classification approaches that warrants further investigation to decipher the factors contributing to this variance.

Our provision of clear Kappa score indicators and analysis of each model's relative performance is designed to deepen the understanding of the reliability of these models. Furthermore, these insights are pivotal for facilitating the refinement of automated classification systems, aiming to enhance their accuracy and their alignment with human evaluative standards.

The analysis highlights the potential of AI in mirroring human cognition in the classification of complex mental health discussions. However, it also emphasise the need for continuous calibration of AI models to ensure they remain consistent with the nuanced human understanding of mental health issues, especially given the variability seen between AutoML and PaLM 2. By iterating on these findings, future research can aim to reduce this gap, optimising AI systems for more precise and human-like performance in the analysis of mental health discourse.

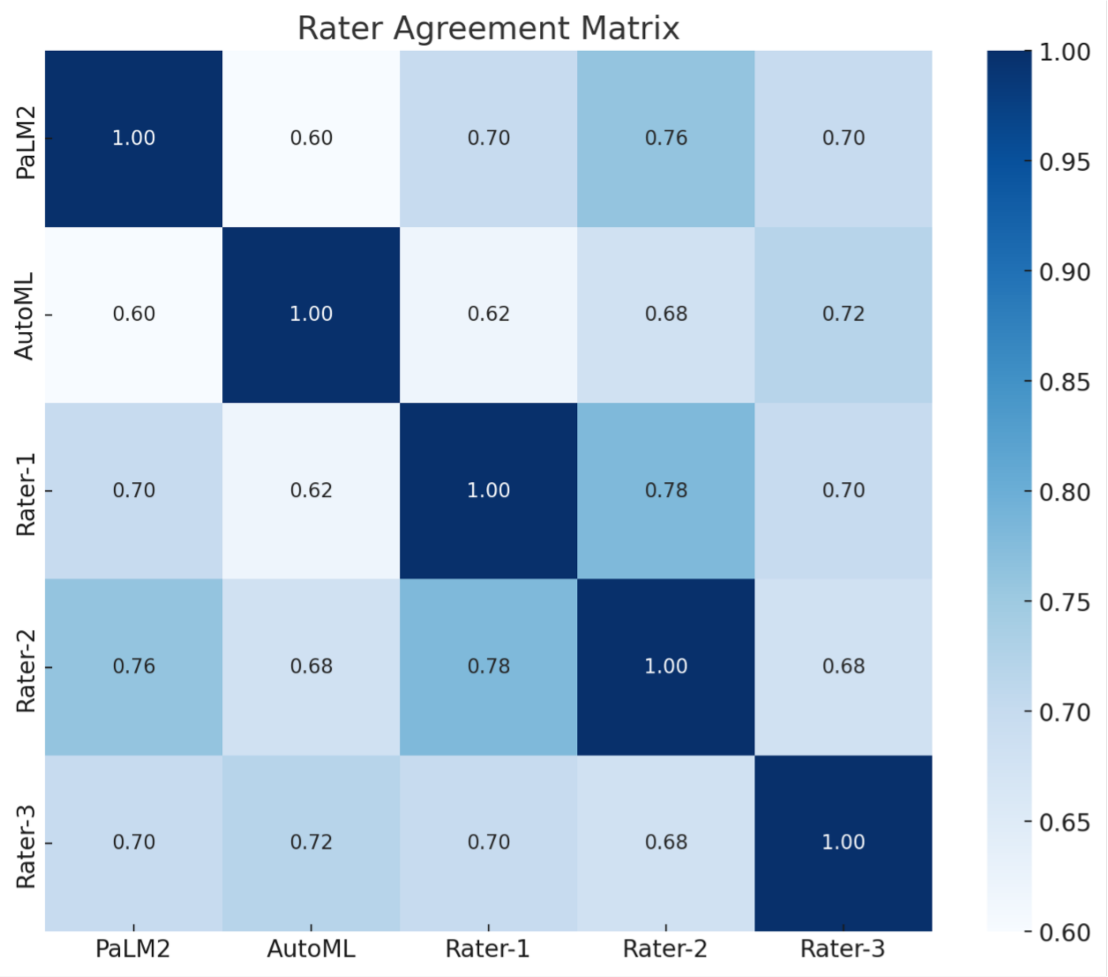


Figure 6.3: IRR Human-Machine Raters Agreement

6.1.5 Quantitative Alignment

Our analysis rigorously evaluated the alignment of the PaLM 2 and AutoML models with a benchmark established by human consensus, dissecting the rate of alignment and the pattern of discrepancies. This critical examination sheds light on the precise nature of the models' predictive capabilities, pinpointing areas ripe for enhancement. As detailed in Figure 6.4, the PaLM 2 model manifests a noteworthy 80 percent concordance rate with the human benchmark, indicating a substantial level of precision in its predictions. In contrast, the AutoML model aligns at a lower rate of 68 percent, signalling a divergence in its predictive accuracy.

This comparative analysis highlights the specific challenges and variances faced by each model in the task of categorising complex mental health discussions. While PaLM 2 demonstrates closer alignment with human evaluators, suggesting a more nuanced understanding of the subtleties in mental health narratives, AutoML's lower alignment rate implies a need for further calibration to reconcile its assessment criteria with human judgement.

These findings not only speak to the current effectiveness of these models in mimicking human analytical patterns but also underscore the potential areas for model optimisation. By identifying the strengths of PaLM 2 in its alignment with human raters and the shortcomings of AutoML in comparison, our research points to a targeted approach for refining AI capabilities in mental health classification tasks. The results, focused strictly on performance metrics, lay the groundwork for a more in-depth discussion on the implications of these findings for the future development of AI-driven diagnostic tools.

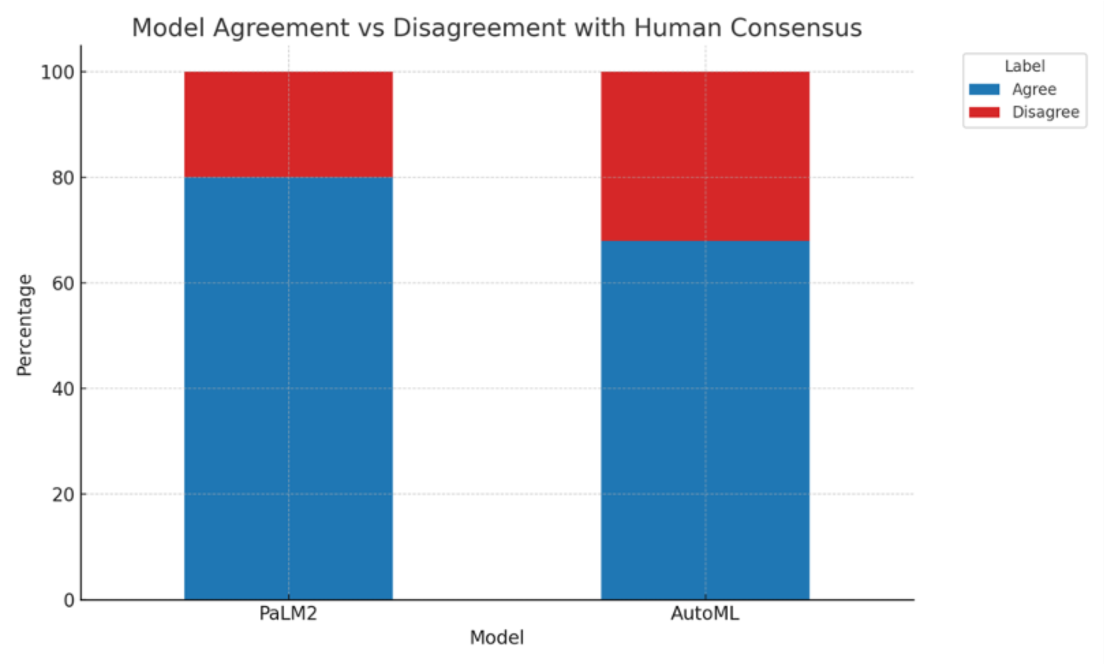


FIGURE 6.4: AI Models Agreement vs Disagreement with Human Consensus.

6.1.6 Misclassification Pattern Assessment

In our quantitative evaluation, we meticulously assessed the pattern of misclassifications made by PaLM 2 and AutoML as juxtaposed with the consensus of human raters. This assessment is vital to pinpointing the specific areas where each model diverges from human judgement and to understanding the nuanced complexities of mental health categorisation. As outlined in Figure 6.5, our analysis of the distribution of errors by label exposes distinct trends.

The 'Early Life' category surfaced as the most challenging for both models, with the highest frequency of misclassifications. This observation suggests a greater inherent ambiguity within the category, or a potential overlap of features that are not distinctly captured by the models. Conversely, the 'Drugs and Alcohol' category was identified with a lower error rate, implying that the characteristics of this category are more uniquely defined and more easily recognisable by AI systems.

In the case of 'Personality' and 'Trauma and Stress', the error rates were found to be equivalent, indicating these categories' intermediate complexity level. The models might struggle with discerning the finer nuances between these two categories, possibly due to shared symptoms or the interplay of contextual elements within the narrative structures of the posts.

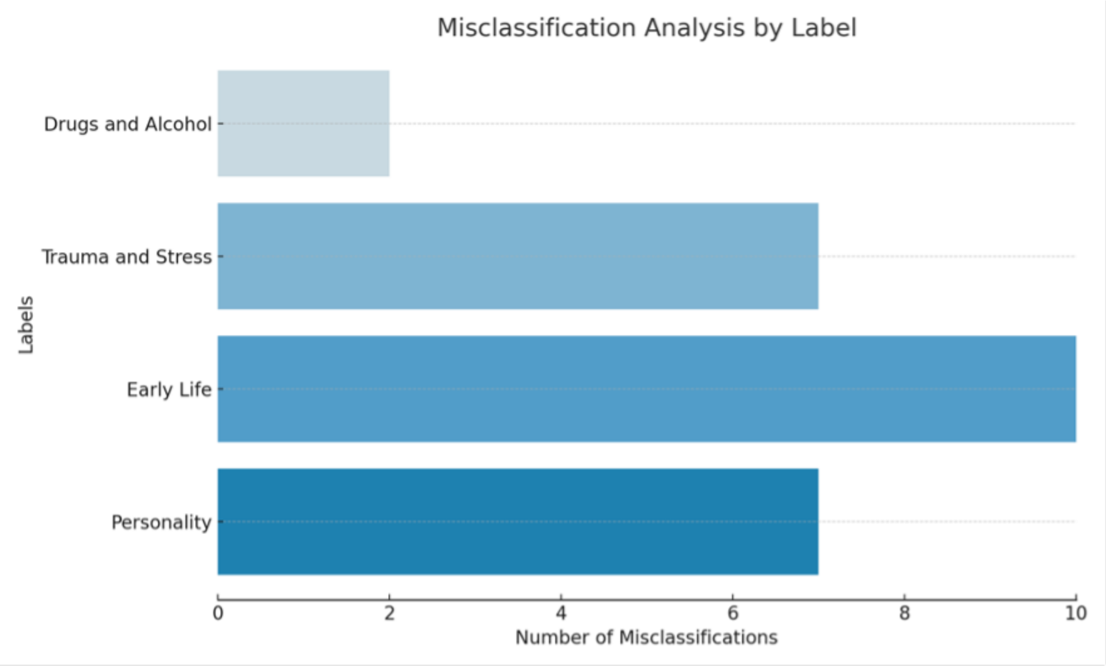


FIGURE 6.5: Total Number of Misclassifications by Labels

The confusion matrices, as represented in Figure 6.6, allow us to delve deeper into the comparative analysis of the two AI models against the human consensus. Here, we see that PaLM 2, although precise in pinpointing 'Drugs and Alcohol' related content, loses accuracy when it comes to 'Early Life' discussions, often conflating them with 'Personality' issues.

This misclassification points towards an intricate blending of narrative elements or symptomatic expressions that PaLM 2 cannot distinctly separate. Similarly, the model's difficulty in clearly differentiating 'Trauma and Stress' from 'Personality' reflects a complexity in distinguishing between the immediate and more ingrained, chronic aspects of these experiences.

The AutoML model parallels PaLM 2 in its competence with 'Drugs and Alcohol' content but displays confusion when discerning 'Early Life' influences, showing an equitable misclassification rate with 'Trauma and Stress' and 'Personality'. This pattern suggests a commonality in language or context between 'Trauma and Stress' and 'Personality' that AutoML is unable to resolve.

These findings highlight the challenges AI models face when dealing with the nuanced language and overlapping symptoms present in mental health discussions. They underscore the need for further refinement of these models to better parse the subtle complexities of mental health categories and align more closely with human expertise. Such enhancements are pivotal to advancing the precision and utility of AI in the nuanced field of mental health assessment.This analysis not only focuses the nuanced performance of AI models in content classification but also serves as a precursor to subsequent qualitative examinations

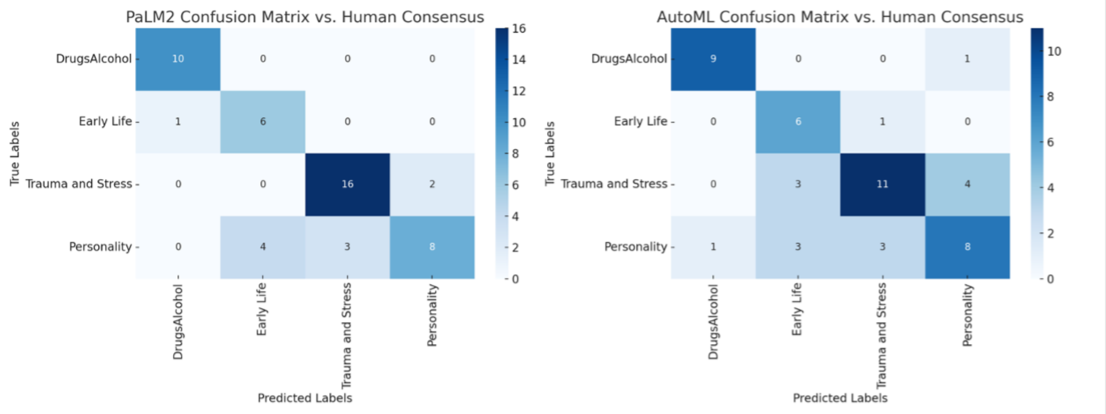


Figure 6.6: Confusion Matrix of AutoML and PaLM 2 against Human Consensus

6.1.7 AI Classification Challenges by Label

The study narrows its focus further to specific labels where AI classification has encountered challenges, providing a crucial pivot point for subsequent qualitative analysis. This detailed examination is pivotal for guiding future qualitative enhancements of the models. As illustrated in Figure 6.7, the data reveals that both models struggle with the 'Trauma and Stress' category, with PaLM 2 incorrectly classifying it three times and AutoML four times. This recurrent misclassification signifies an area where the models are challenged by the complexity of this particular label.

Further scrutiny of the "Personality" label discloses that PaLM 2 has misclassified it on two occasions, while AutoML's misclassification count is higher at five. This difference in performance between the two models indicates a variation in how they process and interpret the distinguishing features of this category. The 'Early Life' category also presents substantial difficulty for both models, evidenced by PaLM 2 misclassifying it four times and AutoML even more frequently, with six misclassifications.

These statistics are telling; they emphasise the struggle that both PaLM 2 and AutoML encounter when attempting to discern the intricate and often subtle narrative threads that define each label, particularly when these labels encompass overlapping themes or similar descriptive elements. The frequency of these errors provides a clear indication of where the models' current classification boundaries are being tested and where their understanding of the narrative content's nuances falls short.

This precise identification of problematic labels is an essential step toward enhancing the AI classification process. It reveals the specific junctures at which the models' algorithms fail to emulate the nuanced comprehension that human raters typically bring to the task. Recognising these areas allows us to pinpoint the most impactful opportunities for improvement in the models' design and training data, with the aim of reducing misclassifications and achieving a more refined and human-like performance in the categorisation of complex mental health narratives.

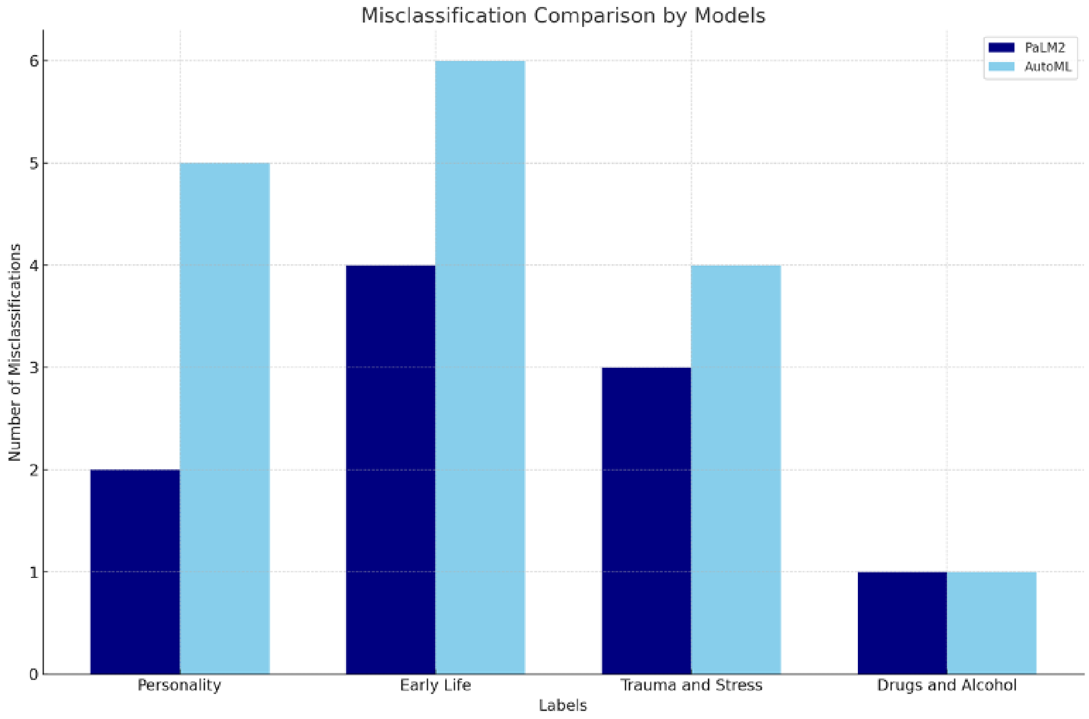


FIGURE 6.7: Missclassification Comparison of AutoML and PaLM 2

In the first instance of Table 6.2, we examine the complex narratives involved, aiming to clarify the confusion arising from comparing AI model interpretations with human raters. This examination is crucial, especially after identifying error patterns where the "Trauma and Stress" and "Personality" categories intersect. Such intersections highlight the challenges in discerning the nuanced boundaries of mental health narratives. To understand the depth of these complexities, we investigated the context within which discrepancies arise, revealing the layers of interpretation that distinguish human raters' nuanced judgements from AI determinations.

The overlapping of the 'Trauma and Stress' and 'Personality' labels in this instance is representative of the complex challenge inherent in mental health classification. On one hand, the label 'Trauma and Stress' is substantiated by the event-driven narrative where the pregnancy and its subsequent psychological impact are framed as a traumatic experience with enduring stress. This interpretation is supported by the presence of keywords and phrases commonly associated with postnatal depression, a specific sub type of trauma-related stress disorder. On the other hand, the 'Personality' label finds its justification in the portrayal of traits such as persistent self-deprecation and motivational decline, which could suggest an underlying personality predisposition. The

Instance	**PaLM 2**	**Auto ML**	**Rater -1**	**Rater -2**	**Rater -3**	**Posts**
1	Personality	Trauma and Stress	Trauma and Stress	Personality	Trauma and Stress	*"My poor mental health has been running my life for the past three years and I'm so tired of it. Ever since I became pregnant with my second child, I haven't been myself. I lost a lot of my motivation and desire to do much of anything and apart from rare moments, it hasn't really changed. I went to therapy and that helped me get out of a desperate place. I'm on meds but still feel numb. Also gained weight and can't find the motivation to diet and exercise like I did before. This past year I tried going back to working full time in the hopes that it would help me feel like myself, butI couldn't handle the job and had to quit. Now I'm just sitting on my couch surrounded by a messy house and it's hard not to hate myself. I don't have the energy to clean it or play with my kids.I'm afraid my husband is going to start resenting me. After three years it's hard to believe I'll ever just snap out of it and be normal again. I don't have any real reason to feel this way. I'm just mentally weak. [Some text omitted for brevity]”*
2	Trauma and Stress	Early Life	Trauma and Stress	Trauma and Stress	Trauma and Stress	*“I'm a freshman in high school and it's already so challenging for me. I don't know how to talk to my counsellor and every day gets harder. The big exams are a week away and I don't understand anything. I'm really considering killing myself but the only thing that's holding me back is fear and the sake of my mother because I know she'd be heartbroken. I'm thinking about overdosing but it's probably not enough. I just don't want to put a burden on my mom anymore.[Some text omitted for brevity]”*

TABLE 6.2: Annotating Posts with Rationale Reflection

language used by the individual reflects a pattern of self-perception and behaviour that extends beyond the immediate context of the traumatic event, potentially indicating a more deeply rooted personality structure. The intersection of these two classifications suggests that such labels do not exist in isolation but often in a state of confluence, where symptoms and experiences that can be interpreted through multiple diagnostic lenses. The challenge for AI models and human raters, therefore, lies in discerning the primacy or interdependence of these factors. Improving accuracy in these complex scenarios necessitates algorithmic advances capable of discerning between language related to personality traits and cues of trauma and stress. It also underscores the need for a multidimensional training dataset that captures a broad spectrum of human experiences. For AI models to accurately mirror human classification practices, they must recognise the complex relationship between individual personality characteristics and situational stressors.

6.1.8 Comparative Model Interpretation Analysis

Following our initial analysis, we proceed with a comparative review of how different AI models interpret specific narratives, focusing on a case highlighted in second instance of Table 6.2. This case involves a high school freshman dealing with academic pressures and suicidal thoughts. We contrast the interpretations of the PaLM 2 and AutoML models to emphasise the importance of contextual understanding in AI classification. PaLM 2 accurately identifies the narrative's main concern as 'Trauma and Stress,' recognising the urgency and specific stressors described. This reflects PaLM 2's ability to grasp the immediate context, showing an understanding of the narrative's acute stressors. In contrast, AutoML's misclassification under 'Early Life' points to a reliance on certain keywords rather than the full context, indicating a need for improvement in how models evaluate narrative content. The PaLM 2 model's correct categorisation as 'Trauma and Stress' indicates its nuanced understanding of the narrative, recognising the severe stress related to academic pressures and suicidal ideation. It appears to correctly weigh the narrative's urgent language and context, such as the mention of exams and thoughts of overdose, as indicators of a current crisis rather than long-term developmental issues. AutoML's incorrect classification suggests it might prioritise keywords like "freshman in high school" over the narrative's situational context, mistaking the current crisis for developmental challenges. This misinterpretation shows a gap in AutoML's ability to analyse the narrative's temporal and contextual elements, possibly due to differences in training data or algorithmic emphasis. The case underscores the critical need for AI models to balance keyword detection with a deeper understanding of context. Training data should include a range of stress-related situations to improve classification accuracy. Models need to be refined to better interpret the timing and severity of stressors, especially when signs of a crisis are evident, as PaLM 2 demonstrated in this instance.

This analysis highlights the differing abilities of PaLM 2 and AutoML in categorising complex mental health narratives. The findings reveal strengths and weaknesses in each model's approach to interpreting human experiences. While AI models can align closely with human judgement in some instances, discrepancies in handling intricate categories like "Trauma and Stress" emphasise the ongoing need for model refinement. Improving AI sensitivity to the nuances of human stories will enhance their effectiveness in mental health assessments and interventions. This study not only advances our understanding of AI's potential in the mental health field but also directs future research and development efforts, stressing the importance of integrating technical innovation with profound psychological insight.

6.2 Unpacking the Complexity: AI and Human Interpretation of Mental Health Narratives

In this section, we delve into the broader implications of our findings, focusing on the complexities of interpreting mental health narratives through both AI and human lenses. This discussion evaluates the performance of different AI models, interprets the consensus between human and AI analyses, and considers the inter-rater reliability across various evaluative frameworks. Through a comparative analysis, we aim to shed light on the specific challenges and successes of each model, offering insights into improving future mental health assessments using AI tools.

6.2.1 Evaluating Model Performance

The training outcomes of the AutoML and PaLM 2 models exhibit contrasting performance characteristics. AutoML, with its high precision and recall metrics, demonstrates a tight adherence to the training dataset's patterns. However, this model's performance did not scale equally when applied to a broader array of narratives, hinting at a tendency toward over fitting. Such a tendency can result in a model that, while accurate in its training context, may lack the flexibility to adapt to new, unseen data.

On the other hand, PaLM 2 showed a remarkable ability to generalise its learning, as evidenced by its substantial accuracy across a varied collection of human evaluations. Despite exhibiting marginally lower precision and recall compared to AutoML, PaLM 2's robustness in validation suggests that it is able to accommodate a wider range of data variability, which is a valuable characteristic when dealing with the diverse expressions found in mental health discussions.

Both models demonstrated challenges when it came to the nuanced language typically found in mental health narratives. This difficulty is particularly pronounced in the domain of trauma, where individuals often use ambiguous or metaphorical language as a means of articulating complex emotional states. The subtleties of such expressions can easily introduce classification uncertainty, as the AI models must navigate through layers of nuanced communication that are often open to multiple interpretations.

These observations underscore an imperative need for advanced linguistic comprehension capabilities within AI models. To effectively parse and understand the complexities of mental health discourse, AI must be capable of discerning not just the literal meanings of words but also the intricate web of context, emotional undertones, and non-literal language that human communication entails. Enhancing AI's ability to process and interpret these complex linguistic elements is essential. This would not only improve the classification accuracy but also allow for a more empathetic and nuanced interaction with the delicate subject matter of mental health, ultimately leading to AI systems that can provide more reliable and meaningful analyses within this critical field.

6.2.2 Interpreting The Human-AI Consensus

The analysis of MHEC aimed at identifying the root causes of mental health issues yielded a notable consensus in 21 instances among all human annotators and both AI models. This unanimous agreement highlights the capacity of both human insight and AI to recognise clear indicators of mental health concerns when certain elements are prominently featured in the discourse. The underlying reason for such a consensus can be attributed to the distinct clarity with which these posts communicate distress, underscored by the prevalent use of specific, impactful keywords and phrases that serve as strong indicators of mental health struggles.

Taking a closer analysis of first exemplar post in Table 6.3 reveals the effectiveness of direct expressions at the very beginning and the presence of significant early life event in facilitating a unanimous understanding among annotators and models. The post discusses major life changes, starting from early life including parental divorce and financial instability, coupled with expressions of deep personal insecurity, academic disengagement, and profound thoughts of self-harm. The explicit articulation of these issues, particularly the mention of contemplating suicide with detailed planning, provides a clear signal that is easily recognisable as long term, deep rooted severe mental health concerns.

The clarity in the post's narrative is critical for a twofold reason: it leverages direct language to convey the individual's emotional turmoil and plans for self-harm, making the mental state of the person unmistakably clear. Moreover, the mention of early life crises, such as parental divorce and its ramifications, furnishes a context that greatly enhances the understanding of potential root causes for the individual's psychological distress.

The consensus on this post, particularly noting the explicit reference to issues starting from early life, confirms the effectiveness of recognising specific keywords and narrative elements that frequently appear in discussions around mental health crises. These elements include straightforward mentions of self-harm, descriptions of intense emotional distress, and references to significant, life-changing early events. The unanimous interpretation by both human annotators and AI models indicates their proficiency in identifying clear expressions of mental health concerns when the discourse includes undeniable indicators of distress and detailed accounts of individual struggles, right from the start of the narrative.

The implications of this consensus are profound, suggesting that enhancing AI models' sensitivity to such key indicators could further refine their effectiveness in recognising mental health issues. This consensus not only validates the approach used in combining human and AI analyses but also offers valuable insights into the characteristics of discourse that facilitate unanimous identification of mental health concerns, paving the way for more nuanced research and targeted interventions in the field.

Second instance in Table 6.3 is also an unanimous agreement among annotators and AI models regarding a post that vividly describes an individual's struggle with gambling addiction and its catastrophic financial and psychological consequences offers a poignant illustration of trauma and chronic stress's role in exacerbating mental health issues. This consensus reflects the effectiveness of both human understanding and artificial intelligence in identifying clear manifestations of trauma and stress, especially when aligned with predefined guidelines on traumatic life events and ongoing stress factors.

The post in question details a harrowing journey of hopelessness, excessive sleep with a wish to never wake up, and a profound sense of regret stemming from significant financial loss due to gambling addiction. The individual's narrative of losing all cryptocurrency savings in an online casino, coupled with a lack of employment, absence of skills, and the background of coming from a country with low economic opportunities, encapsulates a multifaceted view of trauma and stress.

This case aligns perfectly with the trauma and stress guidelines used for annotation and model training, which include factors like financial issues, the burden of ongoing difficulties, and a profound sense of isolation or loneliness. The detailed account of gambling addiction, a chronic stressors that significantly heightens the risk of mental illness alongside the backdrop of financial instability and social isolation, provides a clear framework for understanding the post's alignment with the risk factors for mental health disorders.

The consensus achieved on this account highlights a critical observation: clear, detailed narratives that align with well-defined risk factors for mental health issues enable a precise and unified understanding across both human and AI-driven analyses. Specifically, the story of financial ruin due to gambling addiction embodies a tangible manifestation of key risk factors like financial strain and the ensuing psychological fallout that are universally recognised as exacerbating mental health vulnerabilities.

This unanimous recognition of the outlined risk factors within the narrative underscores the importance of clarity and specificity in mental health discourse. It demonstrates how explicit articulation of life experiences and challenges, particularly those resonating with established risk factors for mental health issues, facilitates accurate and cohesive interpretations by diverse evaluators. Consequently, this consensus not only validates the robustness of the guidelines used for annotation and model training but also exemplifies the efficacy of comprehensive, nuanced narratives in enhancing our understanding of mental health trajectories. Through such detailed and resonant accounts, we gain invaluable insights into the complex interplay of individual experiences with broader socio-economic factors, guiding more effective and empathetic approaches to mental health support and intervention.

6.2.3 Interpreting Misclassification Patterns:The Devil in the Details

Our quantitative analysis has underscored a shared confusion among AI models and human raters in specific scenarios, which is notable despite high overall accuracy. Despite achieving high accuracy rates, there are specific areas where both PaLM 2 and AutoML consistently stumble, pointing to a fundamental challenge in the machine's interpretive faculties.

This recurring confusion signals the need for an introspective look into the decision-making algorithms of these AI models. Understanding where and why these misclassifications occur can offer profound insights into the cognitive processes of AI and the limits of its current understanding. It prompts questions about the AI's ability to contextualise language, to interpret metaphor, and to discern subtleties that human raters typically navigate with intuitive ease.

This particular focus on misclassification patterns serves as an impetus to delve deeper into the cognitive mechanisms at play within AI models. Exploring the 'why' and 'how' behind the AI's decisions in complex categorisation tasks can unveil the areas where machine learning algorithms need to evolve. It opens up pathways for improving machine learning processes, particularly in their handling of the delicate nuances found in mental health discourse, ultimately leading to a more human-like comprehension by these artificial entities.

Instance	**PaLM 2**	**Auto ML**	**Rater -1**	**Rater -2**	**Rater -3**	**Posts**
1	Early Life	Early Life	Early Life	Early Life	Early Life	*" my life was nice before my parents divorced. while growing up I was very insecure about myself, my family is poor because my father doesnt pay his debts which are in total most than half a million € and now my mother has to take care of them. I used to be a great student but i just dont f care now, I feel like im irrelevant and I constantly think about suicide. If I dont do it now i will do it soon, i have though about shooting at my head with a gun which i think would be the less painful. [Some text omitted for brevity]"*
2	Trauma and Stress	Trauma and Stress	Trauma and Stress	Trauma and Stress	Trauma and Stress	*"Every day to me is being a constant torture. I'm sleeping 14 hours a day, hoping someday i never wake up again.I've lost 130k USD on online casino (started with 12k USD, all my cryptocurrency reserves) in February, i'm always thinking myself i could've just stopped and call done, but that wasn't possible since i only got this far by risking it all. Gambling addictionis hell. I never had a job, i have no skills, i'm from a shithole country, national minimum monthly wage here is equivalent to 300 USD. Thankfully to cryptocurrency and NFT games, i made a good money being at right time and place. I was very lucky. But its all gone.I have 2 choices: either suffer everyday, relapsing to gambling addiction and blaming myself for the rest of my life or cease for all.[Some text omitted for brevity]"*

Table 6.3: Human-AI Consensus

6.2.4 Inter-Rater Reliability (IRR):The Human-AI Consensus

Our study's rigorous examination of inter-rater reliability (IRR) and model-to-rater agreement has unveiled informative insights into the dependability of AI in the nuanced task of classifying mental health narratives. The human raters' Cohen's Kappa scores, which range between 0.60 and 0.78, establish a benchmark of human consensus against which the AI models' performance can be gauged.

The PaLM 2 model demonstrates a notable alignment with the human consensus, suggesting that its mechanisms for interpreting complex narratives are particularly effective. This high level of concordance indicates that PaLM 2 may have a superior ability to parse and understand the nuanced language that is often inherent in discussions of mental health, potentially making it a more reliable tool for such analyses.

On the other hand, the AutoML model exhibits a generally good level of agreement with human raters but falls short in some areas. These gaps point to specific facets of content interpretation where AutoML could benefit from refinement. The content that is rich in nuances, such as subtle expressions of feelings or the use of metaphorical language common in personal recounts of mental health experiences, seems to present a particular challenge for AutoML.

The observed discrepancy in performance between the AutoML and PaLM 2 models underscores the presence of varying interpretative strategies employed by each. This difference is pivotal, as it directs attention to the underlying processing capabilities that each model utilises when dealing with complex narrative material. The variance in their performances suggests that while one model may excel in certain aspects of narrative interpretation, it may lack in others where a different model shows strength.

This discovery accentuates the importance of delving deeper into the cognitive-like processes that AI models undertake when interpreting text. It encourages an exploration into how models like PaLM 2 and AutoML parse language, understand context, and discern meaning, especially in the challenging arena of mental health narratives where the depth of context and subtlety of language play critical roles.

6.2.5 Model-Specific Performance

The PaLM 2 model, with its 80 percent concordance rate, reaffirms its advanced interpretive capabilities. In contrast, AutoML's 68 percent alignment rate underlines the necessity for improvements, particularly in the nuanced interpretation of content, where it lags behind PaLM 2.

This differential performance is particularly evident in misclassification pattern analysis, as depicted in Figure 6.5, highlights 'Early Life' as a particularly challenging category for both models. This trend suggests an inherent ambiguity within the 'Early Life' narratives, possibly stemming from overlapping symptomatology or the multifaceted nature of developmental experiences that AI models struggle to disentangle which are not as straightforwardly classified as those in the 'Drugs and Alcohol' category.

The confusion matrices for PaLM 2 and AutoML reveal a nuanced landscape where both models encounter challenges reflective of those faced by human raters, particularly with 'Early Life' and 'Personality' narratives. PaLM 2's confusion between 'Early Life' experiences and 'Personality' traits indicates a potential limitation in linguistic processing. This points to AI content classification difficulty in disentangling the overlapping thematic content and language that often co-occur in personal narratives.

AutoML's performance, mirroring PaLM 2's strengths, also reveals significant challenges. Its difficulty in distinguishing between 'Early Life' and 'Trauma and Stress' suggests a model potentially overfitted to narrower linguistic features, thus struggling to capture the broader spectrum of human experiences.

The emergent pattern of misclassification across the "Trauma and Stress," "Personality," and "Early Life" labels by PaLM 2 and AutoML models raises critical questions about the current limitations of AI in mental health narrative analysis. Notably, the consistent challenges in correctly categorising "Trauma and Stress" narratives may reflect a broader issue within AI systems. PaLM 2's fewer misclassifications in "Personality" traits compared to AutoML suggests a disparity in model sophistication.

The consistent difficulty with the "Early Life" label across both models points to a classification challenge, hinting at the need for a more granular dissection of training criteria to accurately reflect the complexities of early life experiences. These findings suggest that precision in AI classification is heavily reliant on the clarity and specificity of the language used within the training corpus. The near-perfect performance in classifying the "Drugs and Alcohol" label by both models underscores the importance of explicit terminology in enhancing the reliability of AI categorisation.

Our qualitative analysis offers a window into the interpretive challenges that both AI models and human raters face when analysing mental health narratives. The findings indicate that ambiguity within these narratives is a shared obstacle, affecting both human and automated classifications. For instance, specific language markers intended to indicate 'Personality' traits can overshadow broader contextual signals, resulting in misclassifications of narratives that should be categorised under 'Trauma and Stress'. This pattern of confusion, mirrored in human raters, points to a fundamental difficulty in distinguishing between prominent language cues and the underlying narrative context. The qualitative content analysis, particularly illustrated in Table 6.2, underscores the complexity of categorising narratives that straddle the lines between 'Trauma and Stress' and 'Personality'. These categories often blend at the crossroads of mental health diagnostics, revealing a crux where both AI and human judgement must navigate overlapping thematic elements. A narrative infused with keywords suggestive of postnatal depression may simultaneously reflect traits associated with a personality disorder, highlighting the multiplicity inherent in mental health conditions and the need for a nuanced diagnostic lens. This confluence of symptoms and experiences, often clear through multiple diagnostic lenses, reveals a significant challenge common to both models and human raters. It accentuates the necessity for algorithms capable of nuanced sensitivity to context and symptomatology, beyond the recognition of isolated linguistic markers.

The comparative model interpretation analysis further accentuates the distinct approaches of PaLM 2 and AutoML models. PaLM 2's adeptness at capturing the acute stressors in narratives against AutoML's inclination to prioritise developmental keywords over immediate context reflects a divergence in model understanding as seen in the second instance of Table 6.2, where PaLM 2 accurately classified the narrative under 'Trauma and Stress,' acknowledging both the currency and specificity of stressors, AutoML faltered, assigning the narrative to 'Early Life' due to a probable overemphasis on keywords.

This study has brought to light the pronounced disparities in AI performance, especially in narratives entwined with academic pressures and suicidal ideation context, where temporal and situational context is imperative. These insights compel us to consider the implications for AI deployment in mental health settings. The accuracy with which AI models mirror human classification practices is not solely a technical issue but one of understanding the intricacies of human experiences and the subtleties of psychological symptomatology.

Our findings advocate for the larger annotated datasets, embodying a broad spectrum of human conditions, to refine AI models' interpretive capacities. The evidence that AI models like PaLM 2 excel in certain narrative contexts while others such as AutoML encounter difficulties.

The qualitative discrepancies revealed in our analysis highlight the inherent challenge in mental health classification, navigating the blurred boundaries between disparate mental states. The AI models' tendency to conflate 'Trauma and Stress' with 'Personality' suggests a shared linguistic or contextual footprint that current algorithmic interpretations struggle to separate. Such findings resonate with recent discourses on the limitations of AI in complex decision-making, where context and nuance are paramount.

This chapter delved into the patterns of mental health discussions on Reddit during the pandemic as well as it assessed the performance of AI models alongside human analysis in understanding mental health narratives. The focus was on examining the accuracy of data annotations, the effectiveness of AI in mental health classification, and how AI's findings compared with human judgements. We looked at how these models identified and misclassified mental health conditions, and how well they aligned with human evaluations. Through detailed analysis, the chapter highlighted the challenges AI faces in accurately identifying mental health issues, pointing to the need for ongoing improvements to enhance both the accuracy and the sensitivity of these tools. This evaluation helped underscore the benefits of combining AI with human insights to improve mental health diagnostics, aiming for more effective interventions in public health.

Chapter 7

Synthesising Insights: A Culmination of AI and Human Synergy in Mental Health Analysis

In the concluding chapter of this thesis, we synthesise the key insights garnered from our exploration of mental health discourse on social media, particularly within the context of AI integration. This chapter is structured into three pivotal sections, each addressing distinct yet interconnected aspects of our study. The initial section, revisits the significant findings and discusses the impact of AI on understanding and diagnosing mental health issues, highlighting the successes and challenges encountered during the study and provides a detailed reflection on how these findings address our stated research questions and hypotheses, thereby offering a comprehensive evaluation of our theoretical propositions. The subsequent section critically assess the limitations of our research, addressing the constraints and biases inherent in our methods and data sources, and reflecting on the implications these have for the generalizability and applicability of our findings setting the stage for discussions on the need for future research to overcome these challenges. The final section proposes paths forward, suggesting how future research can build upon our findings to further advance the field of mental health diagnostics in the digital age. It aims to chart a course for future inquiry and innovation, inspired by the insights and limitations outlined earlier.

Together, these sections aim to provide a comprehensive overview of our research journey, encapsulate the lessons learned, and guide ongoing and future efforts to harness AI and social media in understanding and treating mental health disorders.

7.1 Concluding Reflections

Our comprehensive investigation into the intersection of mental health discourse on social media and AI has yielded significant insights, marking a pivotal advancement in the understanding of mental health diagnostics.At its core, the study has developed a large-scale dataset that provides an unparalleled in-depth view of mental health discussions on social media platforms, particularly focusing on pandemic era. This dataset serves not only as a testament to the complexity and diversity of mental health narratives but also as a foundational tool for delving into the root causes of mental health issues.

Our study systematically addressed the outlined research questions and hypotheses. Research Question 1 was tackled by developing a large-scale dataset capturing mental health sentiments across various pandemic phases, enabling a profound textual analysis of mental health narratives. This robust analysis supports Hypothesis 1, which posits that social media platforms are invaluable for understanding the root causes of mental health issues, due to their rich, textual data of human emotions. Research Question 2 was explored through a detailed comparison between advanced AI models and human raters, demonstrating that AI can match or even surpass human accuracy in identifying mental health issues, thus validating Hypothesis 2 that suggests a significant correlation between NLP outputs and human judgement. Finally, Research Question 3 was investigated by examining AI model interpretability, showing how AI's understanding compares and contrasts with human cognitive processes in complex scenarios, which informed Hypothesis 3 about the patterns of alignment and divergence in AI and human interpretations.

A pivotal aspect of our research involved a meticulous analysis between advanced AI models and human raters, unveiling the potential of AI to not only complement but, in certain cases, enhance traditional methods of early detection and interpretation of mental health concerns. This comparison revealed that AI models are capable of matching or even surpassing human accuracy in identifying the root causes of mental health issues from narrative texts, suggesting a promising avenue for integrating AI-driven insights into mental health services. This finding underscores the viability of leveraging AI for preliminary screening and support, potentially reducing the burden on healthcare professionals and improving access to mental health care.

By transitioning the focus from the mere detection of mental health disorders to a nuanced analysis of their underlying causes, our study offers a more holistic understanding of the mental health landscape as portrayed online. This shift is crucial for the development of targeted mental health interventions, providing valuable insights into the linguistic cues and contextual factors that signify mental distress. The integration of advanced artificial intelligence models in our analysis has further underscored the potential of AI in complementing traditional diagnostic methods, revealing that AI can match, and in some cases, exceed the accuracy of human raters in identifying mental health root causes.

This synthesis of AI technology and mental health discourse analysis heralds a new era in mental health research, one that emphasises the importance of understanding the intricate dynamics of mental health narratives. Our dataset, enriched by a rigorously annotated subset, emerges as a vital asset for future research, unlocking new avenues for the application of machine learning models and predictive analytics. The capability of AI to offer preliminary screenings and augment mental health services presents an exciting prospect for enhancing care delivery, reducing the strain on healthcare systems, and facilitating better access to mental health support.

7.2 Limitation:Acknowledging the Boundaries

This thesis, encompassing the investigation of mental health discourse on Reddit and the application of artificial intelligence for narrative interpretation, offers substantial insights while acknowledging several limitations that frame the context and implications of our findings.

7.2.1 Dataset Limitations

- **Reliance on Data from Reddit:** Our primary data source, Reddit, has specific demographic characteristics that may not fully represent the global population's diversity. This introduces potential biases, suggesting that while our findings reveal significant trends within Reddit, their applicability to broader mental health dynamics might be limited.

- **Qualitative Depth:** The qualitative depth of user interactions, encompassing discourse quality, tone, or context within online communities, was not fully captured due to the studies' quantitative orientation. This limitation underscores the complexity of mental health discussions, which extends beyond mere engagement patterns.

- **Categorisation Challenges:** Particularly within the "trauma and stress" category, ambiguities in narrative distinction were highlighted. The scope of our annotated subset, limited to 800 posts, restricts a comprehensive analysis of Reddit's multifaceted mental health discourse.

7.2.2 Time frame Limitations

- **Transient Emotional States:** The potential reflection of transient emotional states within our dataset, particularly influenced by the pandemic, may not consistently represent long-term mental health conditions. This highlights the need for caution in interpreting observed trends as indicative of lasting mental health status changes.

7.2.3 Human Evaluator-Related Limitations

- **Subjectivity and Model Selection:** The use of consensus benchmarks and the selection of AI models, namely PaLM2 and AutoML, bring forth the challenges of subjectivity and the rapid evolution of AI technology. The inherent subjectivity in human-derived standards and the current capabilities of AI models underline the variability in classification and interpretation of mental health narratives.

- **Limited Number of Raters:** The evaluation process relied on assessments from a limited number of raters, which restricts the generalizability of our findings. A broader and more diverse pool of human raters would substantiate the comparison metrics between AI models and human judgement.

In light of these limitations, our research underscores the imperative for a multifaceted approach in future studies, incorporating both quantitative and qualitative methodologies and a wider array of data sources. Such comprehensive exploration is essential for a holistic understanding of mental health discussions in the digital age, particularly against the backdrop of global crises like the COVID-19 pandemic. This approach will be instrumental in informing effective interventions and policy development in mental health care, advancing the capabilities of AI in narrative analysis, and ensuring that technological advancements in this domain are both inclusive and reflective of the complexities of human mental health narratives.

7.3 Beyond the Present: Future Directions

Drawing from the detailed investigation of mental health narratives on social media and the application of artificial intelligence in my study, this section articulates the challenges encountered and the insights gained to guide future research endeavour in this evolving field. The journey through analysing vast datasets from Reddit and employing advanced AI and NLP techniques has not only illuminated the complexities of mental health discourse but also opened new avenues for technological and methodological advancements in this critical area of study.

The emergence of transformer-based models and their potential to revolutionise our understanding of context and semantics in mental health discussions on social media represents a direct extension of my work. Future research should delve deeper into these models, leveraging the richly annotated dataset we've developed to uncover deep insights into the changing dynamics of mental health discourse across different pandemic phases. This endeavour could dramatically enhance our predictive models, offering a more nuanced understanding of early mental health issue indicators and facilitating more timely and effective interventions.

Moreover, the limitations encountered in categorising complex mental health narratives, particularly within the broad "trauma and stress" categories, underscore the need for a more refined approach to data annotation. Future studies inspired by this thesis should aim to expand the annotated dataset, incorporating a wider array of narratives to capture the full spectrum of mental health expressions. This expansion is crucial for training AI models to navigate the nuanced interplay of language, sentiment, and thematic context inherent in mental health discourse with greater accuracy and sensitivity.

Additionally, the comparative analysis between AI models and human judgement in interpreting mental health narratives has laid the groundwork for future innovations.

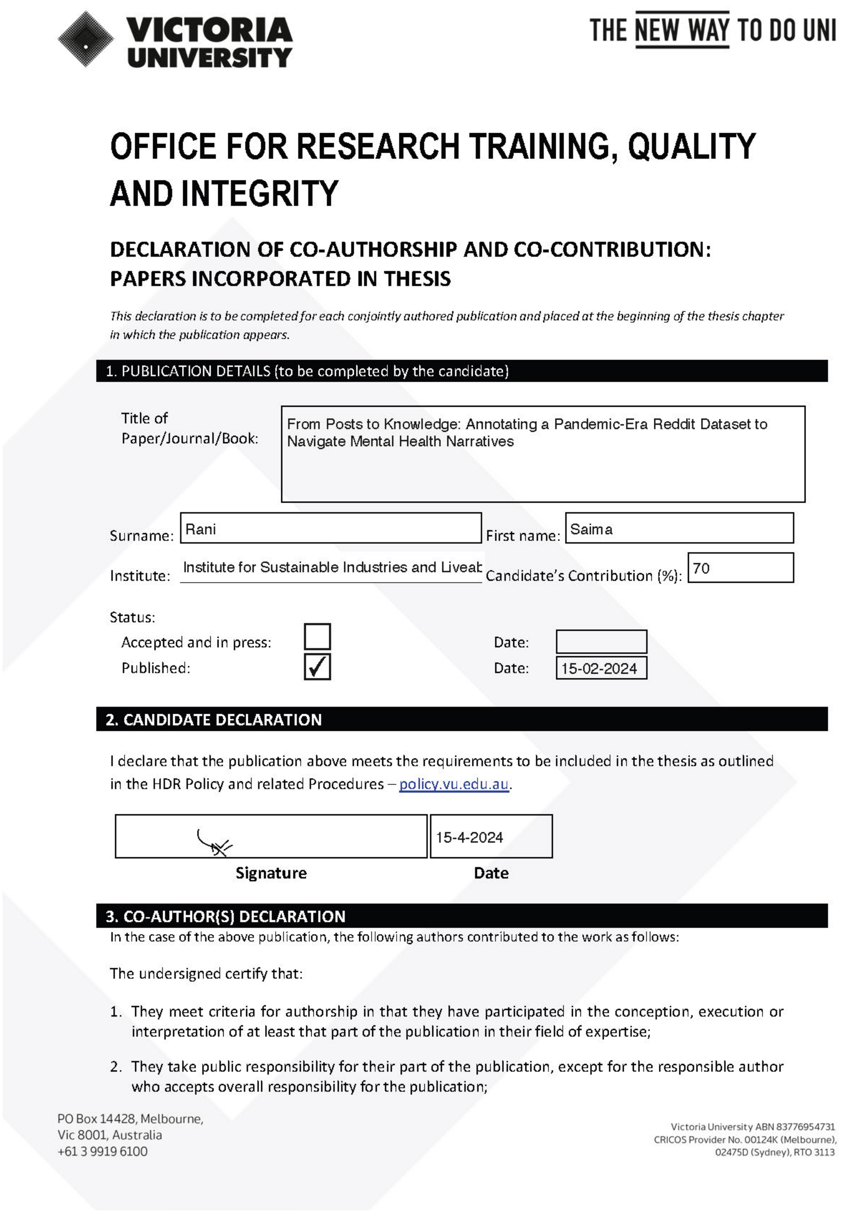
The next phase of research should explore incorporating additional contextual indicators and algorithmic innovations to improve AI's narrative interpretation capabilities. This includes developing AI systems that can more effectively discern narrative tone, sentiment, and underlying themes, thereby bridging the gap between lexical analysis and the sophisticated psychological appraisal akin to human understanding.

In reflecting on the path forward, this thesis not only proposes a road map for technological innovation but also champions a more empathetic and comprehensive approach to mental health care. The anticipated advancements in AI models and annotation methodologies stand to make a significant impact, transforming how we identify, understand, and address mental health issues in the digital age. As we venture into these future directions, the foundation laid by this thesis will serve as a beacon, guiding ongoing efforts to harness the power of AI and social media in fostering a more informed and responsive mental health care ecosystem.

In sum, our exploration into the capabilities of AI in the realm of mental health diagnostics represents a significant stride toward bridging the gap between technological innovations and empathetic, effective mental health care. As we move forward, it is imperative that the research community continues to build upon the groundwork laid by this study, addressing the limitations identified and fostering a collaborative environment for the development of AI tools. Our collective efforts should aim not only to augment the proficiency of mental health professionals but also to empower individuals with deeper insights and enhanced support on their journey toward mental well-being. The potential of our work to influence future research and application in mental health care is immense, and we eagerly anticipate its impact and the further advancements it will inspire.By dissecting the intricacies of these interpretative strategies, we can begin to tailor AI systems more effectively to mirror the sophistication of human cognitive processes. Such advancements would not only enhance the precision of AI classification in mental health discussions but would also contribute to the broader field of AI interpretability and reliability in tasks that require a deep understanding of human language and experience.

Appendix A

Appendix: Co Author Declaration



VICTORIA UNIVERSITY

THE NEW WAY TO DO UNI

OFFICE FOR RESEARCH TRAINING, QUALITY AND INTEGRITY

DECLARATION OF CO-AUTHORSHIP AND CO-CONTRIBUTION: PAPERS INCORPORATED IN THESIS

This declaration is to be completed for each conjointly authored publication and placed at the beginning of the thesis chapter in which the publication appears.

1. PUBLICATION DETAILS (to be completed by the candidate)

Title of Paper/Journal/Book: From Posts to Knowledge: Annotating a Pandemic-Era Reddit Dataset to Navigate Mental Health Narratives

Surname: Rani First name: Saima

Institute: Institute for Sustainable Industries and Liveab Candidate's Contribution (%): 70

Status:

Accepted and in press: □ Date:

Published: ☑ Date: 15-02-2024

2. CANDIDATE DECLARATION

I declare that the publication above meets the requirements to be included in the thesis as outlined in the HDR Policy and related Procedures – policy.vu.edu.au.

	15-4-2024
Signature	**Date**

3. CO-AUTHOR(S) DECLARATION

In the case of the above publication, the following authors contributed to the work as follows:

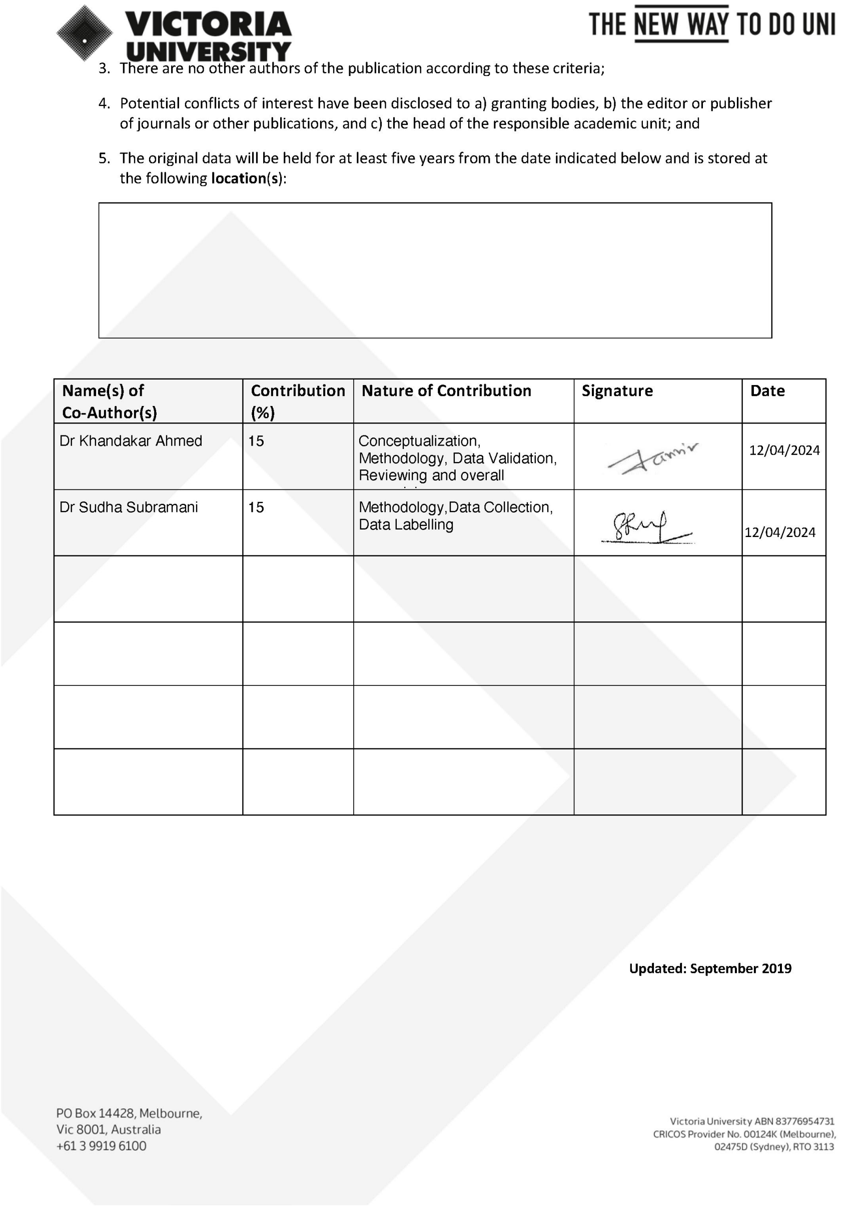
The undersigned certify that:

1. They meet criteria for authorship in that they have participated in the conception, execution or interpretation of at least that part of the publication in their field of expertise;

2. They take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;

PO Box 14428, Melbourne,
Vic 8001, Australia
+61 3 9919 6100

Victoria University ABN 83776954731
CRICOS Provider No. 00124K (Melbourne),
02475D (Sydney), RTO 3113



VICTORIA UNIVERSITY

THE NEW WAY TO DO UNI

3. There are no other authors of the publication according to these criteria;

4. Potential conflicts of interest have been disclosed to a) granting bodies, b) the editor or publisher of journals or other publications, and c) the head of the responsible academic unit; and

5. The original data will be held for at least five years from the date indicated below and is stored at the following **location**(s):

Name(s) of Co-Author(s)	Contribution (%)	Nature of Contribution	Signature	Date
Dr Khandakar Ahmed	15	Conceptualization, Methodology, Data Validation, Reviewing and overall		12/04/2024
Dr Sudha Subramani	15	Methodology,Data Collection, Data Labelling		12/04/2024

Updated: September 2019

Appendix B

Appendix: Annotation Guidelines

Annotation Guidelines for the Analysis of Mental Health Narratives

Introduction

This document outlines the annotation guidelines for a thesis aimed at understanding the root causes of mental health issues through social media narratives. Our hypotheses suggest that social media platforms offer a rich dataset for this purpose and that Natural Language Processing (NLP) can effectively identify these root causes, aligning closely with human judgement. For this purpose, we will take up annotation task according to these guidelines.

Dataset Description

The primary dataset for this project is a collection of 1,000 posts extracted from various mental health-related subreddits on Reddit. These posts were selected based on a preliminary analysis indicating that 99% of the sampled subreddit content directly addresses mental health concerns, making them a valuable source of real, unfiltered narratives for our study.

After the initial selection, a consensus process further refined the dataset to 800 posts, deemed most relevant and informative for our research objectives. This subset, hereafter referred to as the "MH-RAC" will undergo a detailed annotation process to identify specific mental health root causes discussed within each post.

Dataset Goals

The MH-RAC aims to serve as a foundational corpus for:

- Training and testing NLP models in accurately identifying mental health root causes.
- Offering insights into mental health discourse.
- Enhancing our understanding of how individuals articulate and share their mental health experiences online.

This dataset is positioned to be a cornerstone of our research, providing the raw material from which we will extract insights and patterns related to mental health narratives on social media.

Annotation Task Overview

Introduction to Annotation Tasks

As an annotator for this project, you play an essential role in advancing our understanding of mental health narratives as shared on social media. Your task involves meticulously reading through selected posts and classifying each one into one of four mental health root cause categories. This process is vital to our study's success, as it underpins our exploration into the capabilities of NLP technologies in recognizing and interpreting mental health issues, as well as their alignment with human judgment.
Below, you'll find detailed descriptions of each category you will be using in your annotations. These categories have been selected to encompass a wide range of topics within mental health discussions on social media. For each category, we provide a definition and a list of key terms commonly associated with narratives in that category. Your role

involves identifying these key terms within posts and, more importantly, interpreting the context surrounding them to ensure the most accurate categorization.

Categories and Labels

1. Drug and Alcohol

- **Definition**: Focus on posts discussing substance abuse, medications, and its impacts on mental health. This includes narratives about personal struggles with addiction, recovery journeys, and how substance use affects mental well-being.
- **Key Terms**: Look for terms like Alcohol, Drugs, Substance abuse, Addiction, Medication, Dependence, Overdose, Intoxication, Withdrawal, Rehab, Detox, Relapse, Sobriety. It's crucial to consider the context these terms are used in, distinguishing between discussions of personal experience, support-seeking, or advice-giving.

2. Early Life

- **Definition**: This category encompasses posts reflecting on early life experiences and their enduring effects on an individual's mental health. These might include narratives about childhood trauma, parental relationships, bullying and formative experiences that have shaped the poster's mental health.
- **Key Terms**: Key terms include Childhood trauma, Parental neglect, Early loss, childhood abuse, Upbringing, Neglect, Abuse, Trauma, Environment, Family, Poverty, Parent's Divorce, Bullying, School, Early age mention. Context is key here – the focus should be on how these early life experiences are connected to current mental health issues.

3. Personality

- **Definition**: Here, the focus is on posts that delve into personality traits and their correlation with mental health issues. This includes discussions on how certain personality features may predispose individuals to mental health challenges or how they navigate their mental health.
- **Key Terms**: Look for mentions of Perfectionism, Low self-esteem, Self-critical, Negative, Worry, Pessimistic, Impatient, Impulsive, Indecisive, Disrespectful, Aggressive, Arrogant, Emptiness. Consider how the poster relates these traits to their mental health struggles or insights.

4. Trauma and Stress

- **Definition**: This category covers posts about experiences of trauma and stress and their influence on mental health. It includes both acute traumatic events and chronic stressors, and their psychological impacts.
- **Key Terms**: Key terms to identify include PTSD, anxiety, stressors, trauma, coping mechanisms, Depression, Violence, Abuse, Accident, Environment, Disaster, Loss, Grief, Financial Crisis, Work Stress, Breakup. The emphasis should be on understanding the poster's experiences and expressions of trauma and stress.

Guidelines for Accurate Categorization

- **Contextual Reading:** Always read beyond the key terms to grasp the full context of the post. A term like "stress" might be used in various contexts, not always relating directly to trauma.
- **Interpreting Narratives:** Pay attention to how the narrative connects personal experiences with mental health issues. The presence of a key term doesn't automatically place a post in a category unless it's clearly linked to the narrative's main focus on mental health.
- **Handling Ambiguity:** If a post seems to fit into more than one category or if you're unsure about the proper classification, make a note of it for further discussion. Our aim is clarity and accuracy, and your critical thinking is invaluable in achieving this.

Annotation Process

This annotation process is designed to carefully identify the underlying factors contributing to mental health issues as discussed in social media posts. To achieve the consistency and accuracy required for our data annotations, we employ a three-part strategy: semi-automated analysis of keywords, thorough review of entire posts, and manual annotation.

Step-by-Step Annotation Process

1. **Keyword-Based Analysis**
    - **Step 1:** Utilize the search function for semi-automated keyword identification tool provided. This tool scans each post in the dataset using a search function designed to locate predefined keywords that are indicative of mental health root causes.
    - **Step 2:** Refer to the list of keywords of your guideline manual. These keywords have been carefully selected based on their relevance to the root cause categories. For example, in the category of "Personality Factors," look for keywords such as 'perfectionism', 'low self-esteem', and 'pessimistic'.
    - **Step 3:** Each instance where a keyword is found should be flagged for further investigation. This initial flagging is critical as it helps categorize the posts preliminarily, setting the stage for deeper analysis.
2. **Whole Post Analysis**
    - **Step 4:** After the keywords have been identified and flagged, proceed to a detailed examination of the entire content of each post. It is important to read beyond the keywords to grasp the overall context and narrative.
    - **Step 5:** Ensure that the root causes identified through keywords are genuinely reflecting the main focus of the post. This involves verifying that the narrative of the post aligns with the root cause suggested by the keyword(s).
    - **Step 6:** This comprehensive review of the post is crucial not only for confirming the initial findings but also for gaining deeper insights into the complex experiences and expressions of the authors, thereby understanding the content beyond mere keyword occurrence.
3. **Manual Labelling**

- **Step 7:** The final step in the annotation process involves manually assigning a category label to each post, based on the defined root causes and your comprehensive understanding of the post from the previous steps.
- **Step 8:** Maintain the clarity and precision of our dataset by excluding any posts that span multiple categories or do not clearly fit into one category. This selective labelling approach ensures that each post distinctly represents a specific root cause.
- **Step 9:** Record your categorization decision in the annotation tool provided, ensuring each post is labelled according to the root cause categories: Drug and Alcohol, Early Life, Personality, and Trauma and Stress.

Quality Assurance

Resolving Unclear Cases Through Consensus-Based Approach

In the process of annotating social media posts for mental health root causes, situations may arise where the categorization of a post is not immediately clear. These instances require a structured approach to resolve ambiguities and ensure the highest possible accuracy in our dataset. This section outlines the steps you, as annotators, should follow when you encounter unclear cases during the annotation process.
Step-by-Step Process for Resolving Unclear Cases

1. **Initial Review and Annotation**
   - **Step 1**: Each annotator independently reviews and assigns labels to their assigned set of posts, aiming to categorize 50 posts each.
   - **Step 2**: Record your decisions and the rationale behind each categorization in the annotation tool provided. This information will be crucial for discussions in cases of disagreement.
2. **Consensus Meeting Between Annotators**
   - **Step 3**: After the initial annotation phase, schedule a meeting between the annotators assigned to the same posts. The goal of this meeting is to review each post where the categorizations do not match and attempt to reach a consensus.
   - **Step 4**: Discuss each conflicting post, presenting and considering the rationale behind each proposed category. Open dialogue is encouraged to explore the different perspectives and insights that each annotator brings to the table.
3. **Implementing Inter-Annotator Agreement Measures**
   - **Step 5**: Apply the Cohen’s Kappa statistic to evaluate the reliability of the annotation process after the initial consensus meeting. This statistical measure helps to determine the level of agreement between the annotators beyond what would be expected by chance.
   - **Step 6**: Review the Kappa results to identify any posts where agreement is still below acceptable levels (as predetermined in your project guidelines). Posts with Kappa scores indicating insufficient agreement should be earmarked for further review.
4. **Resolution of Disagreements**

- **Step 7**: For posts identified with insufficient agreement from the Kappa analysis, revisit the discussions to see if a further consensus can be reached with the additional statistical context.
- **Step 8**: If both annotators maintain their initial differing opinions after the second discussion, record the post as unresolved. These unresolved cases will be set aside for further review.

5. **Consultation with a Domain Expert**
    - **Step 9**: Compile all unresolved posts and arrange a meeting with the domain expert assigned to oversee the annotation process.
    - **Step 10**: Present the unresolved posts and the rationales from both annotators to the domain expert. Discuss each case in detail to provide the expert with a full understanding of the points of contention.
    - **Step 11**: The domain expert will review the arguments and make the final decision on the categorization of each unresolved post. The expert's decision is final and will be used to update the dataset accordingly.
6. **Finalization of the Annotated Dataset**
    - **Step 12**: Once all posts, including those reviewed by the domain expert, have been finalized, ensure that the dataset of 800 posts reflects consensus-based, expert-validated categorizations.
    - **Step 13**: Perform a final review of the dataset to confirm that all entries are accurately categorised, and that the dataset is ready for further analysis and use in training machine learning models.

Mental Health Root Cause According to Health Direct Australia And Beyond Blue

Health Direct	Beyond Blue
Personality factors - some traits such as **perfectionism or low self-esteem** can increase the risk of depression or anxiety. s**elf-critical and negative and tend to worry a lot.**	**Personality factors** – Some people may be more at risk of depression because of their personality, particularly if they tend to worry a lot, have **low self-esteem**, are perfectionists, are **sensitive to personal criticism**, or are s**elf-critical and negative**.
Drug and alcohol abuse - illicit drug use can trigger a manic episode (bipolar disorder) or an episode of psychosis. Drugs such as cocaine, marijuana and amphetamines can cause paranoia.	**Drug and alcohol use** – **Drug and alcohol use** can both lead to and result from depression. Many people with depression also have drug and alcohol problems. Over 500,000 Australians will experience depression and a substance use disorder at the same time, at some point in their lives.
Early life environment - negative childhood experiences such as **abuse, or neglect** can increase the risk of some mental illnesses.	**Early Life events -** Research suggests living in an **abusive** or uncaring relationship, Bullying.
Trauma and stress - in adulthood, traumatic life events or ongoing stress such as social isolation, domestic violence, relationship breakdown, financial or work problems can increase the risk of mental illness. Traumatic experiences	**Life events -** Research suggests that continuing difficulties – long-term unemployment, **living in an abusive or uncaring relationship**, **long-term isolation**

such as living in a war zone can increase the risk of post-traumatic stress disorder (PTSD). . **continuing difficulties** - Domestic violence - Relationship breakdown - Financial issue - Work Stress - Isolation or loneliness	**or loneliness**, prolonged work stress – are more likely to cause depression than recent life stresses.
Other biological factors - some **medical conditions** or hormonal changes.	**Serious medical illness** – The stress and worry of coping with a serious illness can lead to depression, especially if you're dealing with **long-term management and/or chronic pain**.
Genetic factors - having a **close family member with a mental illness** can increase the risk. However, just because one family member has a mental illness doesn't mean that others will.	**Family history** – Depression can run in families and some people will be at an **increased genetic risk**. However, having a parent or close relative with depression doesn't mean you'll automatically have the same experience. Life circumstances and other personal factors are still likely to have an important influence.

Key words Guide for Selected Mental Health Root causes.

Class	**Potential Keywords**
Personality	Perfectionism, Low self-esteem, Self-critical, Negative, Worry, Pessimistic, Impatient, Impulsive, Indecisive, Disrespectful, Aggressive, Arrogant, Emptiness
Drug and Alcohol	Alcohol, Drugs, Substance abuse, Addiction, Medication, Dependence, Overdose, Intoxication, Withdrawal, Rehab, Detox, Relapse, Sobriety
Early Life	Childhood, Upbringing, Parenting, Neglect, Abuse, Trauma, Environment, Family, Poverty, Parent's Divorce, Bullying, School, Early age mention
Trauma and Stress	Trauma, Stress, PTSD, Anxiety, Depression, Violence, Abuse, Accident, Environment, Disaster, Loss, Grief, Financial Crisis, Work Stress, Breakup

Appendix C

Appendix: Ethics Application

ResearchMaster

Ethics Application

Application ID :	HRE23-005
Application Title :	Mental Health Root Causes Identification from Social Media Posts by Leveraging NLP Technique.
Date of Submission :	12/01/2023
Primary Investigator :	DR KHANDAKAR AHMED (Chief Investigator)
Other Personnel :	MISS SUDHA SUBRAMANI (Associate Investigator)
	Ms Saima Rani (Student)
	PROF MANJULA O'CONNOR (Associate Investigator)

Introduction

Important Information

Form Version: V.16-02. Last Updated: 6.7.2016.

IMPORTANT INFORMATION FOR ALL APPLICANTS:

- Applicants are advised to follow the guidelines provided on the **Human Research Ethics website** prior to submitting this application.
- Ensure all questions are appropriately answered in plain language with correct spelling and grammar.
- All applications must be sighted and approved by all members of the research team and any relevant parties. Applications will not be reviewed without appropriate authorisation.
- To avoid unnecessary delays, please ensure application is submitted in full by the submission deadline for the relevant HREC.

You are reminded that your project may not commence without formal written approval from the appropriate Human Research Ethics Committee.

Contact:

Ethics Secretary
For help and further information regarding ethical conduct, refer to the Human Research Ethics website: *http://research.vu.edu.au/hrec.php* or contact the Secretary for the Human Research Ethics Committee, Office for Research.
Phone: 9919 4781 or 9919 4461
Email: researchethics@vu.edu.au

Quest Service Desk
For technical help, please log a Service Request:

1. Go to the VU Support Hub
2. Scroll down and click "Research" tile
3. Click "Submit a new ticket"
4. Click "Quest Application";
5. Enter "Service" and "Details of Request"
6. Click "Review & Submit" button.

External Resources

- NHMRC: National Statement on Ethical Conduct in Human Research
- NHMRC: Australian Code for the Responsible Conduct of Research

Quest Guide

Quick Tips for Using Quest

Need Help? For help and instructions, we strongly recommend that you download the full Quest Online Ethics Guide (.pdf). Your questions may also be answered in the FAQ page on the Quest Website.

- **Answer All Questions**:
  Most questions are mandatory and must be completed before the application can be submitted. These questions are marked with a red asterisk (*)
- **Access Help and Tips**:
  The ? help icon, found next to questions and at the top of each page, will provide you with detailed advice on ethical content.
- **Remember to Save**:
  Use the 💾 floppy disk icon (and the ✓ green tick in some sections) regularly to avoid losing any answers. Each page will save automatically when you click *Next* ➡ or *Back* ⬅.
- **Print or Save a Copy of Your Application**:
  You can use the ▯ report icon at any stage to generate a printer friendly version of the form. Select HTML to print to screen. To save as a .pdf file to your computer select PDF then save a copy from the pop up screen. *(Don't forget to save a copy before you submit!)*
- **Submit Application**:
  When you have completed your application, click on the *Action* tab in the left-hand column and click *Submit Application.* The system will then convert the form to read-only and send it to the Ethics Secretary for review.
  You will receive an email confirmation at submission. Double check that your application has been submitted by viewing the application status in the *My Applications* page.

Responding to comments (if your application is returned)

There may be stages throughout the application process in which the Ethics Secretary will instruct you to amend your application form. These amendments will be communicated to you via 'Comments' within the eForm.

1. **Generate a List of All Comments**:
   Click the report icon, select *Comments Report* from the Document drop-down field and click *OK*. This list will show all comments created in your application and which page they are applicable to. Click *Cancel* to return to the application form.
2. **Revise your Answers**:
   Open the page which shows a Red Flag icon red flag; these denote an Action Comment which you are required to respond to. Revise the relevant question(s) in your application form as required. Remember to click save!
3. **Respond to Action Comments**:
   AFTER you have revised your answers, you must provide a response to each Action Comment explaining to the Committee how you have addressed their communication. Open the Page Comments window and click New Comment to enter your response into the textbox. Click the ✓ green tick to save your text.
4. **Mark Comments as Responded**:
   Once you have revised your answers AND finished responding to all comments, reopen Page Comments window, use the checkbox to select the *Action Comments* and click *Mark Selected Comments as Responded.* The colour of the flag will change to Yellow Flag icon yellow and the page will become Read Only.
   Important: *DO NOT mark the comments as 'Responded' until you are completely satisfied with your revised answers - you will lose access to edit the page and the comments.*
5. **Submit Revised Application**:
   Once you have addressed all of the Red Flags, open the *Action* tab and click *Submit Revised Application.* The system will then send the form to the Ethics Secretary for review. Remember to save a copy of your application by clicking the Report icon and generating a PDF or printer-friendly version.

[Office Use Only - Administration]

Application ID - *Assign HRE # using "Manage Applications"*

HRE23-005

Clearance Purpose

Research

For Review:

Assigned Ethics Committee

Low Risk Review Panel

Risk Level *(Enter 'High' or 'Low' or 'Neg')*

Low

Students involved in conduct of project? *(Enter 'Yes' or 'No')*

Yes

Date Accepted by Ethics Secretary

24/01/2023

For Finalisation:

Date Approved

29/05/2023

Approved Start Date for Project

29/05/2023

Approved End Date for Project

29/05/2025

Date Rejected

This question is not answered.

Date Withdrawn

This question is not answered.

Application Process Comments

This question is not answered.

[Office Use Only - Risk Assessment]

NEGLIGIBLE RISK INDICATORS
Applicant has responded YES to:

4.5. Does the research only include the collection of anonymous and non-sensitive data (e.g. online survey, observational data) that poses no foreseeable risks or discomfort to participants? *Any foreseeable risk must be no more than inconvenience.*

4.6. Does the research only include the use of non-identifiable and non-sensitive data from an existing database? (e.g., data mining). *Such data should pose no foreseeable risks or discomfort to individuals whose information is contained in the database, or to individuals/organisations responsible for the database.*

HIGH RISK INDICATORS
Applicant has responded YES to:

POSSIBLE HIGH RISK INDICATORS
Applicant has responded YES to:

LOW RISK INDICATOR
If no statements appear under the headings above, the applicant has not responded yes to any negligible or high risk indicators.

SECTION 1 - PROJECT OVERVIEW

General Details

1.1. **Ethics Category***

Human

1.2. **Project Title***

Mental Health Root Causes Identification from Social Media Posts by Leveraging NLP Technique.

1.3. **Project Summary** (Include brief details of aims, methods and significance of the project in plain language. Maximum of 2000 characters)*

Aims
This research aims to propose a framework to identify the mental health root causes from the social media posts using Natural Language Processing (NLP) techniques as these underlying factors lead to serious mental health illness. Assessment of mental health disorder is not possible without identifying underlying cause. Thus, by having support of automated root cause identification, mental health professionals can have support towards treatment.

Methods
1- Proposed research is designed to identify mental health root causes from social media text. Data will be collected from Reddit posts from 2019 to Aug 2022.
2- Limited number of posts will be labelled manually with root causes according to the guidelines of Health direct and Beyond Blue Australia.
3- At next stage extracted content will go through data pre-processing. That includes spellings correction, removing all URLs and any possible identifiable tracks.
4- Next data will be transformed into numerical features that can be processed and inputted into ML model. The model will adopt different classification methods to identify the root cause of mental health issues.
5- Several experiments will be performed to evaluate the performance of model to achieve the desired outcome.
Significance
The need of exploring online tools in observing and offering public health support, is an emerging reality after Covid 19. Not all people reach out to face-to face clinical meetings and prefer sharing their feelings online. Mental health disorder clinical diagnosis is also labour-consuming and time-costing.
Existing studies did not emphasis on root cause identification but provided noteworthy understandings about mental health disorder detection on social media. The implication of our research offers pre-emptive method for mental health root cause identification with formal guidelines from Health Direct Australia. Our study can be beneficial for relevant authorities without being intrusive for health care professionals.

1.4. **Primary College or Institute for Application***

INST OF SUSTAINABLE INDUSTRIES & LIVEABLE CITIES

Timeline and Funding

1.5. **Period for which ethical approval is sought.** *Please refer to the office-use only section at the top of your application to review your application approval dates.*
Project commencement date:*

● Immediately upon receiving ethical approval
○ Other date

1.6. **Date the data collection is expected to be completed:***

15/03/2023

1.7. **How will the research be funded?***

□ External grant
□ VU grant or funding
□ Sponsor
☑ Other
□ Unfunded

Other, provide details:*

This is a student project and part of the Masters of Research program.

1.8. **Is the research a collaborative effort with another organisation?***

○ Yes
● No

SECTION 2 - PROJECT INVESTIGATORS

Investigators

2.1. **Please list all investigators associated with this project.**
The research team is the group of investigators accountable for the conduct of the project. Include details of the Primary Chief Investigator (primary contact for application), as well as all other Chief Investigators and Associate Investigators. *Student details will be requested separately.* Other staff (e.g. technicians) may perform tasks within the project although they are not necessarily investigators. They should be listed as "Other Staff" if appropriate.*

1	ID	E5107953
	Surname	SUBRAMANI
	Given Name	Sudha
	Full Name	MISS SUDHA SUBRAMANI
	College	College of Engineering and Science
	Email	Sudha.Subramani@vu.edu.au
	Role	Associate Investigator
	Phone	0406623832
	Mobile	0406623832
	Qualifications	Qualification: PhD in Information Technology Experience: Dr. Sudha Subramani has 5+ years of experience in working with Natural language processing (NLP) techniques, data analysis, machine learning, deep learning algorithms, social media analysis, and AI for social good applications such as domestic violence, mental health, hay fever prediction, pollen allergy surveillance, antisocial behaviour analysis, COVID19 data analysis, and agricultural spray revolution. Her research work has been published in various Q1 journals and A Rank conferences like IEEE Access, BMC, HIS, and WISE. Her research interest is to apply cutting-edge NLP techniques to a real-world problem which proves beneficial to support groups and the community. Area of Expertise : NLP and AI for Social Good Applications
2	ID	X8881
	Surname	O'CONNOR
	Given Name	MANJULA
	Full Name	PROF MANJULA O'CONNOR
	College	
	Email	manjulao@unimelb.edu.au
	Role	Associate Investigator
	Phone	61396545271
	Mobile	61396545271
	Qualifications	Qualification: MBBS, FRANZCP, DPM, M.MED. Experience and skills related to Project: Dr. Manjula Datta O'Connor is a highly accomplished Consultant Psychiatrist, with over three decades of experience focused on mental health and family violence. Her work addresses the mental health implications of domestic violence in diverse cultural. Dr. O'Connor has actively contributed to research and policy development, all while prioritising mental health aspects. Dr. O'Connor's scholarly work explores the mental health impact on different groups and various situations, such as intimate partner homicide, culturally diverse communities, dowry abuse, and the intersection of mental health with gender and race. Her research emphasises the importance of understanding the unique mental health challenges faced by different populations and the need for tailored support and interventions. She holds professional appointments as Honorary Clinical Associate Professor at the University of Melbourne and Adjunct Professor at UNSW School of Social Sciences. As the Chair of the Family Violence Psychiatry Network at the Royal Australian New Zealand College of Psychiatrists, Dr. O'Connor has been dedicated to enhancing mental health care and support for survivors of family violence. She is the Founding Director of the Australasian Centre for Human Rights and Health (ACHRH), an NGO committed to translating research for communities through public education campaigns addressing sexual and domestic abuse, mental health disorders, and suicide prevention.
3	ID	E5109078
	Surname	AHMED
	Given Name	Khandakar
	Full Name	DR KHANDAKAR AHMED
	College	Information Technology
	Email	Khandakar.Ahmed@vu.edu.au
	Role	Chief Investigator
	Primary CI	Yes
	Phone	0426240101
	Mobile	0426240101
	Qualifications	Qualification: BSc in Computer Science MSc in Networking and e-Business Centred Computing PhD in Electrical and Computer Engineering Experience: Dr. Khandakar Ahmed has an extensive academic and industry background spanning over a decade, with experience working in South Asia, Europe, and Australasia. He has led as a chief investigator in multiple research projects involving a diverse range of fields, including the Internet of Things, smart cities, machine learning, cybersecurity, and biomedical informatics. Presently, he is in charge of the Intelligent Technology Innovation Lab (ITIL) at VU, where his research interests focus on 'AI for Social Good.' The lab's research areas are fascinating, with AI applications in fields like domestic violence, digital abuse, mental health, machine emotion comprehension, and AI for higher education. These research areas hold great potential for solving significant social problems with AI's intervention. Dr. Khandakar has also recently conducted a study (https://doi.org/10.1371/journal.pone.0279743) examining the feasibility and effectiveness of AI-enabled chatbots in detecting depression at an early stage. The outcome is published in a D1 Journal. Area of Expertise Relevant to the Project: Data Mining, Natural Language Processing (NLP), Health Intelligence

Note: Please click the Question Help icon above for instructions on how to search for personnel and use this table.
Once an Investigator record has been added, click on the name in the table above to open the record and edit the information required.

If you are unable to find a personnel record in this system which must be added to your application, please create a "Quest Application" service request under Research section via VU Support Hub.

Student Investigators

2.2. **Will any students be involved in the conduct of this project?***

● Yes
○ No

2.2.a. **If YES, is the project:***

● A STUDENT PROJECT for the degree in which the student is enrolled?
○ A STAFF PROJECT that involves a student(s) undertaking some part of the project?
○ Other

2.2.a.i. **If the research is a STUDENT PROJECT, at what level?***

Masters by Research

* Has this project been approved by the Postgraduate Research Committee? (ie. during confirmation of candidature process)*

● Yes
○ No

2.2.b. **Please list all student investigators involved in this project.**
*Ensure the primary supervisor (not the student), has been marked as the Chief Investigator and primary contact for the application in Q.2.1.**

1	ID	S4665090
	Surname	Rani
	Given Name	Saima
	Full Name	Ms Saima Rani
	College	INST OF SUSTAINABLE INDUSTRIES & LIVEABLE CITIES
	Email	saima.rani@live.vu.edu.au
	Role	Student
	Phone	0432560310
	Mobile	0432560310
	Qualifications	Qualification: Masters In Computer Science Experience: Saima possesses a strong background in the field. With a degree in computer science and 15 years of experience as a lecturer and industry professional, She is currently working as academic sessional in Victoria University. Saima has developed considerable expertise in machine learning and artificial intelligence. This experience is particularly relevant to the project, which primarily involves training a model with labelled data, requiring proficiency in machine learning and coding. Saima is receiving training on application of AI in mental health from Dr Khandakar and Dr Sudha; while learning mental health relevant discipline knowledge under the supervision of Prof Manjula Datta O'Connor.

Note: Please click the Question Help icon above for instructions on how to search for personnel and use this table.
Once a student's record has been added, click on the name in the table above to open the record and edit the information required.

If you are unable to find a personnel record in this system which must be added to your application, please create a "Quest Application" service request under Research section via VU Support Hub.

2.2.c. **What arrangements are in place for the supervision of student(s) when undertaking project activities?***

To ensure constant communication between the supervisor and the student : following arrangements are in place.
1- Scheduled weekly formal Zoom meetings.
2- Informal contact in need via email and mobile.
3- Additional meetings on student's request.
4-Continues feedback to keep student progress on track
5-Assist the student from the selection and planning of the research topic
6-responds in a timely and thorough manner to written work submitted by the student, with constructive suggestions for improvement and continuation

Involvement of Other Individuals/Organisations

2.3. **Will any individuals who are not members of the research team be involved in the conduct of this project?** (e.g., medical personnel involved in procedures, research contractors, teachers) *

○ Yes
◉ No

SECTION 3 - NATURE OF THE PROJECT

Type of Project

3.1.a. **Is the project a pilot study?***

○ Yes
◉ No

3.1.b. **Is the project a part of a larger study?***

○ Yes
◉ No

3.1.c. **Is the project a quality assurance or evaluation project (e.g., related to teaching, health-care provision)?***

○ Yes
◉ No

3.1.d. **Does the research involve a clinical trial (of a substance, device, psychological or physical intervention)?***

○ Yes
◉ No

3.1.e. **Does the research involve the use of therapeutic/intervention techniques or procedures (non-clinical trial)?***

○ Yes
◉ No

Target Population

3.2.a. **Does the research focus on Australian Indigenous (Aboriginal and/or Torres Strait Islander) populations?***

○ Yes
◉ No

3.2.b. **Does the research involve participants under the age of 18 years?***

○ Yes
◉ No

3.2.c. **Does the research involve participants who are highly dependent on medical care?***

○ Yes
◉ No

3.2.d. **Does the research involve participants who have a cognitive impairment, intellectual disability or mental illness? ***

○ Yes
◉ No

3.2.e. **Does the research involve participants in other countries?***

○ Yes
◉ No

3.2.f. **Does the research involve pregnant women (with a research focus on the pregnancy) and/or the foetus (in utero or ex utero) or foetal tissue?***

○ Yes
◉ No

3.2.g. **Does the research involve participants who are likely to be highly vulnerable due to any other reasons?***

○ Yes
◉ No

Intrusiveness of Project

3.3.a. **Does the research use physically intrusive techniques?***

○ Yes
◉ No

3.3.b. **Does the research cause discomfort in participants beyond normal levels of inconvenience?***

○ Yes
◉ No

3.3.c. **Does the research collect potentially sensitive data? (e.g., related to a sensitive topic or vulnerable group; personal health/medical information; sensitive organisational strategies)***

○ Yes
● No

3.3.d. **Does the research involve deception of participants?***

○ Yes
● No

3.3.e. **Does the research involve limited disclosure of information to participants?**

○ Yes
● No

3.3.f. **Does the research involve covert observation of participants?***

○ Yes
● No

3.3.g. **Does the research produce information that, if inadvertently made public, would be harmful to participants?***

○ Yes
● No

3.3.h. **Does the research involve accessing student academic records?***

○ Yes
● No

3.3.i. **Does the research involve human genetic or stem cell research?**

○ Yes
● No

3.3.j. **Does the research involve the use of ionising radiation?***

○ Yes
● No

3.3.k. **Does the research involve the collection of human tissue or fluids?***

○ Yes
● No

3.3.l. **Does the research involve any uploading, downloading or publishing on the internet?***

● Yes
○ No

3.3.m. **Does the research seek disclosure of information relating to illegal activities or is the research likely to lead to disclosure of information relating to illegal activities?***

○ Yes
● No

3.3.n. **Does the research involve procedures that may expose participants to civil, criminal or other legal proceedings?***

○ Yes
● No

3.3.o. **Does the research involve gaining access to medical/health related personal information from records of a Commonwealth or State department/agency or private health service provider?***

○ Yes
● No

3.3.p. **Does the research involve gaining access to personal information (not medical/health) from the records of a Commonwealth or State department/agency or private organisation?***

○ Yes
● No

SECTION 4 - PROJECT DESCRIPTION

General Information

Note: All fields have a maximum of 4000 characters (unless otherwise specified) in plain text only.
If supporting documentation needs to be provided for the following questions (images, graphs etc), please upload as referenced appendices in Section 11 - "Required Attachments" below.

4.1. **Aims of the project.** Provide a concise statement of the aims of the project (maximum 2000 characters in plain language).*

According to WHO, 1 in 8 people in the world suffering from various type of mental disorders like anxiety depression, PTSD and Bipolar etc. There is not one single reason behind these disorders, many of them can be tracked back to certain root causes.
Assessment of any disorder and its treatment is not possible without identifying its underlying root cause and these underlying factors lead to serious mental health illness of an individual. Thus, by understanding the causes, mental health professionals can have a right direction towards treatment for sufferer.
According to health direct Australia following are major root causes which affect individual's mental health.

- Genetic factors
- Drug and alcohol abuse factors
- Other biological factors
- Early life environment factors
- Trauma and stress factors
- Personality factors

This research aims to propose a framework to identify the mental health root causes from the social media posts using Natural Language Processing (NLP) techniques to provide actionable information to assist in initial direction to mental health disorder diagnostic.

4.2. **Briefly describe the relevant background and rationale for the project in plain language.***

Mental health investigation is a continues process of methodical collection and analysis of patient's data to accurate measures indicators to improve mental health[1]. Reliable information of mental health root causes in a timely manner is the key to accurate diagnosis of disorder. Social media data is being considered as a wealth of information for mental health disorders detection and analysis through NLP and ML [2]. Application of NLP and ML represent as supportive tool for decision makers to plan preventive measures to reduce the number of people at risk[3].
A huge presence of research related to detection of mental health issues on social media is evident. However, it is found that the existing studies did not emphasis on root cause identification but provided noteworthy understandings about mental health detection on social media using ML predictors models [4]. To boost detection performance, it is important to understand the cause of the illness as well.
Health care department always requires to identifying all mental health root causes leading to suicides and homicides to be investigated. The process is disliked and considered punitive by mental health professionals[5]. This creates a need of systematic process of data collection and analysis to identify root causes. Our study can be a base for automated clinical diagnoses in terms of identifying the underlying reasons of mental health disorder.
In Covid 19, Health care professional have no focussed trainings of mental health care addressing such pandemics. Despite the grave impact on the mental health, still there is less data have been collected during and after pandemic. Researchers also faced this challenge as deficient amount of data is available for researchers to increase ML model accuracy[6]. Relevant data need to be mined and analysed to support future studies and further professional development of health care professionals [7]. Language used on social media gives a natural lens for mental health research and can be fruitful to fulfil the need of a large, human language dataset for deeper exploration in this area. To achieve the objective, we will also build a large dataset to support aforementioned challenges.
Our study can be beneficial for relevant authorities without being intrusive for health care professionals. This can be used by health services to identify approaches to suicide prevention. It is anticipated that it could be excellent of value to the health care system.

4.3. **Methodology and procedures**
Include specific details relating to any measures, interventions, techniques, and/or equipment used in the research.
Provide step-by-step details of the procedures with particular reference to what participants will be asked to do.
Provide details separately for different phases or conditions of the research or, where appropriate, different participant groups.*

Methodology and Procedure.
In proposed study data will be collected from social media posts without involving any human as subject.
The proposed conceptual framework will consist of following sections:
1- Data Collection
2- Data Labelling
3- Pre-Processing
4- Feature Extraction
5- Model Training
6- Performance Evaluation

Data Collection
Social media networks have provided a new pathway for ground-breaking research to reach enormous textual data of human behavioural dispositions. In our proposed study, we will use Reddit for creating our dataset. Researchers also applied NLP techniques to detect and analyse the data to evaluate the impact of mental health disorders [8, 9] . Reddit is popular as a data source for range of disciplines and there is an upward growth in number of publications using Reddit [10].
On following grounds, we have chosen Reddit as our data collection medium.
1. Anonymity and Content Quality
1.1. One of the unique benefits of Reddit is that all the users participate as pseudonymous users. This help user to write up without being anxious about any social stigma.
1.2. Subreddit moderators also ensure strictly that site rules are being always followed and never allow users to disclose their identities. Therefore, these posts return a high-quality subject related content which is less biased as compared to questionaries and surveys [11].
2. Access to large volume of data related to mental health.
2.1. Reddit is widely used social media site with approximately 430 million users which has range of dedicated single topic forums called 'subreddits', each of which provide focussed information. Reddit extensively harbours range of mental health subreddits as well. Some are well recognised in terms of size, consistency, and members participation, include r/mentalhealth, r/suicidewatch, r/lonely, r/depression, and r/anxiety which provide access to deep insights of user's mental health issues. Domain experts also found them beneficial and prevalent [12, 13].
2.2. Reddit is used to seek or give advice on a variety of mental health issues. One of the most popular subreddit r/IAMA with 22.4M members helping its users in real time with answering their questions including anxiety, trauma, or general mental health under its affiliate r/Iama_Health. This subreddit is particularly being run by professional psychotherapists [14].
2.3. Mental health related subreddits have huge number of active members. These members actively post about their mental health issues without any hesitation. Researchers are already using many of these subreddits [15-17].
2.4. Subreddit moderators delete off topic posts to ensure the authenticity and relevance of the topic.
2.5. Reddit allows long user's posts called 'submissions' to express their sentiments. This makes Reddit ideal for researchers to get deep insights of sensitive topics of human mind.
Studies around mental health have shown focus on Reddit more than other social media platform mostly because of plentiful text allowance for submissions and its pseudonymous user system [18]. These features provide a favourable environment for user to share their honest and forthright feelings which provide lead to researchers about their mental health state. These attributes made Reddit as chosen medium for our study as well.
The first stage of research is to collect reddit posts related to mental health using Pushshift API. Pushshift API is a Reddit data collection and archiving platform and made it available to researchers [19]. Pushshift ingests and collates the Reddit data into public data dumps for researchers to retrieve large amounts of data.
We intend to use Push shift API to collect posts and associated metadata from 5 mental health subreddits. For each of these subreddits, we aim to extract reddit posts relates to following period.

- Pre-Pandemic- 2019
- Mid Pandemic-2020-2021
- Post pandemic- Till Aug 2022

Use this textbox if additional room is required for Question 4.3.

Data Collection :

1. Data source: We will collect data from Pushshift Reddit Dataset (Publicly available) using Pushshift API covering 2019 to 2022. We also already have obtained permission by registering as developer on Reddit developer website. Reddit gave us authentication token to use Reddit API. We will use Pushshift API to collect historical data from Pushshift dataset. We will collect posts from 5 subreddits related to mental health.
2. Data cleansing: As extra layer of privacy, we will clean the data by removing posts which will contain any personally identifiable information or that were from users under the age of 18. This will be done through filtering posts with relevant key words for instance "X year "or "age".
3. Data pre-processing: we will pre-process the data by removing stop words, stemming, and converting all text to lowercase. They also performed tokenisation and feature selection to prepare the data for text classification.
4. Data annotation: We will manually annotate a subset of the data to create a labelled dataset for training and testing their text classification model.
5. Model training and evaluation: We will train a text classification model on the annotated data and evaluated its performance using tools.
6. Data Publication: we will publish the cleaned, pre-processed, annotated data for further research.

Data Labelling
Data will be labelled with accurate root causes for the mental health disorder according to the guidelines from the National mental health resources such as Health Direct Australia and Beyond Blue. We will have mental health expert Dr Manjula in our Supervisory team as well.[23]. There are 6 predominant root causes identified such as personality factors, drug and alcohol abuse, early life environment, trauma and stress factor, biological factors, and genetic factors. Reddit posts will be labelled with these 6 root causes.
Data annotation will be done for 1500 posts initially by 2 annotators, who are familiar with the guidelines and to ensure the data labelling accuracy. If the 2 annotators have given the same label for a given post, then the label will be finalised. Afterwards it will be reviewed by Dr Manjula who is mental helath expert to ensure the data labelling accuracy. If there is any discrepancy between the 2 annotators, the Dr Manjula will have the interactive discussion and decide the label. Through this process, we ensure, all the posts will be labelled accurately and consistently.Afterwards deep learning model will be trained over the dataset at the late stage.

Feature Extraction and Model Training
We will feed text classification model from labelled Reddit posts, to detect if the user is showing the root cause of mental health disorder. The ML model will adopt different classification methods to identify root causes in the given dataset after feature extraction.

Performance Evaluation
At this stage we will evaluate ML models with different methods to have a comparative accuracy level to identify mental health root causes. The results will also reflect the gaps for future research in the field.

Data Collection

4.4. **Indicate all types of data to be collected.***

□ Questionnaire / survey responses*
□ Individual interview responses*
□ Group interview or focus group responses*
□ Participant observations
□ Blood or tissue samples
□ Physiological measures
□ Biomechanical measures
□ Accessed health / medical records or data
□ Accessed student academic records or data
□ Archival data
☑ Other data

Other data, give details:*

Data will be Reddit posts which will collected from PUSHSHUFT Dataset (Public dataset)

4.5. **Does the research only include the collection of anonymous and non-sensitive data (e.g. online survey, observational data) that poses no foreseeable risks or discomfort to participants?** *Any foreseeable risk must be no more than inconvenience.**

● Yes
○ No

4.6. **Does the research only include the use of non-identifiable and non-sensitive data from an existing database? (e.g., data mining).**
*Such data should pose no foreseeable risks or discomfort to individuals whose information is contained in the database, or to individuals/organisations responsible for the database.**

● Yes
○ No

4.7. **Does the research involve photographing or video recording of participants?***

○ Yes
● No

4.8. **Who will be collecting the data?** (give details for all types of data collected and all persons involved)*

Saima Rani. (Student).
DATA TYPE:
1- Reddit posts will be collected from anxiety, depression, suicidewatch,lonely and mentalhealth subreddits.
2- These posts will be collected from PUSHshift Dataset which is available publicly.
3- We will be doing data cleansing as well.

4.9. **Where will the data be collected?** (give details for all types of data collected and all locations)*

R drive of Victoria University.

4.10. **How will the data be analysed?** (give details for all types of data collected)*

After data collection in CSV files which are posts around mental health issues, 1500 to 2000 posts will be labelled manually with 6 predominant root causes such as personality factors, drug and alcohol abuse, early life environment, trauma and stress factor, biological factors, and genetic factors. This will be done by 2 annotators under the guidelines. If the 2 annotators have given the same label for a given post, then the label will be finalised. Afterwards it will be reviewed by Dr Manjula who is mental helath expert to ensure the data labelling accuracy. If there is any discrepancy between the 2 annotators, the Dr Manjula will have the interactive discussion and decide the label. Through this process, we ensure, all the posts will be labelled accurately and consistently.Afterwards deep learning model will be trained over the dataset at the late stage.

4.11. **Who will have access to the data collected?** (give details of all persons who will have access to the data)*

Data will be available to project investigators including student.

4.12. **Will individuals or organisations external to the research team have access to any data collected?***

○ Yes
● No

SECTION 5 - PARTICIPANTS

Participant Group Details

5.1. **Provide details of all distinct participant groups below.**
Please be as precise as possible, if specific details have not been determined you must indicate that they are approximate.

Group 1
Details of specific participant population:*

There are no direct participants in the research. Reddit users are anonymous and on signup user only require to provide an email address which cannot be retrieved or access at any point. Since users will be anonymous so it will be hard to quote specific participant population as well the number of individuals. However, posts anticipated number will be around 1 million.

Number of participants: *

Anticipated number of posts will be around 1 million.

Age range of participants:*

Reddit requires users to be 13+; however, tracing or obtaining user age data is unfeasible

Source of participants:*

Pushshift Dataset of Reddit posts.

*Record details for additional group? (Group 2)**

○ Yes
● No

Participant Selection

5.2. **Provide a rationale for the sample size.***

The anticipated number of posts will be 1 million.

5.3. **Does the project include any specific participant selection and/or exclusion criteria beyond those described above in Question 5.1?***

○ Yes
● No

5.4. **Will there be a formal screening process for participants in the project?** (e.g. medical/mental/health screening)*

○ Yes
● No

5.5. **Does the research involve participants who have specific cultural needs or sensitivities?** (e.g., in relation to the provision of informed consent, language, procedural details)*

○ Yes
● No

5.6.a. **Does the research involve a participant population whose principal language is not English?***

○ Yes
● No

5.6.b. **Will documentation about the research (e.g., Information to Participants form and Consent form, questionnaires) be translated into a language other than English?***

○ Yes
● No

SECTION 6 - RECRUITMENT OF PARTICIPANTS

Recruitment and Informed Consent

6.1. **Will individuals other than members of the research team be involved in the recruitment of participants?***

○ Yes
● No

6.2. **How will potential participants be approached and informed about the research and how will they notify the investigators of their interest in participating?**
Attach copies of the "Information to Participants Involved in Research" form and any flyers or other advertising material to be used in the research in Section 11 - "Required Attachments" below.

We are collecting Reddit posts using publicly available Push shift dataset.

6.3. **Will potential participants be given time to consider and discuss their involvement in the project with others (e.g. family) before being requested to provide consent?***

○ Yes
● No

If NO, provide reasons:*

We are collecting Reddit posts using publicly available Push shift dataset.

6.4. **How will informed consent be obtained from participants?***

□ Participants be required to sign an informed consent form
□ Consent will be implied e.g. by return of completed questionnaire
□ Verbal consent will be obtained and recorded (audio, visual or electronic)
☑ Other

Other, provide details: *

We are collecting Reddit posts using "publicly available Push shift dataset". we will still apply for waiver of consent considering following aspects.
1. Minimal risk
2. No adverse impact
3. Impracticality
4. No alternative methods
5. Confidentiality and data protection

To avoid repetition, please refer to wavier application for detail information.

6.5. **Provide procedural details for obtaining informed consent:***

We are seeking for the waiver of informed consent in light of 6.4 answer.

6.6. **Will you be seeking consent in order to contact participants in the future for related research participation and/or use participants' data for related research purposes?***

○ Yes
● No

Competing Interests

6.7. **Will any dual relationship or conflict of interest exist between any researcher and potential or actual participants?** (e.g., a member of the research team is also a colleague or friend of potential participants)*

○ Yes
● No

6.8. **Does the research involve participants who are in dependent or unequal relationships with any member(s) of the research team or recruiting organisation/agency (e.g. counsellor/client, teacher/student, employer/employee)?***

○ Yes
● No

6.9. **Will you be offering reimbursement or any form of incentive to participants which are not part of the research procedures? Please note:**

- Cash payments by the university will only be provided as a gift card.
- Physical cash payments or bank transfers are not approved methods of payment.
- Gift cards can only be ordered through Procurement using the Gift Card Request Form located at: https://intranet.vu.edu.au/Procurement/Forms.asp.

*

○ Yes
● No

6.10. **Is approval required from an external organisation?** (e.g., for recruitment of participants, data collection, use of premises)*

○ Yes
● No

SECTION 7 - RISKS ASSOCIATED WITH THE RESEARCH

Physical Risks

7.1.a. **Are there any PHYSICAL RISKS beyond the normal experience of everyday life, in either the short or long term, from participation in the research?***

○ Yes
● No

Psychological Risks

7.1.b. **Are there any PSYCHOLOGICAL RISKS beyond the normal experience of everyday life, in either the short or long term, from participation in the research?***

○ Yes
● No

Social Risks

7.1.c. **Are there any SOCIAL RISKS beyond the normal experience of everyday life, in either the short or long term, from participation in the research.** (e.g., possible inadvertent public disclosure of personal details or sensitive information)*

○ Yes
● No

Other Risks

7.2. **Does the research involve any risks to the researchers?***

○ Yes
● No

7.3. **Does the research involve any risks to individuals who are not part of the research, such as a participant's family member(s) or social community (e.g., effects of biographical or autobiographical research)?***

○ Yes
● No

7.4. **Are there any legal issues or legal risks associated with any aspect of the research that require specific consideration (i.e., are significant or out of the ordinary), including those related to:**

- participation in the research,
- the aims and nature of the research,
- research methodology and procedures, and/or
- the outcomes of the research?

*

○ Yes
● No

7.5. **Risk-Benefit Statement:**
Please give your assessment of how the potential benefits to the participants or contributions to the general body of knowledge would outweigh the risks.
*Even if the risk is negligible, the research must bring some benefit to be ethical.**

The potential benefits of this research substantially outweigh the risks, as it seeks to extend our comprehension of mental health underlying causes through Artificial Intelligence.

1. Streamlined research: A well-labelled dataset serves as a foundation for future studies, It will accelerate research by providing a labelled dataset to investigate mental health issues, contributing to knowledge, and further discovery.
2. Comparative analysis: Labelled data will facilitate cross-study comparisons, leading to new hypotheses and insights.
3. Customised interventions: Identifying root causes using machine learning will allow for tailored mental health interventions and support systems applications.
4. Health professionals: It will aid health professionals by providing a foundation for targeted treatments and interventions. By identifying specific root causes through AI, it will lead to improved patient outcomes and more effective healthcare strategies.
5. Policymakers: It will support policy development to address mental health concerns, targeting resources and interventions more effectively.
6. AI advancements: It will foster innovative mental health applications, driving advancements in AI-assisted mental health care and facilitating better access to support for those in need.

SECTION 8 - DATA PROTECTION AND ACCESS

Data Protection

8.1. **Indicate how the data, materials and records will be kept to protect the confidentiality/privacy of the identities of participants and their data, including all hardcopies, electronic files and forms.** *See help for definitions.**

● Data and records will be entirely anonymous
○ Data and records will be coded and non-identifiable
○ Data and records will be coded and re-identifiable
○ Some or all of the retained data and records will include personally identifying information
○ Other

8.2. **Who will be responsible for the security of and access to confidential data and records, including consent forms, collected in the course of the research?***

Research team.

8.3. **Where will data, materials and records be stored during and after completion of the project?** Provide full details of the location for all types of data.
Note: The VU Research Storage provides secure digital storage and long term retention for research project data including graduate research projects.

During the project:*

R Drive

Upon completion:*

R drive

8.4. **Indicate the minimum period for which data will be retained.** See help for definitions.*

○ Indefinitely
○ 5 years post publication
○ 7 years post publication
○ 15 years post publication
○ 25 years after date of birth of participants
● Other

Other, provide details:*

we will retain data until the student completes the degree and relevant publication, including the thesis, is completed.

8.5. **Who will be responsible for re-evaluating the data/materials after the retention period and considering a further retention period for some or all of the data/materials?***

Dr khandakar Ahmed

8.6. **Will you transfer your data or materials to a managed archive or repository during the project, after the project, or after the retention period? Which discipline specific or institutional archives will be considered?**
*Note: Some funding agencies and publishers may require lodgement with an archive or repository. Retain a copy at VU where possible.**

R Drive of Victoria university.
Publish the annotated data and code in Github, as we plan to publish our work to the wider research community.

8.7. **When further retention of data and materials is no longer required, responsible disposal methods should be adopted. Disposal software should also be adopted if digital software, computer hardware, disks or storage media are reused or retired. What methods of appropriate disposal or destruction will be employed?**
Note: Personal, sensitive or confidential information, both digital and hardcopy, will require secure destruction or disposal. For other materials you may need to refer to the Hazardous Materials Policy, Animal Ethics Standard Operating Procedures, or the Ethics and Biosafety site found on the VU Office for Research website. *

We will use software like CBL Data Shredder or Eraser for secure data deletion.

SECTION 9 - DISSEMINATION/PUBLICATION OF RESEARCH RESULTS

Publication Details

9.1. **Indicate how the results of this research will be reported or published.***

☑ Thesis
☑ Journal article(s)
□ Book
□ Research report to collaborating organisations
☑ Conference presentation(s)
□ Recorded performance
□ Other

9.2. **Will any contractual agreement exist between the researchers and a third party that will restrict publication of the research findings?***

○ Yes
● No

9.3. **Are there any other restrictions on publications or reports resulting from this project?***

○ Yes
● No

SECTION 10 - OTHER DETAILS

Comments

10.1. **In your opinion, are there any other ethical issues involved in the research?***

○ Yes
● No

10.2. **Additional information and comments to support this application:**

This question is not answered.

SECTION 11 - DOCUMENTS, ATTACHMENTS AND SUPPLEMENTARY FORMS

Required Attachments

The following documentation must be attached to your application:

- Scanned copy of the Declaration Form for External Investigators (if applicable)

- Copy of the '**Information to Participants Involved in Research**' form *(Please use the templates provided on the Human Research Ethics website)*

- Copy of Consent Forms to be used in the research *(Please use the templates provided on the Human Research Ethics website)*

- Any flyers or other advertising material to be used in the research

11. **Please attach each of the items specifically listed above as well as any other supporting documentation.**
All documentation must be accurately titled and referenced to within the body of your application where appropriate (i.e. "Appendix A - Declaration Form", "Appendix F - Risk Factor Assessment Questionnaire", etc.). Please limit file types to .doc, .docx, .xls, .xlsx, .pdf, or small-medium images (ie, .gif, .jpg).*

1	Document type	Hard copy
	Name	Consent Form
	Description	NA
2	Document type	Hard copy
	Name	Information to Participants Involved in Research
	Description	NA
3	Document type	Soft copy
	Name	Declaration Form for External Investigators
	Reference (Document Title)	VU-EX-DeclarationForm-ExternalInvestigators.docx
	Description	Dr Manjula Declaration Form
4	Document type	Soft copy
	Name	Reference List
	Reference (Document Title)	REFERENCE LIST.docx
	Description	Reference list mentioned in all descriptive answers.
5	Document type	Hard copy
	Name	Advertising Material (flyers etc.)
	Description	NA
6	Document type	Soft copy
	Name	Waiver of the Consent application
	Reference (Document Title)	Wavier of informed consent Apllication.docx
	Description	Waiver of the Consent application

Note: Please click the Question Help icon above for instructions on how to upload documents and use this table.

If you are certain that you do not need to supply a Consent Form or Information to Participants Involved in Research (both of which are mandatory), please tick Hard Copy and type 'N/A' in the Reference field.

SECTION 12 - SUBMISSION DETAILS

Declaration

I / we, the undersigned, declare the following:

- I / we accept responsibility for the conduct of the research project detailed above in accordance with:
  a. the principles outlined in the National Statement on Ethical Conduct in Human Research (2023);
  b. the protocols and procedures as approved by the HREC;
  c. relevant legislation and regulations.
- I / we will ensure that HREC approval is sought using the Changes to the Research Project process outlined on the Human Research Ethics website if:
  a. proposing to implement change to the research project;
  b. changes to the research team are required.
- I / we have read the National Statement on Ethical Conduct in Human Research prior to completing this form.
- I / we certify that all members of the research team involved the research project hold the appropriate qualifications, experience, skills and training necessary to undertake their roles.
- I / we will provide Annual / Final reports to the approving HREC within 12 months of approval or upon completion of the project if earlier than 12 months.
- I / we understand and agree that research documents and/or records and data may be subject to inspection by the VUHREC, Ethics Secretary, or an independent body for audit and monitoring purposes.
- I / we understand that information relating to this research, and about the investigators, will be held by the VU Office for Research. This information will be used for reporting purposes only and managed according to the principles established in the Privacy Act 1988 (Cth) and relevant laws in the States and Territories of Australia.

*

1	ID	E5107953
	Name	MISS SUDHA SUBRAMANI
	Role	Associate Investigator
	Type	Internal
	Declaration signed?	Yes
	Signed on	09/01/2023
2	ID	S4665090
	Name	Ms Saima Rani
	Role	Student
	Type	Student
	Declaration signed?	Yes
	Signed on	11/01/2023
3	ID	E5109078
	Name	DR KHANDAKAR AHMED
	Role	Chief Investigator
	Type	Internal
	Declaration signed?	Yes
	Signed on	09/01/2023
4	ID	X8881
	Name	PROF MANJULA O'CONNOR
	Role	Associate Investigator
	Type	External
	Dec. supplied? (External only)	No
	Supplied on	
	Sighted by	

Note: Please click on your name in the table above to complete your declaration; or click on the name of an External Investigator to acknowledge that their declaration has been supplied.

Declaration Instructions and Information

- A digital signature must be supplied by each and every member of the research team using the declaration table above.
- The 'Needs Signature' icon □ shows which records you are responsible for signing.
- Physical signatures are not required for VU staff and students in applications using form version v.13-07.
- External Investigators do not have access to Quest. The Chief Investigator must supply a completed physical declaration on their behalf by following the steps below:
    1. Send the person a copy of the full application form (including any attachments), as well as the **Declaration Form for External Investigators** document.
    2. Once returned, attach the signed *External Investigator Declaration Form* document in 'Section 11 - Required Attachments'.
    3. Enter into the External Investigator's record in the above declaration table and mark the checkbox to indicate these steps have been completed, include the date you have done so.

       The 'sighted by' field will automatically populate with your name. *(Only the Chief Investigator will have permission to complete this step.)*
- The application cannot be submitted until all members of the research team have logged in and completed this declaration.

Finalise Application

Reminders

- All applications must be sighted and approved by all members of the research team and any relevant parties. Please ensure each member of the research team has completed their declaration in *'Section 12 - Declaration'* above, including any declaration forms supplied on behalf of External Investigators. *Applications will not be reviewed without appropriate authorisation.*
- It is strongly recommended that you save a PDF version of your application before submitting as you will lose access to the electronic record while it undergoes formal review.
- **You are reminded that your project may not commence without formal written approval from the appropriate Human Research Ethics Committee.**

Ready to Submit?

- Once the form is complete and all documents are attached, **click on the *'Action'* tab** above the left-hand form navigation, then **click *'Submit Application'*** to forward the application to the Ethics Secretary to be reviewed and assigned to a Committee meeting.
- You will receive an automatic email notification from Quest when your application has been successfully submitted.
- *Note: Only a Chief Investigator is able to submit an application for ethical approval. The Chief Investigator who is marked as the primary contact for this application is:*

DR KHANDAKAR AHMED

Bibliography

[1] World Health Organization, “Mental Health,” 2024. [Online]. Available: https://www.who.int/health-topics/mental-health#tab=tab_1. [Accessed: 15-Jul-2024].

[2] Z. Steel et al., “The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013,” *International journal of epidemiology*, vol. 43, no. 2, pp. 476-493, 2014.

[3] N. Izadinia, M. Amiri, R. g. Jahromi, and S. Hamidi, “A study of relationship between suicidal ideas, depression, anxiety, resiliency, daily stresses and mental health among Tehran university students,” *Procedia - Social and Behavioral Sciences*, vol. 5, pp. 1615-1619, 2010. doi: https://doi.org/10.1016/j.sbspro.2010.07.335.

[4] D. Bloom et al., *The Global Economic Burden of Noncommunicable Diseases*, 2011.

[5] Department of Health and Aged Care Australia, “Mental health and suicide prevention,” 2022. [Online]. Available: https://www.health.gov.au/health-topics/mental-health-and-suicide-prevention

[6] A. I. i. H. a. welfare, "Death by suicide," 2022, [Online]. Available: https://www.aihw.gov.au/suicide-self-harm-monitoring/data/deaths-by-suicide-in-australia/suicide-deaths-over-time.

[7] WHO, "Mental Disorder," 2022, [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-disorders.

[8] H. direct, "Mental Illness," 2022, [Online]. Available: https://www.healthdirect.gov.au/mental-illness.

[9] N. Boettcher, “Studies of depression and anxiety using reddit as a data source: Scoping review,” *JMIR mental health*, vol. 8, no. 11, p. e29487, 2021.

[10] R. Baheti and S. Kinariwala, “Detection and analysis of stress using machine learning techniques,” *Int. J. Eng. Adv. Technol*, vol. 9, pp. 335-342, 2019.

[11] D. Gillies, D. Chicop, and P. O’Halloran, "Root cause analyses of suicides of mental health clients: Identifying systematic processes and service-level prevention strategies," *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, vol. 36, no. 5, p. 316, 2015.

[12] L. S. Radloff, "The CES-D scale: A self-report depression scale for research in the general population," *Applied psychological measurement*, vol. 1, no. 3, pp. 385-401, 1977.

[13] M. Marcus, M. T. Yasamy, M. v. van Ommeren, D. Chisholm, and S. Saxena, "Depression: A global public health concern," 2012.

[14] P. Y. Collins et al., "Grand challenges in global mental health," *Nature*, vol. 475, no. 7354, pp. 27-30, 2011.

[15] J. Y. Breland, L. M. Quintiliani, K. L. Schneider, C. N. May, and S. Pagoto, “Social media as a tool to increase the impact of public health research,” *American journal of public health*, vol. 107, no. 12, p. 1890, 2017.

[16] R. A. Calvo, D. N. Milne, M. S. Hussain, and H. Christensen, “Natural language processing in mental health applications using non-clinical texts,” *Natural Language Engineering*, vol. 23, no. 5, pp. 649-685, 2017. doi: 10.1017/S1351324916000383.

[17] C. Tannenbaum, J. Lexchin, R. Tamblyn, and S. Romans, “Indicators for measuring mental health: towards better surveillance,” *Healthcare Policy*, vol. 5, no. 2, p. e177, 2009.

[18] E. Lisitsa, K. S. Benjamin, S. K. Chun, J. Skalisky, L. E. Hammond, and A. H. Mezulis, “Loneliness among young adults during COVID-19 pandemic: The mediational roles of social media use and social support seeking,” *Journal of Social and Clinical Psychology*, vol. 39, no. 8, pp. 708-726, 2020.

[19] A. Petrosyan, “Worldwide digital population 2023,” 2023. [Online]. Available: https://www.statista.com/statistics/617136/digital-population-worldwide/.

[20] U. Kursuncu, M. Gaur, U. Lokala, K. Thirunarayan, A. Sheth, and I. B. Arpinar, “Predictive analysis on Twitter: Techniques and applications,” in *Emerging research*

challenges and opportunities in computational social network analysis and mining: Springer, 2019, pp. 67-104.

[21] J. A. Naslund, A. Bondre, J. Torous, and K. A. Aschbrenner, "Social media and mental health: benefits, risks, and opportunities for research and practice," *Journal of technology in behavioral science*, vol. 5, no. 3, pp. 245-257, 2020.

[22] M. L. Birnbaum, A. F. Rizvi, C. U. Correll, J. M. Kane, and J. Confino, "Role of social media and the I nternet in pathways to care for adolescents and young adults with psychotic disorders and non-psychotic mood disorders," *Early intervention in psychiatry*, vol. 11, no. 4, pp. 290-295, 2017.

[23] J. A. Naslund, K. A. Aschbrenner, and S. J. Bartels, "How people with serious mental illness use smartphones, mobile apps, and social media," *Psychiatric rehabilitation journal*, vol. 39, no. 4, p. 364, 2016.

[24] D. Giacco, C. Palumbo, N. Strappelli, F. Catapano, and S. Priebe, "Social contacts and loneliness in people with psychotic and mood disorders," *Comprehensive psychiatry*, vol. 66, pp. 59-66, 2016.

[25] K. Gowen, M. Deschaine, D. Gruttadara, and D. Markey, "Young adults with mental health conditions and social networking websites: seeking tools to build community," *Psychiatric rehabilitation journal*, vol. 35, no. 3, p. 245, 2012.

[26] J. Torous and M. Keshavan, "The role of social media in schizophrenia: evaluating risks, benefits, and potential," *Current opinion in psychiatry*, vol. 29, no. 3, pp. 190-195, 2016.

[27] M. Berger, T. H. Wagner, and L. C. Baker, "Internet use and stigmatized illness," *Social science & medicine*, vol. 61, no. 8, pp. 1821-1827, 2005.

[28] J. C. Badcock et al., "Loneliness in psychotic disorders and its association with cognitive function and symptom profile," *Schizophrenia research*, vol. 169, no. 1-3, pp. 268-273, 2015.

[29] B. J. Miller, A. Stewart, J. Schrimsher, D. Peeples, and P. F. Buckley, "How connected are people with schizophrenia? Cell phone, computer, email, and social media use," *Psychiatry research*, vol. 225, no. 3, pp. 458-463, 2015.

[30] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Business horizons*, vol. 53, no. 1, pp. 59-68, 2010.

[31] V. Rideout and S. Fox, "Digital health practices, social media use, and mental well-being among teens and young adults in the US," 2018.

[32] J. A. Naslund, K. A. Aschbrenner, L. A. Marsch, and S. J. Bartels, "The future of mental health care: peer-to-peer support and social media," *Epidemiology and psychiatric sciences*, vol. 25, no. 2, pp. 113-122, 2016.

[33] H. Haker, C. Lauber, and W. Rössler, "Internet forums: A self-help approach for individuals with schizophrenia?," *Acta Psychiatrica Scandinavica*, vol. 112, no. 6, pp. 474-477, 2005.

[34] P. J. Batterham and A. L. Calear, "Preferences for internet-based mental health interventions in an adult online sample: findings from an online community survey," *JMIR mental health*, vol. 4, no. 2, p. e7722, 2017.

[35] G. Barbier and H. Liu, "Data mining in social media," in *Social network data analytics: Springer*, 2011, pp. 327-352.

[36] M. Brunette et al., "Use of smartphones, computers and social media among people with SMI: opportunity for intervention," *Community mental health journal*, vol. 55, no. 6, pp. 973-978, 2019.

[37] D. E. Losada, F. Crestani, and J. Parapar, "Overview of eRisk at CLEF 2020: Early Risk Prediction on the Internet (Extended Overview)," *CLEF (Working Notes)*, 2020.

[38] M. Park, C. Cha, and M. Cha, "Depressive Moods of Users Portrayed in Twitter," 2012.

[39] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead, "From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses," in *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, 2015, pp. 1-10.

[40] De Choudhury, M.; De, S. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the International AAAI Conference on Web and Social Media* **2014**, *8*(1), 71-80.

[41] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of Depression-Related Posts in Reddit Social Media Forum," *IEEE Access*, vol. 7, pp. 44883-44893, 2019, doi: 10.1109/ACCESS.2019.2909180.

[42] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar, "Discovering shifts to suicidal ideation from mental health content in social media," in *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016, pp. 2098-2110.

[43] K. Hollingshead, M. Ireland, and K. Loveys, "Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality," in *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, 2017.

[44] C. L. Keyes, "Promoting and protecting mental health as flourishing: a complementary strategy for improving national mental health," *American psychologist*, vol. 62, no. 2, p. 95, 2007.

[45] M. A. Moreno et al., "Feeling bad on Facebook: depression disclosures by college students on a social networking site," (in eng), *Depress Anxiety*, vol. 28, no. 6, pp. 447-55, Jun 2011, doi: 10.1002/da.20805.

[46] J. C. Eichstaedt et al., "Facebook language predicts depression in medical records," *Proceedings of the National Academy of Sciences*, vol. 115, no. 44, pp. 11203-11208, 2018.

[47] De Choudhury, M.; Gamon, M.; Counts, S.; Horvitz, E. Predicting depression via social media. In *Proceedings of the International AAAI Conference on Web and Social Media* **2013**, *7*(1), 128-137.

[48] Reece, A.G.; Danforth, C.M. Instagram photos reveal predictive markers of depression. *EPJ Data Science* **2017**, *6*(1), p.15. URL: https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-017-0110-z. DOI: 10.1140/epjds/s13688-017-0110-z.

[49] J. Kim, D. Lee, and E. Park, "Machine learning for mental health in social media: bibliometric study," *Journal of Medical Internet Research*, vol. 23, no. 3, p. e24870, 2021.

[50] H. Lin et al., “Detecting stress based on social interactions in social networks,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 9, pp. 1820-1833, 2017.

[51] M. Thelwall, “TensiStrength: Stress and relaxation magnitude detection for social media texts,” *Information Processing & Management*, vol. 53, no. 1, pp. 106-121, 2017.

[52] B. O’dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, and H. Christensen, “Detecting suicidality on Twitter,” *Internet Interventions*, vol. 2, no. 2, pp. 183-188, 2015.

[53] A. Wongkoblap, M. A. Vadillo, and V. Curcin, “A multilevel predictive model for detecting social network users with depression,” in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, 2018: IEEE, pp. 130-135.

[54] R. Cortés, X. Bonnaire, O. Marin, and P. Sens, “Stream processing of healthcare sensor data: studying user traces to identify challenges from a big data perspective,” *Procedia Computer Science*, vol. 52, pp. 1004-1009, 2015.

[55] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, “Natural language processing applied to mental illness detection: a narrative review,” *npj Digital Medicine*, vol. 5, no. 1, p. 46, 2022.

[56] C. Su, Z. Xu, J. Pathak, and F. Wang, "Deep learning in mental health outcome research: a scoping review," *Translational Psychiatry*, vol. 10, no. 1, pp. 1-26, 2020.

[57] A. Prakash, K. Agarwal, S. Shekhar, T. Mutreja, and P. S. Chakraborty, “An ensemble learning approach for the detection of depression and mental illness over twitter data,” in *2021 8th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2021: IEEE, pp. 565-570.

[58] L. Coello-Guilarte, R. M. Ortega-Mendoza, L. Villaseñor-Pineda, and M. Montes-y-Gómez, “Crosslingual depression detection in twitter using bilingual word alignments,” in *International Conference of the Cross-Language Evaluation Forum for European Languages*, 2019: Springer, pp. 49-61.

[59] S. R. Kamite and V. Kamble, “Detection of depression in social media via twitter using machine learning approach,” in *2020 International Conference on Smart*

Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC), 2020: IEEE, pp. 122-125.

[60] A. Mittal, A. Goyal, and M. Mittal, “Data preprocessing based connecting suicidal and help-seeking behaviours,” in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 2021: IEEE, pp. 1824-1830.

[61] S. Fodeh et al., “Using machine learning algorithms to detect suicide risk factors on twitter,” in *2019 International Conference on Data Mining Workshops (ICDMW)*, 2019: IEEE, pp. 941-948.

[62] A. Shrestha, E. Serra, and F. Spezzano, “Multi-modal social and psycho-linguistic embedding via recurrent neural networks to identify depressed users in online forums,” *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 9, no. 1, pp. 1-11, 2020.

[63] B. Shickel, S. Siegel, M. Heesacker, S. Benton, and P. Rashidi, “Automatic detection and classification of cognitive distortions in mental health text,” in *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, 2020: IEEE, pp. 275-280.

[64] A. Park, M. Conway, and A. T. Chen, “Examining thematic similarity, difference, and membership in three online mental health communities from reddit: A text mining and visualization approach,” *Computers in Human Behavior*, vol. 78, pp. 98-112, 2018.

[65] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532-1543.

[66] S. Boukil, F. El Adnani, L. Cherrat, A. E. El Moutaouakkil, and M. Ezziyyani, "Deep learning algorithm for suicide sentiment prediction," in *International Conference on Advanced Intelligent Systems for Sustainable Development*, Springer, 2018, pp. 261-272.

[67] H. T. Phan, V. C. Tran, N. T. Nguyen, and D. Hwang, "A framework for detecting user’s psychological tendencies on twitter based on tweets sentiment analysis," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, Springer, 2020, pp. 357-372.

[68] S. Ghosh and T. Anwar, "Depression intensity estimation via social media: a deep learning approach," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 6, pp. 1465-1474, 2021.

[69] D. Galiatsatos, G. Konstantopoulou, G. Anastassopoulos, M. Nerantzaki, K. Assimakopoulos, and D. Lymberopoulos, “Classification of the most significant psychological symptoms in mental patients with depression using bayesian network,” in *Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS)*, 2015, pp. 1-8.

[70] M. Gaur et al., “Let Me Tell You About Your Mental Health! Contextualized Classification of Reddit Posts to DSM-5 for Web-based Intervention,” in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018, pp. 753-762.

[71] R. Kavuluru, M. Ramos-Morales, T. Holaday, A. G. Williams, L. Haye, and J. Cerel, “Classification of helpful comments on online suicide watch forums,” in *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2016, pp. 32-40.

[72] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

[73] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2022). PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311.*

[74] Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. (2022). Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416.*

[75] Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al. (2021). Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446.*

[76] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21, 1–67.

[77] Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2021). Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652.*

[78] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903.*

[79] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

[80] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361.*

[81] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300.*

[82] Joshi, M., Choi, E., Weld, D. S., & Zettlemoyer, L. (2017). TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551.*

[83] Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., & Stojnic, R. (2022). Galactica: A Large Language Model for Science. *arXiv preprint arXiv:2211.09085.*

[84] Lewis, P., Ott, M., Du, J., & Stoyanov, V. (2020). Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, 146–157.

[85] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T.-Y. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23.

[86] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3, 1–23.

[87] Hong, Z., Ajith, A., Pauloski, G., Duede, E., Malamud, C., Magoulas, R., Chard, K., & Foster, I. (2022). ScholarBERT: Bigger is Not Always Better. *arXiv preprint arXiv:2205.11342*.

[88] Korngiebel, D. M., & Mooney, S. D. (2021). Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. *NPJ Digital Medicine*, 4, 1–3.

[89] Sezgin, E., Sirrianni, J., Linwood, S. L., et al. (2022). Operationalizing and Implementing Pretrained, Large Artificial Intelligence Linguistic Models in the US Health Care System: Outlook of Generative Pretrained Transformer 3 (GPT-3) as a Service Model. *JMIR Medical Informatics*, 10, e32875.

[90] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, 220–229.

[91] J. Diederich, A. Al-Ajmi, and P. Yellowlees, "Ex-ray: Data mining and mental health," *Applied Soft Computing*, vol. 7, no. 3, pp. 923-928, 2007.

[92] C. Feng, H. Gao, X. B. Ling, J. Ji, and Y. Ma, "Shorten bipolarity checklist for the differentiation of subtypes of bipolar disorder using machine learning," in *Proceedings of the 2018 6th International Conference on Bioinformatics and Computational Biology*, 2018, pp. 162-166.

[93] A. Rios and R. Kavuluru, "Ordinal convolutional neural networks for predicting RDoC positive valence psychiatric symptom severity scores," *Journal of biomedical informatics*, vol. 75, pp. S85-S93, 2017.

[94] Bender, E.M.; Friedman, B. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* **2021**, *6*, 587-604.

URL: https://www.mitpressjournals.org/doi/full/10.1162/tacl_a_00041. DOI: 10.1162/tacl_a_00041.

[95] A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: a scoping review of methods and applications," *Psychological Medicine*, vol. 49, no. 9, pp. 1426-1448, 2019, doi: 10.1017/S0033291719000151.

[96] G. Gkotsis, A. Oellrich, S. Velupillai, *et al.*, “Characterisation of mental health conditions in social media using Informed Deep Learning,” *Sci Rep*, vol. 7, p. 45141, 2017, https://doi.org/10.1038/srep45141.

[97] G. Ansari, M. Garg, and C. Saxena, “Data augmentation for mental health classification on social media,” *arXiv preprint arXiv:2112.10064*, 2021.

[98] S. Sarkar, A. Alhamadani, L. Alkulaib, and C.-T. Lu, “Predicting Depression and Anxiety on Reddit: a Multi-task Learning Approach,” in *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Istanbul, Turkey, 2022, pp. 427-435, doi: 10.1109/ASONAM55673.2022.10068655.

[99] M. Mohd, *et al.*, “Enhanced Bootstrapping Algorithm for Automatic Annotation of Tweets,” *IJCINI*, vol.14, no.2, pp. 35-60, 2020, http://doi.org/10.4018/IJCINI.2020040103.

[100] M. Conway and D. O’Connor, “Social Media, Big Data, and Mental Health: Current Advances and Ethical Implications,” *Current Opinion in Psychology*, 2016.

[101] X. He, *et al.*, “STAT: A web-based semantic text annotation tool to assist building mental health knowledge base,” in *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, IEEE, 2019.

[102] R. Snow, B. O’Connor, D. Jurafsky, and A.Y. Ng, “Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008.

[103] S. Amir, G. Coppersmith, P. Carvalho, M.J. Silva, and B.C. Wallace, “Quantifying Mental Health from Social Media with Neural User Embeddings,” *Machine Learning in Healthcare*, 2019.

[104] M.L. Birnbaum, A.F. Rizvi, J. Confino, C.U. Correll, and J.M. Kane, “Partnering with Insiders: A Review of Peer Models across Community, Health Care, and Research Interventions,” *Psychological Services*, 2017.

[105] A. Cohan, S. Young, N. Goharian, and M. Filannino, “Triaging Content Severity in Online Mental Health Forums,” *Journal of the American Medical Informatics Association*, 2018.

[106] M. Sabou, K. Bontcheva, A. Scharl, and M. Föls, “Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines,” in *Proceedings of the LREC*, 2014.

[107] S.M. Yimam, C. Biemann, L. Majnaric, S. Sabanovic, and A. Holzinger, “An Adaptive Annotation Approach for Biomedical Entity and Relation Recognition,” *Brain Informatics*, 2018.

[108] D. Muzafar, F.Y. Khan, and M. Qayoom, “Machine Learning Algorithms for Depression Detection and Their Comparison,” *arXiv preprint arXiv:2301.03222*, 2023.

[109] M.M. Tadesse, H. Lin, B. Xu, and L. Yang, “Detection of Suicide Ideation in Social Media Forums Using Deep Learning,” *Algorithms*, 2020, 13, 7, https://doi.org/10.3390/a13010007.

[110] K. Sayal, “Annotation: Pathways to care for children with mental health problems,” *Journal of Child Psychology and Psychiatry*, vol. 47, no. 7, pp. 649-659, 2006.

[111] Y. Q. Lim, M. J. Lee, and Y. L. Loo, “Towards A Machine Learning Framework for Suicide Ideation Detection in Twitter,” in *2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, IPOH, Malaysia, 2022, pp. 153-157, doi: 10.1109/AiDAS56890.2022.9918782.

[112] R. Pandey, C. Castillo, and H. Purohit, “Modeling human annotation errors to design bias-aware systems for social stream processing,” in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM ’19)*, Association for Computing Machinery, New York, NY, USA, 2020, pp. 374–377, https://doi.org/10.1145/3341161.3342931.

[113] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan, “Human decisions and machine predictions,” *The Quarterly Journal of Economics*, vol. 2018.

[114] Z. Obermeyer and E.J. Emanuel, “Predicting the Future — Big Data, Machine Learning, and Clinical Medicine,” *The New England Journal of Medicine*, vol. 2016.

[115] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 2017.

[116] A.S. Miner, L. Laranjo, and A.B. Kocaballi, “Chatbots in the fight against the COVID-19 pandemic,” *npj Digital Medicine*, vol. 2020.

[117] J. Torous, M.E. Larsen, C. Depp, T.D. Cosco, I. Barnett, M.K. Nock, and J. Firth, “Smartphones, sensors, and machine learning to advance real-time prediction and interventions for suicide prevention: a review of current progress and next steps,” *Current Psychiatry Reports*, vol. 2016.

[118] N. Rezaii, E. Walker, and P. Wolff, “A machine learning approach to predicting psychosis using semantic density and latent content analysis,” *npj Schizophrenia*, vol. 2019.

[119] J. Torous, G. Andersson, A. Bertagnoli, H. Christensen, P. Cuijpers, J. Firth, *et al.*, “Towards a consensus around standards for smartphone apps and digital mental health,” *World Psychiatry*, vol. 20, no. 1, pp. 97–98, 2020.

[120] M.L. Birnbaum, A.F. Rizvi, K. Faber, J. Addington, C.U. Correll, C. Gerber, and J.M. Kane, “Digital trajectories to care in first-episode psychosis,” *Psychiatric Services*, 2020.

[121] A.N. Vaidyam, H. Wisniewski, J.D. Halamka, M.S. Kashavan, and J.B. Torous, “Chatbots and conversational agents in mental health: A review of the psychiatric landscape,” *The Canadian Journal of Psychiatry*, vol. 64, no. 7, pp. 456-464, 2019.

[122] K.K. Fitzpatrick, A. Darcy, and M. Vierhile, “Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial,” *JMIR Mental Health*, vol. 4, no. 2, e19, 2017.

[123] N.C. Jacobson, B. Summers, and S. Wilhelm, “Digital biomarkers of social anxiety severity: Digital phenotyping using passive smartphone sensors,” *Journal of Medical Internet Research*, vol. 21, no. 4, e11398, 2019.

[124] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, “Causability and explainability of artificial intelligence in medicine,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, e1312, 2019.

[125] A. Abd-Alrazaq, M. Alajlani, A. Alalwan, B. Bewick, P. Gardner, and M. Househ, “Artificial intelligence in the fight against COVID-19: Scoping review,” *Journal of Medical Internet Research*, 2020.

[126] M. Venkatesh, Y.K. Cheung, J. Finkelstein, and D.C. Mohr, “A digital intervention for adults with insomnia: A pilot randomized controlled trial,” *Journal of Medical Internet Research*, vol. 21, no. 6, e12555, 2019.

[127] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, “Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission,” in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

[128] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, “What do we need to build explainable AI systems for the medical domain?” *arXiv preprint arXiv:1712.09923*, 2017.

[129] B. Goodman and S. Flaxman, “European Union regulations on algorithmic decision-making and a ‘right to explanation’,” *AI Magazine*, 2017.

[130] S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, 2017.

[131] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should I trust you?’ Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[132] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” *arXiv preprint arXiv:1312.6034*, 2014.

[133] C. Molnar, *Interpretable Machine Learning*, Lulu.com, 2020.

[134] D. Castelvecchi, "Can we open the black box of AI?" *Nature News*, 2016.

[135] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[136] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, 2019.

[137] K. Usher, J. Durkin, and N. Bhullar, "The COVID-19 pandemic and mental health impacts," International journal of mental health nursing, vol. 29, no. 3, p. 315, 2020.

[138] G. J. Rubin and S. Wessely, "The psychological effects of quarantining a city," Bmj, vol. 368, 2020.

[139] A. Wong, S. Ho, O. Olusanya, M. V. Antonini, and D. Lyness, "The use of social media and online communications in times of pandemic COVID-19," Journal of the Intensive Care Society, vol. 22, no. 3, pp. 255-260, 2021.

[140] A.-S. Uban, B. Chulvi, and P. Rosso, "An emotion and cognitive based analysis of mental health disorders from social media data," Future Generation Computer Systems, vol. 124, pp. 480-494, 2021/11/01/ 2021. [Online]. Available: https://doi.org/10.1016/j.future.2021.05.032

[141] Y.-T. Xiang et al., "Timely mental health care for the 2019 novel coronavirus outbreak is urgently needed," The lancet psychiatry, vol. 7, no. 3, pp. 228-229, 2020.

[142] G. Szmukler, "Homicide inquiries: what sense do they make?," Psychiatric Bulletin, vol. 24, no. 1, pp. 6-10, 2000.

[143] A. Thieme, D. Belgrave, and G. Doherty, "Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems," ACM Transactions on Computer-Human Interaction (TOCHI), vol. 27, no. 5, pp. 1-53, 2020.

[144] Rooshenas L, Paramasivan S, Jepson M, Donovan JL. Intensive Triangulation of Qualitative Research and Quantitative Data to Improve Recruitment to Randomized Trials: The QuinteT Approach. Qualitative Health Research. 2019;29(5):672-679. doi:10.1177/1049732319828693

[145] Inkster B, Digital Mental Health Data Insights Group (DMHDIG). Early warning signs of a mental health tsunami: a coordinated response to gather initial data insights from multiple digital services providers. Frontiers in Digital Health. 2021;2:578902.

[146] Chicco D, Oneto L, Tavazzi E. Eleven quick tips for data cleaning and feature engineering. PLoS Comput Biol. 2022;18(12):e1010718. https://doi.org/10.1371/journal.pcbi.1010718.

[147] Ernala SK, Birnbaum ML, Candan KA, Rizvi AF, Sterling WA, Kane JM, De Choudhury M. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Association for Computing Machinery; 2019. Paper 134, 1-16. https://doi.org/10.1145/3290605.3300364.

[148] Muetunda F, Pais S, Sabry S, Dias G, Pombo N, Cordeiro J. Improving mental disorder predictions using feature-based machine learning techniques. In: 2023 IEEE International Conference on Data Mining Workshops (ICDMW). Shanghai, China; 2023. pp. 1279-1288. doi: 10.1109/ICDMW60847.2023.00165.

[149] Singhal, K. et al. "Large language models encode clinical knowledge," *arXiv preprint* arXiv:2212.13138, **2022**. Available online: https://arxiv.org/abs/2212.13138.

[150] M.L. Birnbaum, S.K. Ernala, A.F. Rizvi, M. De Choudhury, and J.M. Kane, "A Collaborative Approach to Identifying Social Media Markers of Schizophrenia by Employing Machine Learning and Clinical Appraisals," in *J. Med. Internet Res.*, vol. 19, 2017, e289. [Google Scholar] [CrossRef]

[151] A. Benton, M. Mitchell, and D. Hovy, "Multitask Learning for Mental Health Conditions with Limited Social Media Data," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, Spain, 3–7 April 2017, Volume 1. Available online: http://www.aclweb.org/anthology/E17-1015 (accessed on 20 March 2022). [Google Scholar]

[152] Yang, K.; Ji, S.; Zhang, T.; Xie, Q.; Kuang, Z.; Ananiadou, S. "Towards interpretable mental health analysis with large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, **2023**.

[153] "Human Research at VU," Victoria University, 2023. [Online]. Available: https://www.vu.edu.au/researchers/research-lifecycle/conducting-research/human-research-ethics/human-research-at-vu.

[154] V. U. Research, "Research ethics review and using data from social media," 2018. [Online]. Available: https://www.vu.edu.au/sites/default/files/human-research-ethics-data-from-social-media_1.pdf.

[155] T. S. Portal Statistics and Studies, "T. S. Portal Statistics and Studies 2022," 2022, [Online]. Available: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[156] S. Stöckli and D. Hofer, "Susceptibility to social influence predicts behavior on Facebook," *PloS one*, vol. 15, no. 3, p. e0229337, 2020.

[157] M. B. KENNETH OLMSTEAD, "The challenges of using Facebook for research," 2015, [Online]. Available: https://www.pewresearch.org/fact-tank/2015/03/26/the-challenges-of-using-facebook-for-research/.

[158] A. H. P. R. C. STEFAN WOJCIK, "Sizing Up Twitter Users," 2019, [Online]. Available: https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/.

[159] N. Proferes, N. Jones, S. Gilbert, C. Fiesler, and M. Zimmer, "Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics," *Social Media+ Society*, vol. 7, no. 2, p. 20563051211019004, 2021.

[160] E. Yeskuatov, S.-L. Chua, and L. K. Foo, "Leveraging Reddit for Suicidal Ideation Detection: A Review of Machine Learning and Natural Language Processing Techniques," *International Journal of Environmental Research and Public Health*, vol. 19, no. 16, p. 10347, 2022. [Online]. Available: https://www.mdpi.com/1660-4601/19/16/10347.

[161] I. Pirina and Ç. Çöltekin, "Identifying depression on reddit: The effect of training data," in *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, 2018, pp. 9-12.

[162] E. Consulting, "Mental Health AMA," 2021, [Online]. Available: https://www.reddit.com/r/IAmA/comments/oqqb8z/mental_health_ama/.

[163] J. Kim, J. Lee, E. Park, and J. Han, "A deep learning model for detecting mental illness from user content on social media," *Scientific reports*, vol. 10, no. 1, pp. 1-6, 2020.

[164] R. Thorstad and P. Wolff, "Predicting future mental illness from social media: A big-data approach," *Behavior Research Methods*, vol. 51, no. 4, pp. 1586-1600, 2019, doi: 10.3758/s13428-019-01235-z.

[165] E. Chandrasekharan et al., "The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1-25, 2018.

[166] R.M. del Rio-Chanona, A. Hermida-Carrillo, M. Sepahpour-Fard, L. Sun, R. Topinkova, and L. Nedelkoska, "Mental health concerns precede quits: shifts in the work discourse during the Covid-19 pandemic and great resignation," in *EPJ Data Science*, vol. 12, no. 1, 2023, p. 49.

[167] E. Bailey, A. Boland, I. Bell, J. Nicholas, L. La Sala, and J. Robinson, "The Mental Health and Social Media Use of Young Australians during the COVID-19 Pandemic," in *Int. J. Environ. Res. Public Health*, vol. 19, 2022, p. 1077. [Google Scholar] [CrossRef] [PubMed]

[168] D. Valdez, M. ten Thij, K. Bathina, L. Rutter, and J. Bollen, "Social Media Insights Into US Mental Health During the COVID-19 Pandemic: Longitudinal Analysis of Twitter Data," in *J. Med. Internet Res.*, vol. 22, 2020, e21418. [Google Scholar] [CrossRef] [PubMed]

[169] Y. Lee, Y.J. Jeon, S. Kang, J.I. Shin, Y.-C. Jung, and S.J. Jung, "Social media use and mental health during the COVID-19 pandemic in young adults: A meta-analysis of 14 cross-sectional studies," in *BMC Public Health*, vol. 22, 2022, p. 995. [Google Scholar] [CrossRef] [PubMed]

[170] S. Lomborg and A. Bechmann, "Using APIs for data collection on social media," *The Information Society*, vol. 30, no. 4, pp. 256-265, 2014.

[171] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, "The Pushshift Reddit Dataset," in *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 14, 2019, pp. 830–839. [Google Scholar] [CrossRef]

[172] A.K. Poudel and T. Weninger, “Navigating the Post-API Dilemma: Search Engine Results Pages Present a Biased View of Social Media Data,” arXiv 2024, arXiv:2401.15479. [Google Scholar]

[173] K. Nikhileswar, D. Vishal, L. Sphoorthi, and S. Fathimabi, “Suicide Ideation Detection in Social Media Forums,” in *Proceedings of the 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)*, Trichy, India, 7–9 October 2021, pp. 1741–1747. [Google Scholar] [CrossRef]

[174] "Mental health and suicide prevention," *Health.Gov.Au*, 2022. [Online]. Available: https://www.health.gov.au/health-topics/mental-health-and-suicide-prevention/what-were-doing-about-mental-health

[175] M. Garg, C. Saxena, V. Krishnan, R. Joshi, S. Saha, V. Mago, and B.J. Dorr, “CAMS: An annotated corpus for causal analysis of mental health issues in social media posts,” arXiv 2022. [Google Scholar] [CrossRef]

[176] T. Beauvais, “Hybrid Representative Sampling of Social Media,” in *Bull. Sociol. Methodol. Methodol. Sociol.*, vol. 160, 2023, pp. 57–70. [Google Scholar] [CrossRef]

[177] Y. Zhou, J. Zhan, and J. Luo, “Predicting Multiple Risky Behaviors via Multimedia Content,” in *Proceedings of the International Conference on Social Informatics*, Oxford, UK, 13–15 September 2017; Springer International: Cham, Switzerland, 2017. [Google Scholar]

[178] X. Huang, X. Li, T. Liu, D. Chiu, T. Zhu, and L. Zhang, “Topic Model for Identifying Suicidal Ideation in Chinese Microblog,” in *Proceedings of the Pacific Asia Conference on Language, Information and Computation*, Shanghai, China, 30 October–1 November 2015, pp. 553–562. Available online: http://www.aclweb.org/anthology/Y15-1064 (accessed on 11 November 2023). [Google Scholar]

[179] C.M. Homan, “Toward Macro-Insights for Suicide Prevention: Analyzing Fine-Grained Distress at Scale,” in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology*, Baltimore, MD, USA, 27 June 2014, p. 107. [Google Scholar]

[180] J. L. Skeem, J. D. Miller, E. Mulvey, J. Tiemann, and J. Monahan, "Using a five-factor lens to explore the relation between personality traits and violence in

psychiatric patients," *Journal of Consulting and Clinical Psychology*, vol. 73, no. 3, p. 454, 2005.

[181] R. F. Krueger, "Personality traits in late adolescence predict mental disorders in early adulthood: A perspective-epidemiological study," *Journal of personality*, vol. 67, no. 1, pp. 39-65, 1999.

[182] D. Preoţiuc-Pietro et al., "The role of personality, age, and gender in tweeting about mental illness," in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pp. 21-30, 2015.

[183] R. E. Drake and M. F. Brunette, "Complications of severe mental illness related to alcohol and drug use disorders," *Recent Developments in Alcoholism: The Consequences of Alcoholism Medical Neuropsychiatric Economic Cross-Cultural*, pp. 285-299, 1998.

[184] J. C. Skogen, B. Sivertsen, A. J. Lundervold, K. M. Stormark, R. Jakobsen, and M. Hysing, "Alcohol and drug use among adolescents: and the co-occurrence of mental health problems. Ung@ hordaland, a population-based study," *BMJ Open*, vol. 4, no. 9, p. e005357, 2014.

[185] C. Lilley, R. Ball, and H. Vernon, "The experiences of 11-16 year olds on social networking sites," National Society for the Prevention of Cruelty to Children (NSPCC), United Kingdom, 2014.

[186] J. D. Swanson and P. M. Wadhwa, "Developmental origins of child mental health disorders," *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, vol. 49, no. 10, p. 1009, 2008.

[187] Teruel, M.; Cardellino, C.; Cardellino, F.; Alemany, L. A.; Villata, S. (2018). Increasing Argument Annotation Reproducibility by Using Inter-Annotator Agreement to Improve Guidelines. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

[188] Seibold, C.; Jaus, A.; Fink, M. A.; Kim, M.; Reiß, S.; Herrmann, K.; et al. (2023). Accurate Fine-Grained Segmentation of Human Anatomy in Radiographs via Volumetric Pseudo-Labeling. arXiv preprint arXiv:2306.03934.

[189] J. Yang *et al.*, “Harnessing the power of LLMS in practice: A survey on ChatGPT and beyond,” arXiv preprint arXiv:2304.13712, 2023.

[190] Truong, A.; Walters, A.; Goodsitt, J.; Hines, K.; Bruss, C.B.; and Farivar, R. “Towards automated machine learning: Evaluation and comparison of AutoML approaches and tools,” in *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 2019, IEEE, pp. 1471-1479.

[191] Thornton, C.; Hutter, F.; Hoos, H.H.; and Leyton-Brown, K. “Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms,” presented at the *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago, Illinois, USA, 2013. [Online]. Available: https://doi.org/10.1145/2487575.2487629.

[192] Zhu, J.; Yalamanchi, N.; Jin, R.; Kenne, D.; Phan, N. Investigating COVID-19’s Impact on Mental Health: Trend and Thematic Analysis of Reddit Users’ Discourse. *J. Med. Internet Res.* **2023**, *25*, e46867. URL: https://www.jmir.org/2023/1/e46867. DOI: 10.2196/46867.

[193] Alambo, A.; Padhee, S.; Banerjee, T.; Thirunarayan, K. (2021). COVID-19 and Mental Health/Substance Use Disorders on Reddit: A Longitudinal Study. In: Del Bimbo, A., et al. *Pattern Recognition. ICPR International Workshops and Challenges*. Lecture Notes in Computer Science, vol 12662. Springer, Cham. DOI: 10.1007/978-3-030-68790-8_2.