Towards Explainable Artificial Intelligence for Mental Healthcare via Social Media

Yusif Ibrahimov

Doctor of Philosophy

University of York

Department of Computer Science

September 2025

Abstract

Mental health challenges present significant global, social, and economic concerns. Mental health conditions may lead to serious consequences, including self-harm, and suicide. In today's interconnected society, social media platforms provide valuable insight into individuals' thoughts and emotions. This dissertation explores the use of platforms for the mental disorder assessment. Although existing AI-based methods can identify mental disorders, they often overlook explainability, which limits their practical adoption.

Initially, we propose AttentionDep, a domain-aware attention model that drives explainable depression severity estimation by fusing contextual and domain knowledge. Our experiments demonstrate that AttentionDep outperforms state-of-the-art baselines by over 5% in graded F_1 score across datasets, while providing interpretable insights into its predictions. This work advances the development of trustworthy AI systems for mental health assessment from Online Social Media (OSM). Further, we develop RedCoM, a new Reddit dataset for analysing contributors to mental health challenges; and MDCNet, an innovative multi-class learning framework. Through extensive evaluations on existing benchmark datasets, MDCNet outperforms state-of-the-art methods, achieving significant improvements across all evaluation metrics. These results highlight MDCNet's ability to enhance the contextual understanding of contributors to mental disorders and stress, allowing for more effective screening methods. Lastly, we develop a theoretical framework for Explainable Artificial Intelligence (XAI) in Mental Health (MH) applications. We introduce the MentalXAI model for explainable mental health assessment via textual social media data. This contribution allows us to investigate the factors that influencing the model's decision. The evaluation of MentalXAI model has been conducted based on the common mental health conditions. The model consistently outperforms state-of-the-art explainability baselines, demonstrating robust performance across all metrics.

To the best of our knowledge, this dissertation proposes the first comprehensive explainability framework for mental health that covers disorder severity, contributors, and the key features that influence the model's decisions which play a vital role in developing responsible data-driven systems for assessing mental disorders.

A.M. Celal Şengör
for illuminating my early steps in the pursuit of science

Acknowledgments

This dissertation is dedicated to Prof. Dr. A.M. Celal Şengör, who inspired me to become a scientist and enlightened me through my early steps, especially in moments when I lost motivation. I am still at the very beginning of my scientific journey, and I know that my most significant contributions are yet to come. Yet I remain confident, knowing that Herr Şengör is there for me.

I would like to express my deepest gratitude to my mom, my dad, and my sister. I have never been an easy son or sibling, but their endless trust, love, and both emotional and financial support have made me stronger and made this journey possible. I do not know how I can fully express my love for them, but I am certain that they are the true architects of this path.

I am sincerely grateful to my supervisors, Dr. Tarique Anwar and Dr. Tommy Yuan, for their continuous guidance throughout my PhD. They have always supported me, even in the most difficult times. I would also like to express my gratitude to Dr. Dimitar Kazakov, my TAP advisor, for following my research closely over the past three years and for motivating me with his valuable advice.

Finally, I would like to thank my friends: Fahmin, Turan, Ayla, Rauf, Tural, and Toğrul for standing by me throughout this long and challenging period. Their friendship and support have been an essential part of this journey, and I feel truly fortunate to have shared these years with them.

I would like to express my gratitude to the Ministry of Science and Education of Azerbaijan for supporting my PhD studies through their scholarship.

This thesis was prepared using the UniPd modern thesis template by Francesco Pio Barone [1].

[1] https://github.com/baronefr/unipd-thesis-modern

Contents

Declaration of Authorship

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

I have presented parts of the work in this thesis previously at the following venues:

- **IEEE Transactions on Artificial Intelligence (Currently under review)** The survey paper *Explainable AI for Mental Disorder Detection on Social Media: A Survey and Outlook* presents the major part of **Chapter 2**.
- **IEEE Transactions on Artificial Intelligence (Currently under review)** The paper *Rethinking AI-Driven Mental Health: Explainability First* presents the major part of **Chapter 5**.
- **W3PHIAI-25 Workshop (AAAI-25)** The paper *DepressionX: Knowledge Infused Residual Attention for Explainable Depression Severity Assessment*, accepted for the 9th International Workshop on Health Intelligence, presents an earlier version of the research in **Chapter 3**.
- **Expert Systems with Applications (Currently under review)** Our paper *AttentionDep: Domain-Aware Attention for Interpretable Depression Severity Assessment* presents the current work in **Chapter 3**.

1 Introduction

1.1 Background

Mental health (MH)is a state of emotional and social well-being that enables people to control their cognitive and behavioural abilities to cope with daily stressors, make decisions, and engage effectively with their social circle. [1]. MH presents a major public health concern, affecting approximately 12.5% of individuals at some point in their lives. Approximately 1 billion people were affected by the societal, individual, economic, and healthcare-related burdens [2] of mental disorders worldwide in 2016 [203]. In the United States alone, 52.9 million adults live with a mental disorder[3]. In 2014, 16% of youth and adults had experienced symptoms of a common mental health problem in England [4]. Moreover, according to the EU's Health at a Glance: Europe report [5], COVID-19 and its associated economic crisis negatively affected citizens' mental well-being [219]. Depression, anxiety, bipolar disorder, eating disorders, and schizophrenia are among the most widespread and serious mental health challenges [285]. In the US, each year, 6.7% of the population is affected by depression, 1.6% by anorexia and bulimia nervosa, and 2.6% by bipolar disorder [154, 292]. Furthermore, the consequences of mental disorders are not limited to psychological dimensions, they are associated with critical outcomes such as suicide, self-harm, cancer, and

[1]https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response

[2]https://www.who.int/news-room/fact-sheets/detail/mental-disorders

[3]https://www.nami.org/about-mental-illness/mental-health-by-the-numbers/

[4]https://commonslibrary.parliament.uk/research-briefings/sn06988/

[5]https://health.ec.europa.eu/document/download/513157b0-2853-4317-baed-fcd8e2f3d66a_en?filename=2020_healthatglance_rep_en.pdf

cardiovascular disease [203, 3]. Mental disorders are a primary risk factor for suicide, which accounts for approximately 800,000 deaths annually [6]. Only in 2023, 6,069 suicide records were registered in England and Wales which is the highest rate seen since 1999 [7].

Moreover, mental illness can be associated with the mass shooting in the US that takes innocent lives each year. [155, 292].

Traditionally, mental health challenges are diagnosed by the healthcare providers, such as therapists, psychologists, and psychiatrists through face-to-face clinical interviews with patients. Additionally, patients may complete standardised disorder-specific questionnaires as part of the assessment [24, 201, 53, 80]. Figure 1.1 illustrates the tree style diagram of the traditional diagnostic methods.

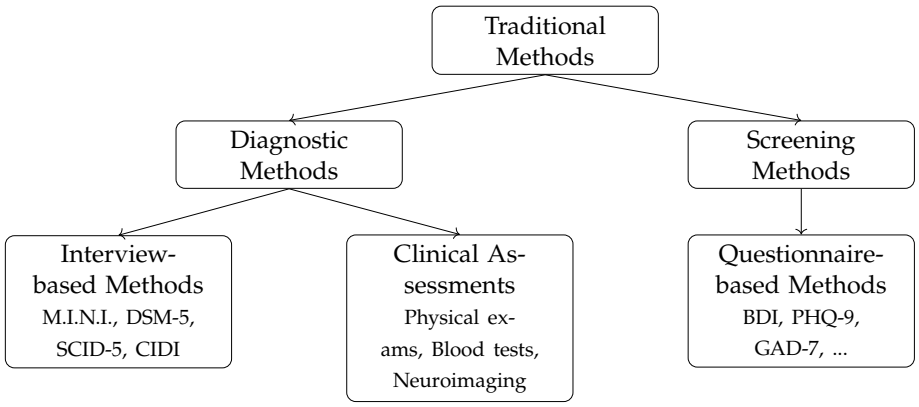


Figure 1.1: Overview of traditional methods for assessment

Although traditional methods are well established, current diagnostic approaches have important limitations [67, 120], including limited session frequency, human factors such as social stigma, fear of prejudice, potential expert subjectivity, and patient response biases that can affect questionnaire accuracy. Given these limitations, there exists a significant demand to explore accessible, fair, and available alternatives for early identification of mental disorders.

Numerous applications of Artificial Intelligence (AI) technologies, such as natural language generation [69], medical imaging [180], and social network analysis [245], have made significant progress. The application of AI technologies has been the subject of extensive study in MH. For MH analytics, AI can be applied in a variety of ways, including early diagnostics [166], AI-driven therapy [89], monitoring and regular screening [162]. A variety of data sources are used

[6] https://iris.who.int/bitstream/handle/10665/131056/9789241564779_eng.pdf?sequence=1

[7] https://tinyurl.com/onsmh

for the mental health analytics and detection, such as Magnetic resonance imaging (MRI) [126], voice [46], Electrocardiogram (ECG) signals [246] and biosignals [11]. Several studies have been conducted on the application of AI algorithms to the detection of mental health disorders such as expert systems [46], bayesian networks [174], logistic regression [62], support vector machines [32], ensemble methods [183] and deep learning algorithms [186]. However, the majority of the previously mentioned data sources are limited in quantity and accessibility.

According to the Datareportal [8], 59.9% of the population uses social media, which equates to more than 4 billion people actively using social media. People convey their emotions, daily routines, problems, ideas, and health states via social media accounts, resulting in an enormous quantity of data that can be used for mental health analytics [83].

There are several research studies focused on mental health disorder detection using social media data [38, 179, 35, 269]. Considering the advantages of deep learning architectures, modern research is mainly conducted with deep learning models [215, 167, 199, 238, 165]. DepressionNet [290] is an effective framework for predicting depression among users by extracting behavioural and textual information using advanced deep learning models. Given the volume of historical user posts, which poses challenges for deep learning models, DepressionNet employs an Extractive-Abstractive Summarisation module to generate input representations. The summarisation process begins with BERT embedding and K-Means clustering to identify the most relevant tweets before content condensation with the Bidirectional and Auto-Regressive Transformers (BART) model, while preserving depression-related content. The model adopts parallel modules for handling user behaviours and posts simultaneously. The behavioural module processes social network features, emotional features, domain-specific features, and topic-related features using Bidirectional Gated Recurrent Unit (GRU) layers, while the textual module employs advanced convolutional, Bidirectional, and attention layers to uncover potential latent indicators of depression within textual posts.

Mihov et al. [156] proposed the MentalNet framework to detect signs of depression by analysing not only textual posts but also associated user interactions on 𝕏 (formerly Twitter). The authors augmented the PsycheNet [195] dataset by including information about replies, mentions, and quote-tweets between users and their social circles. After constructing a heterogeneous social network, they developed a social network graph for further analysis. The associated node features are obtained by sequence-to-sequence Bidirectional Long Short-Term Memory (LSTM) layers before being stabilised by doubly stochastic normalisa-

[8] www.datareportal.com/social-media-users

tion for feeding into graph convolutional networks, resulting in valuable features for successful depression classification. While the previously discussed models focus on depression detection, other researchers have addressed different mental health concerns. For instance, Sawhney et al. developed the SISMO model [215], which detects the level of suicidal ideation in Reddit posts by leveraging Bidirectional LSTM layers and attention mechanisms. Considering the potential volume of social media posts, the authors utilised the pre-trained Longformer model for textual representation, which is then processed through Bi-LSTM and attention models to identify the severity of suicide risk. Notably, the authors adopted ordinal regression to maintain the ordinal nature of the severity labels.

Moreover, Naseem et al. [166] proposed an advanced deep learning model combining Bidirectional LSTM layers with textual graph convolutional networks (TextGCNs) to assess stress and depression severity in OSM posts. Harnessing the representative power of graph neural networks, they adopted the TextGCN model for textual representation. Zhang et al. [284] developed a beneficial deep learning model for estimating depression severities in social media posts by representing textual inputs with Mental RoBERTa [64] (a fine-tuned version of RoBERTa [287] on mental health-based textual data) and SentiLare [122] (a word-level linguistic knowledge-aware transformer model). These advanced contextual and linguistic knowledge-aware embedding features are then processed by an attention-based architecture for contrastive analysis of depression severity patterns. Abuhassan et al. [2] introduced the EDNet model to categorise types of 𝕏 users engaged with Eating Disorder (ED) based content. EDNet learns social media-based input representations by infusing users' historical tweets, biographies, and online behaviours through advanced neural network components. The model employs the Transformer-based BERT model for contextual input representation. Moreover, convolutional temporal attention (CoTA) and bi-convolutional temporal attention (BiCoTA) sublayers are used in ensemble for advanced temporal feature extraction, potentially useful for identifying user categories.

1.2 Research Questions

This dissertation primarily addresses three essential gaps identified through our extensive literature review:

1. **Domain Knowledge in MH Assessment** Advanced deep learning models often follow traditional representation and learning paradigms, even in critical domains such as mental health. Individuals without prior education in mental healthcare are neither capable of nor legally permitted to perform screening, assessment, diagnosis, or intervention, as these activities

require specialised education, practical training, and domain knowledge. However, current deep learning models try to learn correlations between input social media posts and their labels using classical statistical learning mechanisms, such as gradient descent, without incorporating any mental health–related knowledge. Limited studies in the literature consider mental health–related domain information in their analyses. For instance, Zhang et al. [283] developed a context and Patient Health Questionnaire 9 (PHQ-9)-aware model for identifying depressive symptoms using cross-attention modules. They infused descriptive information from the PHQ-9 into the model. Although a few similar studies exist in the literature, they typically employ disorder-specific information that is not generalisable across different mental health conditions. Therefore, such models can focus exclusively on a single disorder, which further raises potential sustainability concerns as they are computationally expensive as training separate model per condition is resource intensive. Moreover, most studies formulate mental disorder assessment as a binary task [225, 209, 290], overlooking severity levels critical for clinical interventions. There are some attempts to address this concern by using regression or classification for severity, however, these models often lack domain-specific depth [166].

Considering these circumstances, the following question emerges:

Can general mental health related domain knowledge enhance the predictive performance of AI based assessment models?What is the role of infusing such domain knowledge in the estimation of depression severity?

2. **MH Contributors** While these models primarily focus on detecting mental health issues, they often overlook the contributing factors of these conditions, which are crucial for precise screening and effective intervention [49, 218, 79]. To address this gap, Mauriello et al. [153] analysed the SAD dataset using Support Vector Machines (SVM), achieving consistent results. Furthermore, Garg et al. [79] applied traditional machine learning, deep learning, and hybrid models to the SAD dataset. Kaur et al. [121] experimented with transformer models on the CAMS dataset, though they did not achieve significant performance improvements. In parallel, Yang et al. [272] fine-tuned LLMs for mental health analysis, aiming to explain the reasoning behind mental health-related causal analysis and daily stressors using both the CAMS and SAD datasets. Despite their potential, LLMs often produce inconsistent results due to the inherent instability and limitations of datasets like CAMS.
   In the light of the aforementioned details, the following research questions can be formulated:
   How can we effectively prepare a dataset that reflects the contributing factors of mental disorders? How can we train a model to achieve unbiased detection of the contributors to mental health challenges by reducing the effects of spurious variables using contextual attention?

3. **Explainability of the proposed Summarisationarchitectures** The rapid emergence of AI across various domains has heightened ethical concerns regarding assurance, transparency, fairness, responsibility, and privacy, particularly in health-related fields such as mental health analysis [117,

114]. Assurance, in this context, refers to establishing justified confidence in healthcare systems for their intended use [114]. The black-box (opaque) nature of many mental disorder detection and screening methods potentially violates assurance and transparency principles. To address these limitations, the AI community has actively pursued explainable and interpretable models, especially for healthcare applications [114]. Explainability is one of the key concepts in artificial intelligence that facilitates understanding of the AI model's decision-making mechanism. Interpretability refers to the ability to easily understand a model's decision by examining its parameters.

Ahmed et al. [4] employed attention mechanisms to explain their Bidirectional LSTM-based mental health analysis model, leveraging attention values to reflect feature importance. Similarly, Zogan et al. [292] utilised a hierarchical attention model to explain their depression detection system, capturing both word-level and sentence-level attentions to elucidate their respective roles in the detection process. Naseem et al. [167] utilised a transformer model for explainable suicide risk identification on social media, by examining attention values from the multi-head attention component of transformer encoder. Yang et al. [272] fine-tuned the *Llama 2* LLM for interpretable mental health screening, focusing on mental disorder detection, severity estimation, and causal classification by generating reasoning behind each assessment.

Based on the provided information, the following research questions can be framed:

How can we formally define and evaluate explainability, particularly within the mental health context? Do baseline explainability models, particularly attention-based models, perform sufficiently well in terms of explainability? How can we tune attention mechanisms to generate context-aware explanations?

1.3 Contributions

The contributions of this dissertation are reported as follows:

- **C1:** We performed extensive bibliographic review to analyse the traditional and automated methods for the mental health assessments. Particularly, we deep dived into the recent AI based trends used for the detection of mental disorder, considering their performance and ethical dilemmas.
- **C2:** We constructed the MH Knowledge Base as a form of knowledge graph (MHKG) that grasps key mental health related entities and the relations among them. MHKG is generated with the help of the LLMs and humans, which further having promising role as a key source of of information that can be injected to the models.

- **C3:** We propose AttentionDep - a domain knowledge-infused attention model for explainable depression severity detection. Building upon prior work, AttentionDep introduces hierarchical attention over unigram and bigram representations, enriched with domain knowledge via a cross-attention mechanism over a curated depression-specific knowledge graph. Empirical results show that AttentionDep outperforms state-of-the-art baselines by over 5% in F_1 score across datasets with varying configurations. This work contributes to the development of trustworthy and interpretable AI systems for mental health analysis via social media.
- **C4:** We prepare RedCoM, a new Reddit dataset for analysing contributors to mental health challenges. Annotation process adopts human-in-the-loop mechanism, with involving LLMs and trained graduate students as annotators, and each labelled post is further validated by an independent annotator (graduate student) who performs a review to confirm the labels. High agreement among annotators is ensured during certain validation phases.
- **C5:** We present MDCNet, an innovative multi-class learning framework that integrates advanced pre-trained transformer-based representations with a causality inspired contextual attention mechanism to distinguish between task relevant and spurious features, enabling the extraction of task-relevant and fine-grained features to identify potential contributors to mental health challenges. Through extensive evaluations on existing benchmark datasets, MDCNet outperforms state-of-the-art methods, achieving significant improvements across all evaluation metrics. These results highlight MDCNet's ability to enhance the contextual understanding of contributors to mental disorders and stress, allowing for more effective screening methods.
- **C6:** We propose MentalXAI, a mechanism forcing attention modules to generate faithful and robust explanations together with the formal definition of XAI for MH domain. We analysed our methodology against the prominent baselines. MentalXAI achieved best trade-off between the detection and explainability measures.

Overall, this dissertation illustrated the XAI as a form of the "trinity" of *disorder severity* (**C2, C3**), *disorder contributor* (**C3, C4**) and *disorder indicators* (**C5,C6**). These contributions can further support mental health practitioners by providing evidence based insights and reasoning from the social media posts of the users, which enable clinicians to prioritise cases that require genuine human evaluation.

1.4 Dissertation Outline

The remainder of this dissertation is organised as follows:

- **Chapter 2.7** introduces the necessary background, covering both traditional and automated methods for assessment of mental health challenges. We start with a review of traditional diagnostic methods, highlighting their main drawbacks and the need for data-driven automated approaches, including an extensive discussion of classical and contemporary machine learning models. Moreover, we survey the available datasets and evaluation metrics. Finally, our review is culminated by emphasising the essential role of explainability in the methodologies discussed.
- **Chapter 3** presents the details of the Mental Health Knowledge Graph (MHKG) and AttentionDep model. It first introduces the proposed hierarchical attention mechanism, then provides the details of the construction of the domain-specific knowledge graph and its infusion into the classification model. We conduct detailed experiments to demonstrate the effectiveness of our module against strong baselines, supported by ablation studies.
- **Chapter 4** reports our methodology to analyse the major contributors to the mental disorders. It first introduces the the role of attention as a measure of feature importance, before presenting the technical details of our RedCoM dataset annotation mechanism. The chapter then presents the MDCNet framework. Moreover, it provides a detailed presentation of the experimental results, including performance evaluation and ablation studies.
- **Chapter 5** focuses particularly on explainability in the mental health domain. It begins with the theoretical background of explainability in the field of mental health. This is followed by the methodology section, where we provide the technical details of the MentalXAI model. The chapter then presents experimental results, including performance evaluation, sensitivity analysis, and the psycholinguistic relevance of the generated explanations.
- **Chapter 6** concludes the dissertation, along with its limitations and potential directions for future work.

2 Literature Review

2.1 Introduction

Mental Health (MH) is a complex phenomenon that significantly influences psychological well-being, affective states, and behavioural patterns. Globally, mental disorders pose a significant public health challenge, affecting approximately 12.5% of individuals at some point in their lifespan. They affected an estimated 1 billion people worldwide in 2016 [203], contributing to a substantial societal, individual, economic and healthcare-related burdens [1]. Depression, anxiety, bipolar disorder, and schizophrenia are among the most prevalent and severe mental health conditions [285]. Like other serious illnesses, mental disorders often require medication, hospitalisation, and emergency care, placing significant strain on healthcare systems.

The implications of mental disorders extend beyond psychological dimensions, serving as causal factors for other critical conditions including self-harm, cancer, and cardiovascular diseases [203, 3]. The World Health Organisation (WHO) identifies mental disorders as a primary contributor to suicide, accounting for approximately 800,000 fatalities annually [2]. These challenges affect individuals across all genders, ethnicities, nationalities, and belief systems, with profound impacts on both mental and physical well-being, frequently manifesting inappropriate behaviours [3]. Delayed or avoided help-seeking for mental health concerns remains prevalent, largely influenced by societal stigma and various contextual factors [243]. The situation is particularly serious in low-income countries, where only 1 in 27 individuals with mental health disorders receives

[1] www.who.int/news-room/fact-sheets/detail/mental-disorders

[2] www.who.int/news-room/fact-sheets/detail/mental-disorders

adequate care [244]. This treatment gap exacerbates critical situations, especially in advanced stages where individuals experience significant functional decline, impaired cognition, self-harm behaviours, and potentially suicidal acts. Stigmatisation, social exclusion, and restricted access to educational and employment opportunities further compound challenges for those with mental health conditions [100]. Notably, mental health problems impose greater economic burdens than many chronic and somatic illnesses, including cancer and diabetes, costing the global economy approximately 2.5 trillion USD in 2010 [248]. These factors contribute to secondary issues including poverty, unemployment, and security concerns, highlighting the extensive societal impact of untreated mental disorders [100]. Additional complexities arise from accessibility challenges during critical scenarios such as pandemics and armed conflicts [2, 83].

Consequently, there exists an urgent need to explore novel paradigms and develop innovative methodologies for early identification and management of mental health challenges. Such initiatives would not only alleviate individual suffering but also support healthcare professionals in addressing mental health crises, ultimately reducing healthcare, economic, societal, ethical, and personal burdens globally. In recent years, the intersection of data science, Artificial Intelligence (AI), and MH has witnessed remarkable advancements [282]. Researchers are making rapid progress in developing novel AI- and data-driven solutions for mental healthcare, including applications such as early diagnosis [166], AI-driven therapy [59], and monitoring through regular screening [48].

A significant portion of our society (approx. 59.9% [3]) actively engages in Online Social Media (OSM) platforms [16]. Particularly among the youth, there is a notable inclination to discuss sensitive topics on online platforms rather than in-person interactions [193]. Individuals share their emotions, daily routines, problems, ideas, and health conditions on OSM, generating vast amounts of personal data that holds potential for mental health analytics [83, 34, 57]. Recent studies propose OSM as a futuristic solution for continuous mental healthcare [211]. Significant advancements are being made in deep learning for the development of models like DepressionNet [290] and EDNet [2] for detecting mental disorders.

However, while deep learning models are valuable, their adoption in healthcare demands explainability [292]. Many of these models operate as black boxes, rendering the reasoning behind their decisions unclear. Given the critical nature of healthcare decisions, reliance on black-box models raises safety concerns, as they cannot guarantee 100% accuracy [114]. Explainable Artificial Intelligence (XAI) models offer a solution by shedding light on the decision-making mecha-

[3] www.datareportal.com/social-media-users

nisms of AI models.

Therefore, investigating deep learning models at a low level is crucial for achieving explainability and developing robust XAI models for detecting mental disorders.

2.1.a Existing surveys and our contributions

A number of surveys have examined the use of AI in Mental Disorder Detection (MDD) using social media data [242, 97, 268, 282, 222, 43]. However, these works are often limited in scope, technical depth, or relevance to explainable MDD on social media platforms. Some, such as Hasib et al. [97] and Yu et al. [268], focus on specific disorders (e.g., depression, suicidality), which limits generalisability across mental health conditions. Others, like Zhang et al. [282], include social media as a major data source but treat its platform-specific characteristics only briefly. Their analysis of deep learning models remains descriptive, lacking critical discussion of technical limitations or compatibility with explainability. Chancellor et al. [43] offer a more focused perspective on predictive methods for mental health using social media but omit detailed analysis of predictive model architecture and underemphasise the role of domain-specific evaluation strategies and explainability. Likewise, Thieme et al. [242] provide a broad review of machine learning in mental health but treat social media only as a subdomain within general Natural Language Processing (NLP), with minimal attention to its distinct structures, challenges, or opportunities.

In contrast, our chapter provides a comprehensive, technically detailed, and explainability-driven review of mental disorder detection using social media data. The novel contributions of our work are as follows:

- *Dedicated analysis of feature extraction techniques specific to social media*: We present a detailed taxonomy of features relevant to MDD, including linguistic cues, user posting behaviours, temporal patterns, and social network-based features (e.g., follower/followee structure, retweet graphs). We also discuss domain-informed features derived from psychological theories and mental health literature.
- *Comprehensive technical review of predictive models*: We examine both classical machine learning and deep learning approaches, offering a comparative analysis of their strengths, limitations, and suitability for the mental health domain. Importantly, we discuss the compatibility of each model type with explainability frameworks, an aspect often overlooked in prior reviews.
- *In-depth exploration of XAI methods*: We critically analyse XAI techniques already applied in MDD (e.g., Local Interpretable Model-agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), attention mecha-

nisms) and introduce promising but underutilised methods such as Graph Neural Network (GNN) explainability tools, Individual Conditional Expectation (ICE) and Partial Dependence Plots (PDP). We outline how these methods can enhance transparency and trustworthiness in mental health applications.

- *Incorporation of domain-specific evaluation metrics*: In addition to general machine learning metrics, we highlight the importance of context-aware performance evaluation by discussing specialised metrics such as Early Risk Detection Error (ERDE), Average Hit Rate (AHR), Average Closeness Rate (ACR), and Depression Category Hit Rate (DCHR). These metrics are critical for evaluating the timeliness and reliability of early MDD systems.
- *Structured synthesis of open challenges and future research directions*: We conclude with a clear outline of the technical, ethical, and methodological challenges in this area, and propose actionable research directions to advance the field of explainable MDD.

The chapter is structured around four guiding research questions:

- **RQ1**: What are the limitations and challenges of traditional methods for diagnosing mental disorders?
- **RQ2**: What are the advantages and disadvantages of data-driven approaches in the detection of mental disorders via social media?
- **RQ3**: What is the current state of explainability in data-driven models for mental disorder detection?
- **RQ4**: What are the future research directions and potential challenges in the field of XAI for mental disorder detection?

2.1.b Our review methodology

This study follows Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to ensure a transparent and structured approach to the selection and analysis of relevant literature. I conducted the entire PRISMA process, including the search, screening, and inclusion stages, using Covidence[4] software to manage and record decisions. I used the Scopus [5] database for searching due to its comprehensive interdisciplinary coverage across computer science and mental healthcare

[4] https://www.covidence.org/

[5] https://www.scopus.com/

Figure 2.1 shows our PRISMA flow diagram, which outlines the steps taken during the study selection process. The literature search was conducted on the Scopus database using the query shown in Table 2.1. We acknowledge that our review methodology is narrow, focusing on studies that explicitly applied AI driven methods for the assessment of mental health conditions. However, this was not the only literature we considered. For each paper included in the review, I also checked its related work and reference list to find other studies that offered useful theoretical, clinical, or methodological insights. These additional papers, even if not captured by the original search query, helped broaden the review and strengthen the understanding of mental health and explainable AI research. Moreover, the main focus is on explainability within the context of mental health intentionally, where understanding model trustworthiness is especially important for reliability and interpretability. Therefore, the review intentionally concentrated on AI-driven systems applied to mental healthcare. However, I also examined well-known XAI methods that are frequently discussed in the wider XAI literature and textbooks but not specifically applied to mental health. Their potential applicability to this domain is discussed in later chapters.

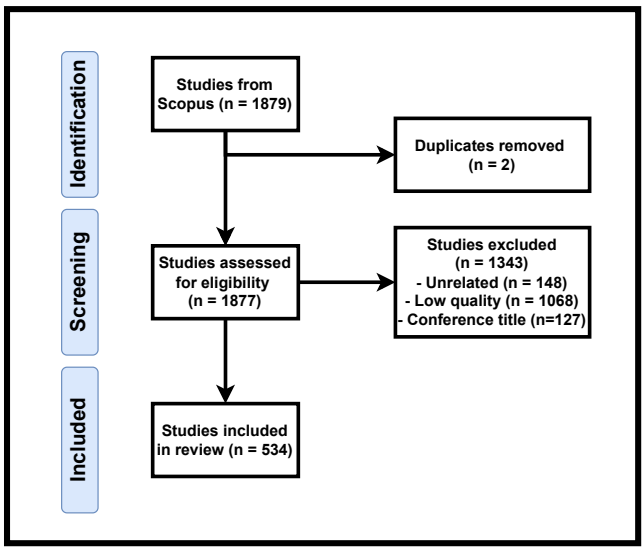


Figure 2.1: PRISMA flow diagram for study selection

The query returned 1,879 papers (before removing two duplicates). We then conducted a multi-stage screening process, assessing titles, abstracts, and full texts where necessary. In total, 148 studies were excluded for being off-topic

Category	Keywords
Mental disorder (**A**)	mental health, mental disorder, depression, anxiety, bipolar disorder, Post-Traumatic Stress Disorder (PTSD), schizophrenia, MDD
Methodology (**B**)	artificial intelligence, AI, machine learning, deep learning, neural network, natural language processing, NLP, computer vision, explainable AI, XAI, interpretable machine learning
Data Source (**C**)	social media, Twitter, Facebook, Instagram, Reddit, online social network, OSN
Objective (**D**)	detection, prediction, diagnosis, classification, screening, monitoring
Query	(**A**) AND (**B**) AND (**C**) AND (**D**)

Table 2.1: Search query used for literature retrieval

(for example, those that did not involve social media data or did not employ AI-based methodologies). A further 127 entries were removed as they were not actual research papers (e.g., conference names or indexing artefacts). To ensure the quality of the selected literature and to keep the review focused on well validated and reproducible methods, we applied a quality filtering step. For journal articles, only those published in Q1 or Q2 journals (based on the SCImago Journal Rank) were retained. For conference proceedings, we included only papers from venues ranked A*, A, or B by the CORE ranking. This quality screening excluded 1,068 studies. I acknowledge that quality based screening is not part of the standard PRISMA procedure. However, one of the main goals of this review was to reliably compare AI models rather than simply list every available study which would further introduce the unnecessary sparsity, and would have made the analysis both time intensive and less reliable. I also recognise that some preliminary cutting edge work might appear in lower-ranked venues, therefore, I kept the boundaries reasonably broad by including Q2 journals and B ranked conferences.

In total, 534 studies met all inclusion criteria and were included in the final review. This selection process ensures that the survey captures the current state of the art in AI-driven mental disorder analysis on social media, with an emphasis on methodologically sound and high-impact research.

2.1.c Organisation

To provide a comprehensive understanding of the domain, we commence with an overview of traditional diagnostic methods for treating mental disorders in Section 2.2. Subsequently, we examine recent data and AI driven research studies in Section 2.3. This section covers the State-Of-The-Art (SOTA) Machine Learning

(ML) methods (specifically modern Deep Learning (DL)) for MDD, outlining their respective advantages and disadvantages, and offering our perspectives on their applicability. Notably, research on XAI for mental healthcare, as well as general healthcare, is in its early stages. Section 2.4 presents a review of existing research on explainability, accompanied by our insights that contribute to the development of XAI solutions for mental healthcare. To facilitate a deeper understanding of the experimental setup, Section 2.5 outlines prevalent practices in experimental evaluation and existing datasets. This encompasses the selection of experimental datasets and evaluation approaches. Section 2.6 identifies and discusses the key issues and challenges inherent in this research. Finally, Section 2.7 concludes the chapter.

2.2 Traditional Diagnostic Methods

Mental health conditions are typically diagnosed by qualified healthcare professionals, such as psychiatrists and clinical psychologists, through structured face-to-face clinical interviews. These interviews often include a mental status examination and can be supported by standardised questionnaires that play a vital role in assessing mental conditions. The Depression Inventory by Beck et al. [24] and the The Center for Epidemiologic Studies Depression (CES-D) Scale [201], each comprising 21 and 20 questions respectively, assess the mental status for depression diagnosis. Gold standards for estimating depression severity include the Beck's Depression Inventory (BDI)-II [25], Hamilton Depression Rating Scale (HDRS) [94], and Patient Health Questionnaire 9 (PHQ-9) [129], focusing on key depression symptoms such as sadness, pessimism, loss of interest, and fatigue. Recognised questionnaires for anxiety detection include the Generalised Anxiety Disorder 7 (GAD-7) [233] and Beck Anxiety Inventory (BAI) [234]. For the assessment and diagnosis of Eating Disorders (ED), the Eating Disorder Examination Questionnaire (EDE-Q) [53] and the Eating Attitudes Test 26 (EAT-26) [80] are employed.

Psychiatrists and clinical psychologists follow international standards such as the Diagnostic and Statistical Manual of Mental Disorders 5 (DSM-5) [19], a manual listing all mental disorders, along with their standard diagnostic methods, which provides the required symptom and associated duration criteria of each mental disorder for the diagnosis [19]. Main symptoms of various mental disorders are identified through these standards. For example, depression is characterised by a loss of interest in previously enjoyed activities, feelings of hopelessness, reduced motivation concentration, energy, and libido, and changes in cognition [19]. Anxiety manifests through symptoms like tachycardia, dizzi-

ness, shortness of breath, trembling, worry, anger, and fear [19]. Eating disorders are identified through signs such as problematic eating habits, abnormal changes in weight, changes in mood and unhappiness with body shape [19].

Despite the well-established area of mental healthcare relying on traditional diagnostic methods, there are inherent limitations in the current approaches [67, 120]. (i) The *frequency of monitoring sessions* is typically limited to once per week. It hinders the ability to track the course of mental conditions in real-time. (ii) The subjective nature of clinical judgement introduces variability between diagnoses, as the outcome depends on how each clinician interprets and evaluates the symptoms. (iii) *Accessibility to therapy sessions* may be negatively affected by factors such as financial constraints and societal circumstances, particularly during crises like pandemics and wars. (iv) *The phenomenon of patient prejudice* is seen quite often, as patients' responses to self-questionnaires may exhibit bias due to individual interpretations, contextual factors, and potential misunderstandings of the questions, including intentional concealment of information. (v) Additionally, *human related barriers*, including cultural diversity, can pose challenges for patients in effectively communicating their needs due to language and traditional limitations.

For a timely detection of mental disorders while addressing the existing limitations, it is crucial to explore widely accessible, impartial, consistently monitored, and immediately available alternatives. The ubiquitous OSM platforms are used by a vast majority of individuals and therefore serve as a repository of substantial volumes of individual data, as people share their intra- and inter-personal experiences and communicate with others [84]. Moreover, individuals with a tendency toward mental conditions use distinctive language patterns with salient negative emotions and tend to discuss relational and health-related concerns [61, 293]. Social media platforms provide anonymity, which gives individuals the opportunity to share their sensitive personal experiences and issues more openly than in clinical settings [118], and they openly describe their challenges regarding their intimate life, such as loss of libido and reduced arousal [26]. Overall, as social media can reveal mental health based symptomatic information [256] such as emotions [293], low social activity [61], sleep disturbances, concerns regarding sexual life [26], and concentration issues [91], they can be used as a potential tool for the analysis of mental health conditions. To address the limitations observed in conventional diagnostic approaches, this study aims to examine preliminary diagnostic methods for MDD using data-driven techniques. Additionally, we will conduct a thorough evaluation of the strengths and weaknesses of current data-driven methods, followed by providing our views on potential strategies to mitigate identified limitations.

Table 2.2: Some common mental disorders, their definitions, and indicative examples of OSM posts

Mental disorder	**Definition**	**Example of OSM post**
Depression	Negative changes marked by affect, cognition, mood, and neurovegetative functions lasting at least 2 weeks [19, 83].	*Can Someone Cheer Me Up? - I'm diagnosed with depression. I will sometimes out of nowhere...*
Anxiety	Conditions characterised by excessive fear and anxiety, along with associated mental and behavioural disorders [19, 205].	*Anxiety is seriously affecting my life and it is ruining my life and I don't know how I'm going to provide for myself in the future*
Eating Disorder	A persistent disturbance in eating behaviour that leads to altered food consumption and significantly impairs mental and behavioural functioning [19, 2].	*Getting treatment for an eating disorder is awful and I wish I never started it. (Vent / cry session haha)*
Bipolar Disorder	Abnormally and persistently elevated, expansive, or irritable mood, along with increased activity or energy [19, 120].	*Finally after years long struggle I considered seeking a psychiatrist. And just diagnosed bipolar disorder. She prescribed me with two medicines*
Schizophrenia	Abnormalities in delusions, hallucinations, disorganised thinking, and grossly disorganised or abnormal motor behaviour [19, 161].	*Hi. I'm **** and I'm a ********* with schizophrenia and I take a medication. I also have evil hallucinations and it scares me. Is it my schizophrenia or is it really the Devil?*

Continued on next page

Mental disorder	Definition	Example of OSM post
Suicide Ideation	Thoughts or contemplation of taking one's own life or self-inflicted harm [19, 39].	*I'm having a really hard time with a lot of things in my life right now. My father started to verbally abuse me and my sisters since I was 6 or 7. I want to commit suicide.*
Obsessive Compulsive Disorder (OCD)	Characterised by the presence of recurrent and persistent thoughts, urges, or images, along with repetitive behaviours or mental acts [19, 187].	*I've been going to a doctor to see if I had autism and while there they also diagnosed me with OCD.*

2.3 Data-driven Methods for Detection

Healthcare approaches are rapidly evolving with technological advancements and the integration of data- and AI-driven techniques, contributing to enhanced outcomes [282, 290, 2]. A particularly promising area where technology can play a pivotal role is the timely detection of mental disorders. Recent research progress suggests a foreseeable future in this direction. While various individual data sources can contribute to MDD, our primary focus in this chapter centres on text-based OSM platforms, such as 𝕏 (formerly Twitter) and Reddit. These platforms, known for their widespread usage and expressive nature, contain valuable content such as users' posts, online behaviours, and social network interactions. Table 2.2 summarises the most frequent mental disorders along with their definitions and examples.

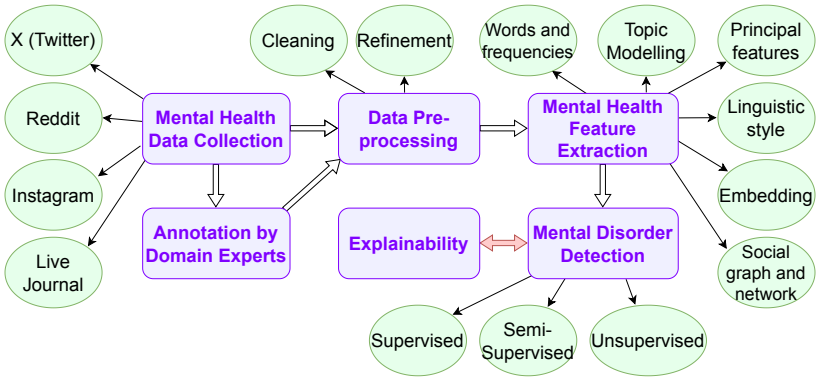


Figure 2.2: Workflow of data-driven methods for MDD

In our exploration of current methodologies for MDD via OSM, we observe a common workflow, depicted in Figure 2.2. This standard process involves several key steps: MH data collection, preprocessing, MH feature extraction, MDD modelling and evaluation, and explainability. The MH data collection involves sourcing necessary data from various OSM platforms. Following this, data preprocessing cleans and refines the raw data through basic cleaning and denoising procedures. MH feature extraction transforms it into a modelling-friendly representation of MH-related features using various techniques (Section 2.3.a). The MDD modelling step generates a model that is able to extract interesting patterns and insights specific to MH, providing a basis for accurate assessments for MDD. This modelling may follow supervised, unsupervised, and semi-supervised approaches. However, the literature has found supervised models as the most promising (Section 2.3.b).

The subsequent explainability step generates insights into the decision-making

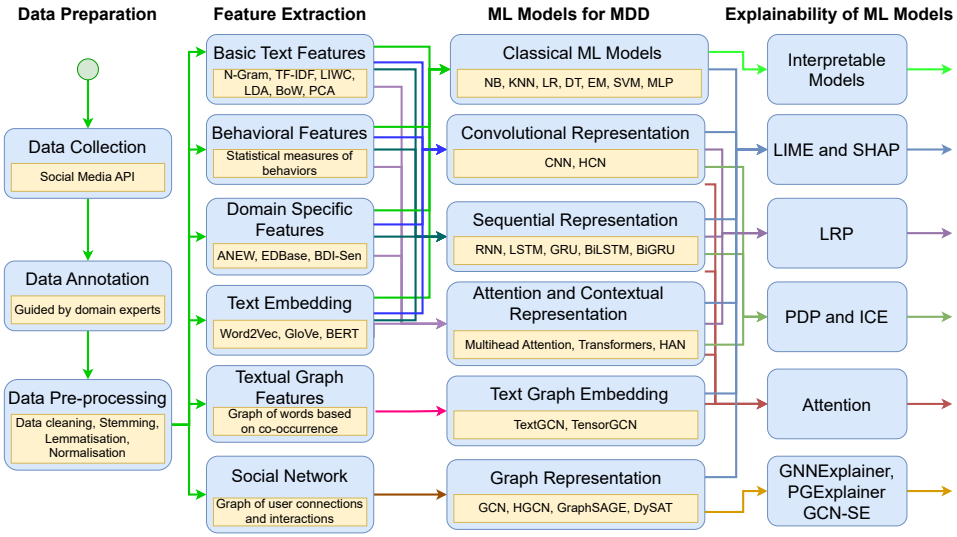


Figure 2.3: XAI pipeline for MDD via social media

process of MDD models, enhancing their overall interpretability (Section 2.4). Figure 2.3 shows a detailed pipeline of XAI methods for MDD from a technical perspective. All the techniques shown in the pipeline are discussed in sufficient detail in the following sections.

2.3.a Mental health feature extraction

The OSM data used for MDD contain lots of information in various formats, including predominantly textual content in unstructured form, associated metadata, user behavioural data, and social communication networks. Transforming this raw data into a set of relevant features or a numeric representation is essential for ML models to process and make sense of it. An effective feature extraction and representation enables a deeper understanding of the underlying patterns and trends within the OSM data.

Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and N-gram [27, 45, 213, 270, 227, 231, 280, 137, 17] are fundamental feature extraction techniques in NLP. BoW emphasises word existence, TF-IDF considers word frequency within documents, and n-gram extracts contiguous sequences of n tokens for vectorisation.

Latent Dirichlet Allocation (LDA) [149, 93] is an iterative probabilistic model that assigns topics to documents, which makes it useful for uncovering latent thematic structures within textual data. The extracted topics can be used as features for ML modelling.

Principal Component Analysis (PCA) [134, 247] is a dimensionality reduction technique used to identify and retain the most important or principal features, while avoiding the curse of dimensionality from a dataset. These features can be used for ML modelling.

Linguistic Inquiry and Word Count (LIWC) [54, 247, 231] is a text analysis tool that examines psychological and linguistic content by categorising words into various categories, such as emotions, social processes, cognitive processes and linguistic style. LIWC scores are values assigned to the words and can play the role of features in data analysis.

Word2Vec [157] leverages neural networks to learn numeric representations of different words from a large corpus. It is widely used today for its simplicity and effectiveness. There exist two primary approaches: Continuous Bag of Words (CBOW) and skip-gram. There are pre-trained versions of the Word2Vec model which can easily be used to vectorise the textual inputs for modelling [235, 142].

Global Vectors for Word Representation (GloVe) [191] is a statistical model used for text embedding similar to word2vec. It is an unsupervised learning algorithm that maps textual features into meaningful vector spaces. Like word2vec, GloVe is a pre-trained model as well, and can be used for textual vectorisation of textual mental health related documents [235, 4, 66, 231, 232].

Bidirectional Encoder Representations from Transformers (BERT) [64]. Contextualised embedding is an important phenomenon in mental health detection if the input is textual data since words can have different meanings depending on the other words in the same documents. BERT [64] is transformer based pre-trained language model can be used to generate contextualised embedding of the textual inputs which further can be used for the classification of the different mental health disorders with machine learning and deep learning models. The details about the contextualised representations will be explored in further chapters. Contextualised embedding is very important in MDD [290, 23], especially for textual data, as words may carry distinct meanings based on their context. BERT (detailed later), a transformer-based pre-trained language model, offers contextualised embeddings for textual inputs. These embeddings enhance the classification of various mental disorders using machine learning models. Further details of contextualised representations are presented later.

Graph representation of text [166]. Representing textual inputs as a graph structure enables capturing intricate structural patterns within words and documents. Examples like TextGraph Convolutional Network (GCN) [275] and TensorGCN [141] demonstrate effective graph-based representations that can be adopted for mental health analytics. These models analyse graphs constructed with words, sentences, and their relations using GCN [127] (detailed later).

Social network representation [156, 131]. OSM platforms are actively used by a significant number of individuals worldwide [6]. Research suggests a correlation between an individual's mental health and the mental health status of their friends, highlighting the potential spread of mental health issues within peer networks [206, 7]. In fact, studies by Rosenquist et al. [206] indicate that a person's depression is influenced not only by their friends but also by their friends' friends, and even by their friends' friends' friends. Similarly, Kiuru et al. [128] found that, over time, individuals' symptoms tend to shift toward the average mental health levels of their peer group. Considering these findings, we believe that the infusion of social network features into individual-level features can offer clear benefits for MDD.

2.3.b Machine learning on social data for MDD

The task of MDD from social data poses the challenges of dealing with high-dimensional, imbalanced, and non-linear data, along with the need to understand complex mental health related properties and social relationships. As classical ML techniques often rely on robust feature engineering, it is difficult for them to deal with such complicated data and the associated challenges. Therefore, recent advancements have seen a paradigm shift towards DL-based methods, surpassing those based on classical ML in their abilities without the need for feature engineering [282]. Given that much of the OSM data is textual, current studies often employ Deep Neural Network (DNN), such as Recurrent Neural Network (RNN), and their variants like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) to learn sequential representations. Additionally, various other types of DNNs are utilised to capture different properties. In this section, we explore the different types of representation learning using these DNNs, and discuss the studies that leverage them for MDD.

Convolution-based representation learning

The convolution operation is a fundamental concept that represents the combination of two signals to produce a third signal. The power of convolution paved the way for Convolutional Neural Network (CNN) [132]. They have found applications across various domains, including natural language processing. CNN models are capable of recognising patterns in textual input embeddings which can be beneficial to identify signs of the mental disorders in OSM posts. Several studies have harnessed their capabilities to explore and understand OSM texts related to mental health [291, 81, 240]. Gaur et al. [81] developed a dictionary

[6] www.datareportal.com/social-media-users

of suicide risk severity using medical knowledge bases and suicide ontology to detect clues about suicidal thoughts and actions. They also generated a labelled dataset of suicide risk severity from Reddit. For their multiclass classification, they considered a simple CNN model. Similarly, Zogan et al. [291] employ CNNs along with other networks to detect depression on 𝕏 and analyse the impact of depression during the pandemic. The authors introduce an advanced model based on CNN and hierarchical attention, incorporating both word- and tweet-level encodings to identify depressive tweets. Nevertheless, there are notable limitations to this study. The classification focuses solely on texts, treating depression detection as a binary classification task. However, assessing the severity or level of depression in a fine-grained manner would be more helpful. The social and communication network of users and textual posts are not considered in the study.

Sequential representation learning

In handling long-term sequential data generated on OSM and exhibiting effective parameter-sharing capabilities, RNNs outperform classical Multi Layer Perceptron (MLP)s and CNNs. Among the various RNN types, our focus is on three main types: vanilla RNN, LSTM, and GRU. The utilisation of sequential representation learning is crucial, as representing textual posts as sequences allows for a more precise detection of mental disorders. This approach enables an accurate capture of relationships among words and sentences.

Vanilla RNNs. In these networks, the hidden layer and output at time t are calculated as $y_t = f(W_{xh}x_t + W_{hh}y_{t-1})$ and $h_t = g(W_{hy}y_t)$, respectively, where W_{xh}, W_{hh} and W_{hy} are trainable parameters, x_t is the input at time t, and $f(\cdot)$ and $g(\cdot)$ are differentiable activation functions. Vanilla RNNs are susceptible to the problem of vanishing gradients.

Long Short-Term Memory (LSTM) [102]. It is an RNN variant that addresses the vanishing gradient problem inherent in vanilla RNNs and handles long-term dependencies more effectively. An LSTM node comprises three gates - *input* gate, *forget* gate, and *output* gate - which collectively determine how to update the cell and hidden states. The *candidate cell state* and *cell state* are responsible for short-term and long-term memory storage, respectively. *The hidden state* is the output of the LSTM at each time step. In a regular (unidirectional) LSTM network, information flows from the first element of a sequence to its last element in a single direction. It may be limiting its potential in cases where the meaning of a word or sequence depends on both preceding and succeeding elements (e.g., OSM texts). Bidirectional Long Short-Term Memory (BiLSTM)s, on the other hand, process the input data in both forward and backward directions simultaneously,

allowing them to capture contextual information from both sides.

Gated recurrent unit (GRU) [50]. It is another RNN variant that addresses the vanishing gradient problem with a simpler architecture than LSTM using *reset* and *update* gates to regulate the flow of information. In GRUs, there is no explicit memory cell with long-term memory features. Instead, a *candidate hidden state* is introduced to update the *hidden state*, which is the output of the GRU model. Similar to BiLSTMs, Bidirectional Gated Recurrent Unit (BiGRU)s capture the contextual information from both forward and backward directions.

MDD using RNNs. Due to their capabilities of learning sequential representations from OSM data, RNNs have been widely employed in previous studies on MDD [83, 250, 232]. In their study, Ghosh and Anwar [83] approach MDD as a regression problem, aiming to provide a fine-grained assessment based on the disorder's intensity. They extract a variety of features and process them through an LSTM network with Swish activation function to predict the intensity of depression. The study utilised an existing 𝕏 dataset labelled with binary classes related to depression diagnosis. Given its binary nature, the authors applied a weak-labelling approach to assign depression intensities using measures of sadness and semantic similarity with depression. They employed a self-supervised learning approach to predict depression severity. While the model yielded promising results, there are several aspects that *require further consideration*. *First*, the model employed an NLP-based relabelling technique without the involvement of mental health professionals and without a deep exploration of the psychological underpinnings. *Second*, for tweets known for their informal linguistic style, a more in-depth analysis of semantic meaning with advanced deep learning architectures and NLP techniques could enhance the model's performance *Finally*, the model does not account for user interactions

Attention and contextual representation

OSM posts often exhibit linguistic, behavioural, and symptomatic indicators associated with mental health difficulties [256, 84], and assigning higher importance to these indicators can be vital for accurate MDD. In the context of DL, emphasising "attention" to specific language patterns within OSM posts is likely to enhance the performance of MDD models [251]. The attention mechanism was first proposed by Bahdanau et al. [21] to enhance sequence-to-sequence (Seq2Seq) performance by focusing on specific important parts of the input sequence. In 2017, Vaswani et al. [252] employed attention to introduce the transformer architecture. The attention mechanism involves three key matrices - query (Q), key (K), and value (V) - learned from the input data during the training process. The query represents a specific aspect of the input data that

the attention mechanism wants to focus on. The key provides information that can be matched against the query. It helps in deciding which information is most relevant to the query. The value contains the actual information associated with the key, which is used to respond to the query. These matrices are defined as $Q = W^q \times \mathbf{X}$, $K = W^k \times \mathbf{X}$ and $V = W^v \times \mathbf{X}$, where W^q, W^k and W^v are trainable weight matrices corresponding to the query, key, and value components and $\mathbf{X}$ is the input data or the sequence of elements to be used. The attention values are computed using scaled dot product attention shown in Equation 2.1, where d_k is the dimension of the query and key. The scaling by d_k is applied to achieve gradient stability and mitigate vanishing/exploding gradient problems.

$$Attention(Q, K, V) = softmax(\frac{Q \times K^T}{\sqrt{d_k}}) \times V \tag{2.1}$$

By stacking the scaled dot product attention, the Transformer introduces a key component called multi-head attention shown in Equation 2.2, where $h_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$, and the parameter matrices vary for each head (self-attention layer).

$$MultiHead(Q, K, V) = Concat(h_1, h_2, ..., h_h)W^O \tag{2.2}$$

Transformers and BERT. After the introduction of Transformers [252], it sparked a revolution in NLP. Unlike traditional sequence-to-sequence models, Transformers rely on the powerful attention mechanism and do not require recurrent connections, making them highly effective for a wide range of NLP tasks. Transformers play an important role in capturing long-term dependencies and developing a contextualised understanding of OSM posts, which enables the detection of anomalies in users' mental health. The Transformer architecture comprises an encoder and a decoder. The encoder consists of six stacked layers, each containing multi-head attention, feed-forward, and layer normalisation units. This architecture allows the model to capture complex dependencies in the input data. The decoder, on the other hand, also consists of six identical layers but employs *masked* multi-head attention to prevent the model from attending to future tokens in the sequence, avoiding data leakage. The Transformer architecture marked a milestone in modern NLP, which led to the development of pre-trained language models. BERT [64] is one of the most widely used pre-trained language models for obtaining contextualised embeddings for words and sentences. It adopts a bidirectional multilayer transformer model, comprising a stack of transformer encoders. There are two primary versions of BERT: $BERT_{BASE}$ and $BERT_{LARGE}$. The $BERT_{BASE}$ model consists of 12 trans-

former layers with a hidden size of 768 and 110 million trainable parameters, while $BERT_{LARGE}$ features 24 transformer layers with a hidden size of 1024 and 340 million parameters. The BERT model can be fine-tuned by adding task-specific layers, such as a fully connected layer. Then it can be used for a wide range of tasks, including mental health research with OSM texts. There have been some follow-up works on extending BERT with unique features and improvements, catering to different requirements in natural language understanding and processing. For example, *DistilBERT* [212] is a distilled, smaller and faster variant of BERT with fewer parameters. *RoBERTa* [287] is an optimised variant of BERT. It is trained on more data, omits the next sentence prediction task of BERT, utilises dynamic masking during training and improves the contextual understanding of texts.

MDD using transformers. The transformer-based models have had a significant impact on research in mental health analytics [290, 2, 202, 40, 151, 1]. Ragheb et al. [202] developed a comprehensive ensemble method for detecting at-risk users with transformer-based models. In their method, they introduce some noise into the textual feature representation generated by transformer-based contextual embedding models. The resulting noise-induced representation is then used for classification to identify the at-risk users. The noisy learners help in generalising the model for multiple disorders such as depression, anorexia, self-harm and suicide, and reducing overfitting. However, the proposed method focuses only on post content relying on clear expressions of disorders, which makes it less suitable for early detection. Furthermore, although their model is designed to detect three different disorders, there is no parameter sharing among them and each disorder is treated independently. In [40], Cao et al. adopted masked language structures to feed transformer-based contextual embedding models in order to detect latent signs of suicide ideation with an LSTM model. A limitation of their approach is the disregard of severity degrees associated with suicide ideation, a critical consideration, especially when assessing latent risk detection. Additionally, the method neglects the interactions among users and their social network friends, which often provide useful clues about the mental state of users. Utilising the 𝕏 dataset [225], Zogan et al. [290] developed DepressionNet. It leverages both user behaviour and posts, and extracts a wide range of features including social network activities, emotional content, domain-specific information, and topic-related attributes. DepressionNet uses extractive-abstractive summarisation technique to summarise the historical posts of each user. This process begins by embedding each post with BERT. Subsequently, important posts are identified using K-Means clustering. Then Bidirectional and Auto-Regressive Transformers (BART) [135], a denoising autoencoder for pre-training sequence-to-sequence

models, is employed to generate abstractive summaries of the posts. These summarised posts are further embedded and passed through BiGRU and attention layers before being concatenated with the stack of BiGRU vectors representing user behaviours. While DepressionNet takes a step forward in depression detection, it has *two major limitations*. *First*, this model treats the problem of depression detection as a binary classification task, the classes being whether a person is depressed or not. Instead of binary classes, a more fine-grained severity levels or a severity measure will be a more informative outcome. *Second*, it utilises the textual posts and user behaviours for representation learning, but does not consider any kind of social or communication network among users and posts. The network contains some very useful information for MDD, given that the mental health related thoughts and experiences are propagated through such networks on OSM platforms. Abuhassan et al. [2] introduced EDNet, a multimodal deep learning model designed to identify the different types of OSM users engaged in Eating Disorders (EDs). This model integrates various data sources, including historical posts, user biographies, and online behaviours from 𝕏, and employs a multi-class classification approach to differentiate between distinct user types, such as ED-users, healthcare professionals, communicators, and non-ED users. EDNet consists of several deep neural layers, including an input layer, an embedding layer, a representation layer, a behaviour modelling layer, and an output layer. This layered architecture facilitates a comprehensive understanding of diverse modalities within OSM data, encompassing users' historical tweets, biographies, and online behaviours. The authors employed the BERT model to generate contextual embeddings for textual tweets and user biographies, while incorporating variations of temporal convolutional layers and BiGRU layers to capture intricate patterns associated with EDs. However, the model specifically focuses on English language, which limits its effectiveness when applied to other languages. Another challenge arises when a user undergoes changes in their engagement type over time. For example, an ED user becoming non-ED a month later, or a communicator starting to experience ED. This kind of occasional behaviours leads to misclassifications, especially when users frequently use phrases associated with EDs in their posts. Furthermore, it does not give a fine-grained information about the severity of an ED user.

Graph representation learning

Current research emphasises the significant impact of social connections on mental well-being [206]. There exists a correlation between an individual's mental health and the mental health status of their friends. Individuals are 93% more likely to experience depression if one of their friends is experiencing

it [206]. To comprehensively understand users' mental well-being, capturing the graph representation of their social networks is essential. This approach discovers structural patterns, information flow dynamics, and community-level interactions. It aids in identifying their support systems, detecting isolation, and understanding the dissemination of mental health-related information within the network. Integrating this graph representation with textual information of posted content provides a holistic view, revealing insights into users' contextual interactions. Additionally, the posts themselves also contain useful structural information that can be captured by a graph representation of their textual content [275]. Recognising the capabilities of neural networks, it is imperative to explore DL models that are good at working with graph structures. It takes our attention to graph neural networks (GNNs). They utilise message passing to gather information from neighbouring nodes and edges, enabling the learning of node and edge representations based on complex structural information inherent in the input graph. There are several types of GNNs depending on the message passing and node/graph based representation, including GCN, Graph Attention Network (GAT), GraphSAGE and Dynamic Self-Attention Network (DySAT).

Graph Convolutional Network (GCN). An ordinary GNN aggregates representations from neighbouring nodes, and therefore the nodes with large degrees tend to have a significant influence in their representations. On the other hand, those with smaller degrees have a minor influence. This can lead to a gradient explosion problem. This issue is addressed by GCNs [127] using fast approximate spectral graph convolution. They are able to capture long-term dependencies and complex representations over the graph effectively. The updated hidden layer of nodes (or activations) $H^{(l+1)}$ in the $(l+1)$-th layer of a GCN, after applying a graph convolution operation, is computed as $H^{(l+1)} = \sigma\left(\tilde{D}^{-0.5}\tilde{A}\tilde{D}^{-0.5}H^{(l)}W^{(l)}\right)$, where $H^{(l)}$ denotes the node activations in the l-th layer, $\sigma(\cdot)$ denotes the activation function, $\tilde{A} = A + I_N$ is an adjacency matrix with added self-connections, $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ is a diagonal matrix of $\tilde{A}$, $W^{(l)}$ is a layer-specific trainable weight matrix, and $H^0 = X$ is the initial node representations. As GCNs use normalised Laplacian matrix for convolution and aggregation of node features, they have to face scalability challenges, especially in large graphs. This limitation has led to the development of more scalable and flexible GNN architectures.

Text graph embedding. While textual contents are generally represented with text-based embeddings, they also contain some structural information. It enables them to be represented as graphs and be processed through GNNs. For example, TextGCN [275] and TensorGCN [141] adapt GCN on graph structures extracted from textual contents for text classification. TextGCN [275] constructs a

graph based on the given set of documents and the words in them. A node is created corresponding to each word as well as document, and connections among them are established using pointwise mutual information (PMI) and TF-IDF. Similarly, TensorGCN [141] aims to enhance the comprehension of syntactic, semantic, and sequential relationships within text data. Operating on graph structures where nodes represent words and textual documents, TensorGCN establishes connections between words and documents using TF-IDF. The model integrates tensors representing three distinct graphs: a semantic-based graph, a syntactic-based graph, and a sequential-based graph. Semantic-based connections are determined by LSTM-generated embeddings' similarity, syntactic-based edges are extracted using the Stanford CoreNLP parser for grammatical dependencies, and sequential-based edges are formed through PMI. Following graph construction, TensorGCN employs inter-graph and intra-graph propagation processes. Intra-graph propagation gathers and consolidates information within a given graph, while inter-graph propagation facilitates information exchange between distinct graphs within the tensor. These graph creation and propagation methods enable the application of GCN for text classification. The scalability issues present in GCN are inherently present in both TextGCN and TensorGCN, making them difficult to work with long texts.

Graph Attention Network (GAT) [253]. GATs address the limitations of GCNs in terms of scalability by incorporating attention mechanism within the GNN. The attention enables the nodes to selectively aggregate information from their neighbours based on their importance. Each neighbouring node is assigned an attention coefficient, which is used to make more contributions from more important nodes to the aggregation process. The attention coefficients α_{ij} between nodes i and j is computed using Equation 2.3, where $||$ represents concatenation, h_i and h_j are hidden representations of i and j, and $\mathbf{a}$ and $\mathbf{W}$ are trainable parameters.

$$\alpha_{ij} = \frac{\exp(LeakyReLU(\mathbf{a}^T[\mathbf{W}h_i||\mathbf{W}h_j]))}{\sum_{k\in\mathcal{N}_i}\exp(LeakyReLU(\mathbf{a}^T[\mathbf{W}h_i||\mathbf{W}h_k]))} \tag{2.3}$$

The information from neighbouring nodes $j \in \mathcal{N}_i$ of i are then aggregated based on their attention coefficients α_{ij} and learnable parameters $\mathbf{W}$, and passed through a non-linear activation function σ, to generate its updated feature representation h'_i from h_i, as shown in Equation 2.4. To stabilise the learning process of self-attention, the original GAT model employs multihead attention [252], with K = 3, using Equation 2.5, where $||$ represents concatenation, α^k_{ij} is the attention coefficient computed by the k-th attention mechanism, and $\mathbf{W}^k$ is the corresponding weight matrix. Using multiple heads allows the model to capture

different aspects of node relationships.

$$h_i' = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W} h_j\right) \tag{2.4}$$

$$h_i' = ||_{k=1}^{K} \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^{k} \mathbf{W}^{k} h_j\right) \tag{2.5}$$

GraphSAGE [94]. Traditional GNNs like GCN and GAT require knowledge of all nodes in the graph during training, including those that will be part of the test set, leading to data leakage concerns. GraphSAGE tackles this problem by adopting a novel approach. The main idea behind GraphSAGE is to learn feature aggregation methods from local neighbourhoods by utilising node sampling techniques. It significantly enhances the efficiency, flexibility, and scalability of GNNs, making them suitable for larger graphs. In a nutshell, GraphSAGE takes as input a graph $\mathcal{G}(V, E)$, node features, a depth parameter (K), trainable weight matrices, a non-linear activation function, and an aggregator function. The aggregator function plays a pivotal role. It aggregates information from a node's sampled neighbours and concatenates it with the node's own representation from the previous layer. Subsequently, these concatenated representations are normalised at each layer. The final embeddings are obtained from the last layer at depth K. Depending on the given context and specific requirements, different aggregation methods can be employed in GraphSAGE, including mean aggregators, pooling operations or aggregation based on neural networks.

MDD using GNNs. Some recent studies have adapted GNNs for MDD, in order to learn structural patterns within data. This is done by utilising users' textual graphs generated from posts [166, 168] and social networks [140, 195, 216, 156, 131].

Depression severity assessment is a pressing concern within MDD. Naseem et al. [166] treat this problem as a fine-grained classification task, in which they consider four classes of varying severity from *minimal* to *severe*. Their method utilises several layers of DNNs and leverages a dataset extracted from Reddit. The method begins with the preprocessing of user posts, which are then processed through a representation layer with TextGCN. It generates their numeric embeddings by constructing a textual graph of posts (as nodes) based on their constituting words, and applying GCN on it to learn the structural post (node) representations. These are fed into a BiLSTM layer augmented with an attention mechanism to learn sequential properties. As the severity classes are ordinal in nature, the resultant vectors are passed into an ordinal classification layer to

generate the output severity class. While this method exhibits promising results, it's important to note that textual data alone may fall short in identifying the depression severity level. Depressed individuals may not always explicitly post about their mental health status on OSM platforms. A more holistic approach that incorporates historical and behavioural data is important for a generalised model. Furthermore, leveraging social interactions and communications between OSM users can enhance the reliability and effectiveness of such models. Along a similar direction, GHAN [168] utilises textual posts extracted from Reddit to identify the various levels of suicide ideation, including *support*, *indicator*, *ideation*, *behaviour*, and *attempt*. TensorGCN is employed to exploit semantic, syntactic, and contextual information for embedding, after which an attentive transformer encoder is used to capture temporal information across the user's posts. As the classes are ordinal, an ordinal classification layer is used to enhance the model's performance. While the model demonstrates great performance compared to the baselines, its generalisation may face challenges related to factors such as language and demographics. Although linguistic elements have proven beneficial, there are instances where they may be insufficient.

Social interactions (daily connections and relationships) can influence an individual's mental health in both positive and negative ways. Supportive connectedness within family and close friend environments can protect individuals from the symptoms of mental disorders [262, 130], whereas studies show that severity of mental health conditions may converge toward the peers of an individual and can even correlate across several degrees of separation [128, 206]. Building upon this, Liu et al. [140] developed a heterogeneous network approach by taking into account an array of factors, encompassing social interactions, personality traits, social status, physical health, and overall well-being. A heterogeneous graph is constructed by amalgamating these diverse facets, in which each user is represented as a node and multiple types of links are established between different pairs of users. Thus, the MDD task is treated as a node-classification problem. The leveraged data is obtained from various sources including smartphones, social media accounts, and wearable devices. While this multi-centric approach is comprehensive and promising, it introduces unique challenges. Specifically, the collection of non-public personal data (personality traits, physical health metrics, well-being indicators) requires access to sources outside standard OSM platforms. Furthermore, the fusion of dissimilar data types collected through distinct mechanisms (wearable devices, psychological assessments, and social media) presents significant integration challenges not encountered when working solely with social platform data. Introduced by Pirayesh et al. [195], MentalSpot is another method incorporating social interactions. To enable this

study, the authors firstly generated a dataset (PsycheNet) with users' social contagion network from 𝕏. The method begins by embedding user tweets with GloVe, followed by deploying 1D convolutional maps, which considered the top-k friends of each user. While the authors considered the social contagion among users, they did not examine the role of GNNs. The adoption of GNN models would deepen the analysis of the user's social circle's role by employing message passing, thereby leveraging the advantages of the graph structure. However, they use only the textual data of tweets and treat MDD as a binary classification task. To identify suicide ideation among OSM users, Sawhney et al. [216] developed a model (hyper-SOS) based on hyperbolic GCN (HGCN) [42]. Hyper-SOS considers the historical OSM posts of users and uses them together with social interactions through replies, comments, and quotes. The textual posts are embedded with BERT, and Hawkes temporal emotion aggregation (HEAT) mechanism is used to synthesise the posting history. This method adopts the Hawkes stochastic process for aggregating the historical posts using an exponential kernel. Once the posts are embedded and the connections are established, the generated graph is processed through an HGCN for detecting suicide ideation cases. HGCN is a variant of GCN that uses hyperbolic geometry to capture long range connections and achieves enhanced robustness, superior performance, and improved interpretability. The model, in its current form, is applied on static graphs, which is a hurdle when considering real-time dynamic data of evolving OSM.

Some studies exploit advanced graph structures. MentalNet [156] utilises graph-structured data by forming an ego network of users as a heterogeneous graph. Each user is denoted as a node, and connections are established based on interactions through replies, mentions, and quote tweets. The study extends the existing PsycheNet dataset [195] by incorporating interaction data to create PsycheNet-G. The problem is framed as heterogeneous graph classification. Node features are derived by applying an LSTM autoregressive sequence-to-sequence embedding once the graph is constructed. Given the adoption of heterogeneous graphs with repeated convolution, a normalisation process, known as doubly stochastic normalisation, is employed. The constructed graph is then processed through a stack of GCN and convolution layers. MentalNet's strength lies in its holistic approach of considering both tweet content and user interactions. However, it has a few limitations with possible areas of enhancement. It solely relies on textual posts for initial node features, without considering user-specific and behavioural features that can provide further insights into depression detection. One potential direction forward is to shift its focus from binary classification for detection to estimating the intensity level. Furthermore,

model complexity is a concern. Lastly, the model relies on static graphs, but the dynamics of social contagion can shift rapidly in evolving OSM platforms. Incorporating dynamic graph structures and properties could bolster the model's real-world applicability. Kuo et al. [131] extend the MDD research into dynamic settings with their ContrastEgo network. They leverage the PsycheNet-G dataset [156] to construct a heterogeneous graph based on social interactions. Once the temporal graph is generated, interpersonal dynamics are established using GNN models, and a transformer encoder captures temporal dynamics. The final layer employs contrastive learning with binary cross-entropy loss and supervised contrastive loss to maximise user agreements. While ContrastEgo can capture the temporal evolution of communication patterns to some extent, it relies solely on textual posts for node features, neglecting user-specific and behavioural attributes. The model performs binary classification for detection, but a fine-grained analysis of severity levels or intensity estimation is crucial for a comprehensive mental health assessment. Lastly, as ContrastEgo is a complex model, challenges arise in terms of its explainability.

Table 2.3: Significant studies on ML for MDD from OSM (×: Not explainable, ✓: Can be explained but the paper does not demonstrate its explainability, ✓✓: Explainability demonstrated in the paper)

Study	Disorder	Methodology	Our comments		Dataset	XAI
			Pros	Cons		
Kadkhoda et al. [120]	Bipolar Disorder	Behaviour change graph, Stochastic gradient descent, Random forest, KNN, DT, LR, SVM	Mood changes are considered	Non-robust classifier, Unconsidered OSM post contents and social network. Focuses only on disorder existence	Self collected - 𝕏	×
Sawhney et al. [216]	Suicide Ideation	BERT, Hawkes, Hyperbolic GCN	User interaction, Contextualised embedding	Static social network, only post content based node features. Focuses only on disorder existence	SNAP-BATNET [160] - 𝕏	×
Abuhassan et al. [2]	Eating Disorders	BERT, TCN, BiGRU	User-type identification, Multimodality, Contextual embedding	No social network	Self collected - 𝕏	×

Continued on next page

Study	Disorder	Methodology	Our comments		Dataset	XAI
			Pros	Cons		
Zogan et al. [290]	Depression	BERT, KMeans, BART, BiGRU, CNN	Multi-modality, Contextual embedding	No social network, Focuses only on disorder existence	by Shen et al. [225] - 𝕏	×
Mihov et al. [156]	Depression	GCN, BiLSTM, CNN	Augmented dataset, User interaction	Static social network, Only post contents as node features, Focuses only on disorder existence	Augmented PsycheNet [195] - 𝕏	×
Kuo et al. [131]	Depression	RoBERTa, GCN, Transformer	Dynamic social network, User interactions, Contextual embedding	Only post contents as node features, Focuses only on disorder existence	PsycheNet-G [156] - 𝕏	×
Sawhney et al. [214]	Suicide Ideation	BERT, LSTM, Transformer	Contextual Embedding, Time-aware detection, Considers historical posts	No social network, Only post contents, Focuses only on disorder existence	by Sinha et al.[230] - 𝕏	×

Continued on next page

Study	Disorder	Methodology	Our comments		Dataset	XAI
			Pros	Cons		
Cao et al. [39]	Suicide Ideation	Knowledge graph, BERT, CNN, LSTM	Graphical representation, Contextual embedding, Multimodal data	No social network, Focuses only on disorder existence	Self collected - Sina Weibo and Reddit	×
Ragheb et al. [202]	Depression, Anorexia, Self-harm and Suicide	Transformer-based encoders, Negatively correlated noisy learners	Contextual embedding, Noisy learners, Multiple disorders	No social network, Only post contents, Focuses only on disorder existence	eRisk-2018 (Depression) [144], eRisk-2019 (Anorexia and Self-Harm) [146] - Reddit, UMSD [228]- Reddit	×

Continued on next page

Study	Disorder	Methodology	Our comments		Dataset	XAI
			Pros	Cons		
Zhang et al. [286]	Depression	BERTopic, MLP, GNN, Temporal LSTM, Attention, PLM	Contextual Embedding, Historical user posts, Graphical representation	No social network, Focuses only on disorder existence, Relies only on textual data.	Self collected - Sina Weibo	×
Anshul et al. [14]	Depression	LDA, ResNet50, OCR, Ensemble Learning	Multi-modal features, Ensemble of classifiers, User and depression specific features	Focuses only on disorder existence, Non-contextual embedding, No social network	by Shen et al. [225] - 𝕏 and Self Collected - 𝕏	×
Cao et al. [40]	Suicide Ideation	LSTM, Attention, ResNet	Extracts latent information, Contextual embedding, Multimodal data	No social network, Focuses only on disorder existence	Self collected - Weibo	✓

Continued on next page

Study	Disorder	Methodology	Our comments		Dataset	XAI
			Pros	Cons		
Ansari et al. [13]	Depression	LSTM, Attention, LR, Ensemble Methods	Sentiment Lexicons, symbolic and subsymbolic models	Only text based features, Focuses only on disorder existence, Non-contextual embeddings	CLPsych 2015 [56] - 𝕏, by [196] - Reddit, and eRisk 2018[146] - Reddit	✓
Sawhney et al. [215]	Suicide Ideation	Longformer, BiLSTM, Attention	Ordinal classification, Contextualised embedding	No social network	by Gaur et al. [81] - Reddit	✓
Naseem et al. [166]	Depression	TextGCN, BiLSTM, Attention	Ordinal classification, Contextualised embedding	No social network, Only post contents	Re-annotated Dreaddit [249] - Reddit	✓
Zogan et al. [291]	Depression	CNN, MLP, Attention	Considers both word and sentence importance	No social network, Only post contents, Focuses only on disorder existence	Self collected + by Shen et al. [225]	✓

Continued on next page

Study	Disorder	Methodology	Our comments		Dataset	XAI
			Pros	Cons		
Zhang et al. [284]	Depression	Mental RoBERTa, Transformer, Multi-head soft attention	Ordinal classification, Contextual embedding	No social network, Only post contents	by Naseem et al.[166] - Reddit and by Kayalvizhi [208] - Reddit	✓
Schoene et al.[221]	Suicide Ideation	LIWC, LSTM, Attention	Considers multi-sources, Comprehensive textual analysis	No social network, Only post contents, No contextual embedding, Focuses only on disorder existence	Self Collected + by Schoene et al. [220], Pirinal and Çöltekin [196] - Reddit	✓
Wu et al. [264]	Depression	BERT, Transformer, Self attention	Real time detection, Contextual mood and content embedding	No social network, Only post contents, Focuses only on disorder existence	Self collected - 𝕏	✓

Continued on next page

Study	Disorder	Methodology	Our comments		Dataset	XAI
			Pros	Cons		
Sawhney et al. [217]	Suicide Ideation	BERT, BiLSTM, MLP, Attention	Contextual embedding, Ordinal classification, Robust model	No social network, Only post contents	by Gaur et al.[81]	✓
Wang et al. [258]	Suicide Ideation	Doc2Vec, Multi-head attention,	Extracts emotions, Detects disorder several months earlier	No social network, Only post contents, No contextual embedding, Focuses only on disorder existence	CLPsych-2021 [150]	✓
Amini et al. [9]	Eating Disorder (Anorexia Nervosa)	ELMo, CNN, Attention	Contextual embedding	No social network, Only post contents, Focuses only on disorder existence	eRisk 2019 [146] - Reddit	✓✓
Han et al. [95]	Depression	BERT, HAN	Adopts metaphor generation, Contextual embedding	No social network, Only post contents, Focuses only on disorder existence	by Shen et al. [225] - 𝕏	✓✓

Continued on next page

Study	Disorder	Methodology	Our comments		Dataset	XAI
			Pros	Cons		
Zogan et al. [292]	Depression	GloVe, MLP, HAN	Multi-modality	No social network, No contextual embedding, Focuses only on disorder existence	by Shen et al. [225] - 𝕏	✓✓
Naseem et al. [167]	Suicide Ideation	TensorGCN, Longformer, Transformer	Ordinal classification, Contextualised embedding, Performs well on long sequences	No social network, Only post contents	by Gaur et al. [81] - Reddit	✓✓

In summary, DL models, as illustrated in Table 2.3, have demonstrated promising outcomes for MDD. However, it is noteworthy that some of the datasets used in these studies, particularly those achieving high accuracy, contain anchor posts that predominantly contain explicit mentions and evidence of mental health conditions (e.g., 'I was diagnosed with depression'). While such posts offer clear evidence, they are typically limited in number. This reliance on explicit self-declarations rather than detecting latent behavioural patterns may not offer a robust approach to MDD, especially for early intervention. To advance holistic methodologies, it is important to account for various factors such as historical posts and user interactions that might enhance latent indicators of mental health challenges.

Existing methods that encompass these features tend to be complex and less explainable. Striking the right balance is important, particularly in applications like healthcare, where explainability plays a crucial role.

2.3.c Importance of domain knowledge

While data-driven methods rely heavily on machine learning for learning complex patterns, additional domain knowledge allows the method to further enrich and enhance the learning process according to the context. Taking proper consideration of domain knowledge offers a transparent view of decision-making processes with minimal reliance on computational methods. Nguyen et al. [170] conducted a study to explore the characteristics of depression among LiveJournal users by employing statistical learning methods. LiveJournal is an OSM platform that provides 132 predefined mood labels such as *depressed*, *happy*, *hungry*, and *cheerful*, and allows users to tag these labels in their posts. The researchers harness the Affective Norms for English Words (ANEW) [30], a lexicon of words with their affective measures in terms of valence, arousal, and dominance, for extracting sentiment and mental health related information. They extract Psycholinguistic features using LIWC, and topics using LDA. These features were found to be significantly different between clinical and control groups. Statistical tests confirmed substantial differences between the two groups. Notably, users from the clinical group were found to use negative emotions more frequently. The study observed an increased risk of clinical cases associated with the use of suicide-related words such as coffin, kill, and bury. Interestingly, the posts in the suicide and depression communities displayed similarities. These findings can serve as a foundation for creating guidelines to detect depressed users on social media platforms.

Our society has a social stigma that impedes the application of traditional methods for MDD. To this end, Perez et al. [192] introduced a computational

approach for estimating depression intensity of an OSM user by automatically responding to the BDI-II questionnaire (See Section 2.2) on users' behalf. The authors leveraged eRISK-2019 [146] and eRISK-2020 [145] datasets, which contain Reddit posts as well as answers to the BDI-II questionnaire for a sample of users. BDI-II has 21 questions, each having four options. For all these questions, a unique representative vector is learned with respect to each available option from the training data, resulting in a total of 84 vectors. The learning process utilised Word2Vec embeddings and averaging operations. To infer severity for a test user, the authors calculate Pearson correlation between text embeddings of the user and the representative option vectors. The option with the highest correlation is then assigned to the related question, thus automating the questionnaire filling process. Once all the questions are answered, they can be aggregated for severity assessment following traditional methods.

SOTA approaches and models for MDD exhibit limited capability for their generalisation and interpretation within the context of clinical settings. In order to achieve this objective, Perez et al.[193] developed a sentence dataset known as BDI-Sen, which is specifically designed to capture the clinical symptoms associated with depression. The dataset utilised in this study is derived from the eRISK-2019 dataset and encompasses a comprehensive range of symptoms associated with depression as assessed by the BDI-II questionnaire. In order to create this dataset, four distinct queries are formulated and the relevance is determined by calculating the cosine similarity of their sentence embeddings. Similarly, Anwar et al. [15] developed a lexicon for EDs called EDBase, comprising a comprehensive collection of ED-related terminology used in OSM and their ED relevance scores. The lexicon enables a domain-relevant interpretation and generalisation of ED-related OSM contents. It is particularly useful for domain-specific feature extraction, which can potentially enrich DL models with domain knowledge [2]. Transfer learning on domain-specific data is another essential component in enhancing detection performance. Within this context, deep embedding approaches pertaining to mental health hold significant importance. *MentalBERT* [113] is a pre-trained BERT model tailored for mental health. It is trained on a dataset of Reddit posts related to mental health, making it a valuable resource for analysing mental health related texts.

2.4 Explainability is what we need

Though modern DL models achieve remarkable performance, their architectures are often as complex as they are effective. Explainability is the key to unlocking the black box nature of a model, shedding light on the model's internal decision-

making processes. Transparency and trustworthiness can be a matter of life and death in mental healthcare. Since only mental health practitioners can verify and judge the decisions of AI systems regarding a user's mental state, we explicitly consider them the primary stakeholders of explainability. There exist two different types of explainability [18, 114].

- *Local explainability* focuses on clarifying the rationale behind a single prediction made by the model.
- *Global explainability*, in contrast, provides insights into the entire decision-making process, with a holistic view of arriving at conclusions.

Some machine learning models are inherently able to explain themselves, referred to as *self-explainable* or *directly interpretable* [18, 114]. Decision trees, for instance, offer transparent decision paths that can be easily understood by humans. However, the scenario changes when it comes to deep learning models. Their intricate neural networks do not readily provide a comprehension. Instead, they necessitate a process known as *post-hoc explainability* [18, 114]. In this post-processing phase, additional techniques are employed to investigate and elucidate the model's decisions.

In terms of their applicability, the explainability techniques can be categorised into two approaches [279, 60].

- *Model-agnostic* explainability techniques are versatile. They can be applied to unravel the mysteries of any machine learning model, regardless of its specific architecture.
- *Model-specific* explainability techniques, on the other hand, are customised to work with particular models, providing insights into their internal processes.

In the quest for explainability and transparency, researchers explore both approaches and employ diverse techniques to decipher advanced DL models. There is a fundamental trade-off between a model's predictive performance and its explainability. Models that achieve SOTA performance typically rely on highly complex architectures—such as multi-head attention—which results in an opaque, black-box nature. In contrast, interpretable (intrinsically explainable) models often exhibit lower predictive performance due to their simpler structural design.

Existing post-hoc techniques, such as LIME and SHAP, attempt to address this challenge by approximating feature importance without revealing the underlying architectural complexity. However, it remains unclear whether these general-purpose approximations provide faithful and reliable explanations. Therefore,

this study uses these SOTA techniques as an initial point of reference, evaluating their effectiveness before determining whether novel, domain-specific architectures are necessary to achieve the level of transparency required by practitioners.

Current research on MDD lacks sufficient attention to the crucial aspects of model explainability and interpretability, limiting the real-world applicability of these models. The complex and diverse nature of mental disorders poses a significant challenge in developing comprehensive explanatory models. Some existing studies use self-attention [9], hierarchical attention [95, 292] and multi-head attention [167] models for generating post-hoc explainability. We aim to bridge the existing gap between the black-box nature of current models and the demand for transparent mental health assessments. Through a comprehensive exploration of SOTA explainability techniques, we seek to advance research on XAI for MDD, providing practitioners and end-users with valuable insights into how MDD models arrive at their predictions. This, in turn, will empower informed decision-making processes and establish a foundation for responsible implementation in real-world scenarios.

Table 2.4: Overview of Explainability methods

Explanation Type		**Scope**	**Model Type**	**Example**	**Applied for MDD?**
Interpretable Models		Global	Model Specific	Linear and Logistic Regression, Decision Trees	✓
Post-hoc	Feature relevance	Local	Model Agnostic	SHAP, KernelSHAP	×
		Local	Model Specific	Layer-wise Relevance Propagation (LRP)	×
		Local	Model Specific	A-Grad and RePAGrad	×
	By Approximation	Local	Model Agnostic	LIME	✓
		Global	Model Agnostic	PGExplainer (for graphs)	×
	By Example	Local	Model Specific	GNNExplainer and GCN-SE (for graphs)	×
		Local	Model Specific	Attention	✓
	By Vis. Explanation	Local	Model Agnostic	ICE	×
		Global	Model Agnostic	PDP	×

2.4.a Interpretable models

Some models, such as linear regression, logistic regression, and decision trees, stand out for their inherent interpretability. Their internal architectures make it easy to understand the rationale behind the models' decisions. Linear Regression and Logistic Regression provide transparency through the generation of feature weights. These weights assign significance to individual features, enabling us to recognise their importance in influencing predictions. Decision Trees, on the other hand, adopt a hierarchical structure of questions, forming a tree-like visualisation. This representation offers an intuitive way to grasp the decision logic. By following the branches of the tree, we can trace the series of questions and conditions that lead to a particular outcome. Inherently interpretable models have been applied in various mental health contexts. For instance, Kelly et al. [123] proposed an interpretable model for predicting mental health treatment outcomes, combining probabilistic methods with a self-interpretable logistic regression approach. Their dataset included patient responses from standardised questionnaires such as the PHQ-9 (depression), GAD-7 (anxiety), Social Interaction Anxiety Scale (SIAS) (social anxiety), Panic Disorder Severity Scale (PDSS) (panic disorder), and the Fear Questionnaire (FQ), along with demographic and medical history data. Similarly, Liu et al. [138] developed an interpretable machine learning model to screen for depression in functionally disabled older adults in China. They analysed data from the 2020 China Health and Retirement Longitudinal Study (CHARLS) database, identifying key depression predictors using Least Absolute Shrinkage and Selection Operator (LASSO) regression and logistic regression techniques. While both studies contribute to explainable mental health analysis, their approaches have limitations. First, the required data (clinical questionnaires, medical records) are not as easily accessible as public sources like social media. Second, these methods rely solely on statistical correlations, ignoring semantic and sentiment-based insights. As a result, their explanations highlight feature importance statistically rather than capturing deeper mental health context. Finally, the models' simplicity restricts their ability to analyse complex, high-dimensional features.

2.4.b LIME and SHAP

LIME [204] is an acronym for local interpretable model-agnostic explanations. It is a widely-used model-agnostic method for providing local explanations. Its primary aim is to explain complex ML models by approximating their behaviour with a simpler and more interpretable model, such as linear regression or logistic regression. The local neighbourhood is created by generating a dataset through

a perturbation method, which introduces controlled noise to the input data. Mathematically, given a prediction function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and a data point x, LIME seeks a simpler model $g(x)$ within a local neighbourhood of x. This is done by minimising the loss between the original complex model and the simpler model using Equation 2.6, where $\mathcal{L}(f, g, \pi_x)$ represents the loss function shown in Equation 2.7, and $\Omega(g)$ is a regularisation term applied to the simpler model g. The goal is to find the parameters for this simpler model that best approximate the behaviour of the complex model within the local neighbourhood. In Equation 2.7, $\mathcal{Z}$ denotes the set of points within the local neighbourhood of x, $\pi_x(z)$ denotes the weight assigned to data point $z \in \mathcal{Z}$ determined using a perturbation method, and $f(z)$ and $g(z')$ are the predictions made by the complex and simpler models, respectively.

$$\xi(x) = \arg\min_{g \in G} \left[\mathcal{L}(f, g, \pi_x) + \Omega(g)\right] \tag{2.6}$$

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \cdot \left(f(z) - g(z')\right)^2 \tag{2.7}$$

LIME has proven valuable in various applications, including causal analysis of depression cases [219]. The study used LIME to compare model-generated explanations with human-generated explanations. Figure 2.4 illustrates an example of the explanation generated by the LIME model. Overall, explanations are often represented as important features of the social media posts.

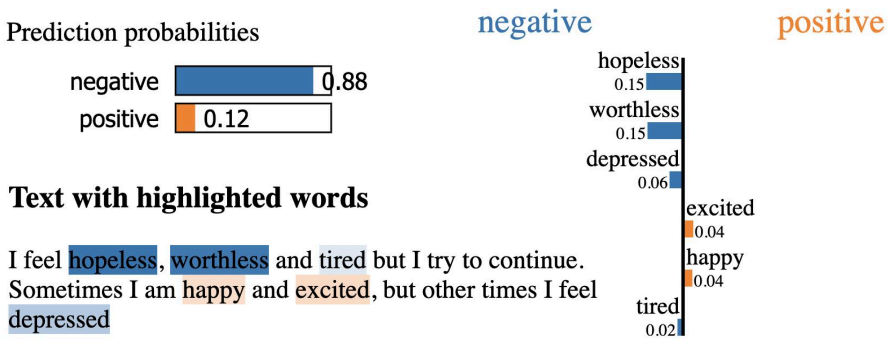


Figure 2.4: LIME explainability example (not-real)

SHAP [147], an acronym for SHapley Additive exPlanations, can be used to obtain the importance (called Shapley value) of individual features used in predictive models. The Shapley value $\phi_i(f, x)$ of a specific feature i of a data point x for a complex model f is calculated using Equation 2.8, where x' denotes the simplified input set (often a subset of the complete set of features), M is the

total number of available features, $|z|!$ calculates the factorial of the size of subset z, $f_x(z')$ denotes the model's prediction for z' subset given as input, and $f_x(z' \backslash i)$ denotes the model's prediction after removing feature i from the subset z'.

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z|!(M - |z'| - 1)!}{M!} \left[f_x(z') - f_x(z' \backslash i) \right] \tag{2.8}$$

Shapley values tend to perform well with linear models, but may encounter limitations with complex models. An extension of SHAP, known as KernelSHAP, is tailored to handle complex models by breaking down Shapley values into smaller subsets of features and employing Monte Carlo sampling techniques to approximate these values.

LIME and SHAP based post-hoc explainability methods have been frequently applied in the analysis of the various mental health issues. For instance, Kerz et al. [124], Saxena et al. [219], and Alghawazzi et al. [6] developed machine learning models to analyse mental disorders using social media data, employing LIME and SHAP to explain their predictions. Alghawazzi et al. [6] proposed an explainable model for detecting suicidal ideation in social media text using an ensemble learning approach. They applied TF-IDF vectorisation to extract text features and used LIME to highlight important words in their predictions. Similarly, Kerz et al. [124] studied mental health challenges, including Attention Deficit Hyperactivity Disorder (ADHD), anxiety, bipolar disorder, depression, and stress, by analysing linguistic and lexical features (e.g., readability, stylistic patterns, ANEW, and LIWC metrics). These features were then processed using a BiLSTM model, with LIME and SHAP providing explanations. However, both Kerz et al. [124] and Alghawazzi et al. [6] rely on statistical existence of features, meaning LIME and SHAP only highlight important statistical features without capturing deeper semantic or sentiment-based insights. Saxena et al. [219] also used a BiLSTM model, but with pre-trained FastText word embeddings, to predict mental disorder causes from social media text. While LIME identifies key tokens in individual posts, its explanations remain limited. It provides local, post-specific approximations rather than a comprehensive understanding of the model's decision-making process.

2.4.c Explanation by Attention values

Attention mechanisms [21, 252] are widely used to determine feature importance in NLP tasks by generating word-level importance scores, contributing to model explainability [263]. Several studies [167, 9] have leveraged attention mechanisms to detect suicide risk from Reddit data while providing explanations based on attention values. In these models, important words in social

media posts that indicate suicidal ideation are highlighted, as they play a crucial role in suicide detection.

However, attention values are not limited to word-level importance. They can also be used to assess the significance of entire sentences through hierarchical attention networks (HAN) [274]. In this approach, word-level attention values are first computed, followed by sentence-level attention values. For example, MDHAN [292] detects depressed users while providing model explanations using HAN. It encodes user posts at both the tweet and word levels, offering explanations at multiple granularities. While MDHAN is effective for both classification and explainability, it focuses solely on textual features and does not consider user behaviour patterns. Similarly, [95] applied hierarchical attention mechanisms to explain depression detection on Twitter. However, their approach incorporates metaphor mapping, but the role of metaphors in explainability is not well justified. Additionally, attention-based explanations (word- and tweet-level) do not clarify whether a word or tweet has a positive or negative influence on model decisions.

To address this limitation, [139] introduced the Attention Gradient (AGrad) mechanism, which measures whether a contextual embedding contributes positively or negatively to a model's decision. [124] further analysed the impact of linguistic and lexical features (e.g., readability, stylistic patterns, ANEW, and LIWC metrics) using AGrad to study mental health challenges such as ADHD, anxiety, bipolar disorder, depression, and stress.

Attention mechanisms can be helpful in interpreting the inner workings of transformer-based black-box models like BERT and Generative Pre-trained Transformer (GPT) [104, 5, 255]. However, their effectiveness in explainability remains debatable, as attention values do not always provide a reliable understanding of model decisions. Moreover, attention mechanisms can produce unstable outcomes that are sensitive to outliers, requiring further analysis and training to improve robustness [107].

2.4.d LRP, PDP, and ICE

LRP [20] is an explainability method used to analyse the feature contributions of complex models. It works by propagating relevance scores backward through the layers of a neural network model, determining the importance of each neuron all the way back to the input features. Let's consider two consecutive layers, i and j. The propagation of relevance score, R_j, can be calculated as $R_j = \sum_k \frac{\theta_{jk} a_{jk}}{\sum_j a_{jk}} R_k$ where a_{jk} is the activation vector of neuron j, and θ_{jk} is the weight vector of the connection from neuron j to neuron k. R_j quantifies the importance of neuron j in the model's predictions, and it simply aggregates the relevance scores R_k from

the subsequent layer. This process allows us to trace back and understand the contribution of each neuron and input feature in the model's decision-making.

PDP visualises the marginal effect of model features using a partial dependence method [76]. Based on this, Greenwell et al. [87] introduced *importance measure* (IM) to compute the importance of a feature. For continuous variable x_i in a model $f()$, $IM(x_i)$ is calculated using Equation 2.9, where k is the number of unique instances within the feature. For categorical variable x_i, $IM(x_i)$ is computed using Equation 2.10.

$$IM(x_i) = \sqrt{\frac{1}{k-1}\sum_{j=1}^{k}[f_i(x_{ij}) - \frac{1}{k}\sum_{j=1}^{k} f_i(x_{ij})]^2} \tag{2.9}$$

$$IM(x_i) = \frac{\max_j(f_i(x_{ij})) - \min_j(f_i(x_{ij}))}{4} \tag{2.10}$$

While PDP illustrates the average effect of a feature, it does not focus on the prediction changes based on individual instances. This limitation is addressed by ICE plots [85]. They depict one line for each instance and each feature by manipulating the feature of interest while keeping the other features fixed.

Although LRP, PDP, and ICE are not commonly used in the analysis of mental health challenges and social media contexts, they still have potential for explaining these analytical models. They may be especially useful for analysing the effects of social, behavioural, emotional, and sentimental features, which could provide insights into model decisions.

2.4.e Explainability of GNNs

While GNNs are able to deliver great results for mental health analytics, achieving explainability in the context of MDD remains a challenge due to the inherent complexity of these models. While ongoing research is striving to enhance the explainability of GNNs, it is still an open problem [279]. Existing explainability techniques, including LIME, attention mechanisms, and LRP, have been applied to elucidate the decision-making processes of GNNs. Given the complex structure of GNNs, some non-traditional explainability methods have also been developed. Recent noteworthy contributions to GNN explainability include GNNExplainer [278], PGExplainer [148], GCN-SE [72].

GNNExplainer [278] is a powerful model-agnostic technique for explaining GNN predictions across diverse graph-related machine learning tasks, including node classification, link prediction, and graph classification. It furnishes explanations in the form of a concise subgraph of the input graph and a subset of node features that exercise the most significant influence on GNN predictions.

GNNExplainer is adaptable to both single-instance and multi-instance scenarios. In single-instance scenarios, it elucidates a GNN's prediction for a specific instance, be it a node label, link, or graph-level label. For multi-instance scenarios, it provides a coherent explanation covering a set of instances, such as nodes belonging to a specific class. Given a trained GNN model Φ and its predicted label distribution Y, GNNExplainer seeks to identify a subgraph $G_S \subseteq G_c$ (computation graph) and the associated node features $X_S = \{x_j | v_j \in G_S\}$ that maximise mutual information with the GNN's prediction, as illustrated in Equations 2.11 and 2.12. Here, $H(Y)$ represents the entropy of the predicted label distribution Y, and $H(Y|G_S, X_S)$ is the conditional entropy. Equation 2.11 strives to minimise uncertainty when the GNN is confined to the explanation subgraph G_S. Equation 2.12 quantifies how much information about the GNN's prediction is present in the explanation subgraph and node features, ensuring that the identified subgraph G_S maximises the probability of the GNN's prediction $\hat{y}$.

$$\max_{G_S} MI(Y, (G_S, X_S)) = H(Y) - H(Y|G_S, X_S) \tag{2.11}$$

$$H(Y|G_S, X_S) = -E_{Y|G_S, X_S}[\log P_\Phi(Y|G_S, X_S)] \tag{2.12}$$

Although GNNExplainer is effective, it suffers from several limitations that impact its applicability in the real-world. *Firstly*, the method primarily focuses on providing local interpretability by generating customised explanations for individual instances, like nodes or graphs, independently. This limits its effectiveness in the inductive setting, as the explanations can not be generalised to other unexplained nodes. *Secondly*, GNNExplainer requires retraining for every single explanation, making it time-consuming and impractical when dealing with a large number of nodes. *Lastly*, as GNNExplainer was designed for interpreting individual instances, the explanatory motifs are not learned end-to-end with a holistic view of the entire GNN model. This lack of a comprehensive global perspective can lead to suboptimal generalisation performance.

PGExplainer [148] recognises the importance of collective and inductive explanations for GNN predictions, and addresses these limitations of GNNExplainer. Though PGExplainer works in a similar fashion as GNNExplainer, maximising the mutual information between the original input graph and the subgraph, PGExplainer offers explanations that collectively cover multiple instances and provide a more global understanding. It leverages a parameterised generative model for graph data, and uncovers underlying structures crucial for GNN predictions. This generation process with a DNN enables collective explanations with shared parameters. Its improved generalisation enhances its practicality, allowing it to infer explanations for unexplained nodes in an inductive setting

without the need for retraining. Both GNNExplainer and PGExplainer are able to explain GNN predictions when the graphs are static. However, they are unable to explain the decision-making process of dynamic GNNs.

GCN-SE [72] addresses the explainability in dynamic GNNs with the help of attention values. It treats graph snapshots at different times of dynamic graphs as different channels of data and attaches a set of learnable attention weights with them based on GCN [127] and SE-Net [106].

However, graph-based explainability methods have not yet been applied to mental disorder analytics, but they hold significant promise. Given that an individual's social circle has a strong influence on mental health, GNN-based explainability methods could provide insights into the sources of mental disorders within a community. Additionally, dynamic approaches like GCN-SE can be useful for analysing the temporal evolution of mental disorders in users.

2.4.f Large Language Model (LLM)-based explanations

Although LLMs are complex and often work like black boxes, they show strong potential for explaining mental health problems in social media content [271]. One example is MentaLLaMA [272], a fine-tuned version of the LLaMA model. It has been used to detect mental health issues like depression and stress, and to identify their possible causes. This model can also generate reasoning behind its predictions, which can be seen as a form of explainability. However, the internal decision-making process of LLMs is still not well understood, so they remain mostly opaque. LLM explainability has become an important research topic due to concerns around ethics, safety, and trust. Many works use prompt-based probing methods to test how these models respond, but such techniques, like those used in MentaLLaMA, do not reveal how the models actually work inside [119, 22, 63, 277, 266, 259]. Attention-based methods offer another option, but the large number of stacked multi-head attention models in LLMs makes this approach difficult to apply and interpret. Some researchers also use probing methods, where LLMs are kept frozen and extra trainable layers are added [119]. These techniques can give hints about how certain parts of the model behave. However, we believe these methods only estimate what the model is doing and do not fully explain how it works. They may find patterns between model behaviour and language features, but they still fall short of fully opening the black box.

2.5 Evaluation Methods and Datasets

2.5.a Evaluation measures

Evaluating an ML model on experimental datasets using specific evaluation metrics is important in order to assess its prediction performance. For MDD models, standard metrics like *Accuracy*, *Precision*, *Recall*, *F-Score*, and *Receiver Operating Characteristic Area Under the Curve (ROC-AUC)* score are commonly employed. Moreover, there also exist problem-specific and domain-specific metrics tailored to consider relevant additional factors. Definition 2.5.1 is applied to evaluate the performance of a model working on early prediction, Definitions 2.5.2-2.5.5 are applied to evaluate models that predict answers of MDD questionnaires, and Definition 2.5.6 is applied to evaluate models that perform a multi-class classification of mental disorder severity.

Definition 2.5.1 (Early Risk Detection Error (ERDE))**.** *This measure introduces a penalty for late correct predictions by considering the number of posts seen before the mental disorder alert is issued by the model [143].*

A lower ERDE value means a better model performance for early detection. For each instance, it is calculated using Equation 2.13, where k denotes the delay (number of posts seen before decision), o, c_{fp}, c_{fn} and c_{tp} are external parameters, and $lc_o(k) = 1 - \frac{1}{1+exp(k-o)}$. The overall $ERDE$ measure for an MDD model is calculated by averaging the measures of all the users.

$$ERDE_o = \begin{cases} c_{fp} & \text{for False Positives (FP);} \\ c_{fn} & \text{for False Negatives (FN);} \\ lc_o(k) \times c_{tp} & \text{for True Positives (TP);} \\ 0 & \text{for True Negatives (TN);} \end{cases} \tag{2.13}$$

Definition 2.5.2 (Average Hit Rate (AHR))**.** *Hit rate (HR) computes the ratio of items the model has estimated the same answer option as the user, and AHR is the average of HR of all users [192].*

Definition 2.5.3 (Average Closeness Rate (ACR))**.** *Closeness rate (CR) is computed as CR $= mad - ad$, where mad and ad represent the maximum absolute difference and absolute difference, respectively [192]. ACR is then computed as the average of closeness rates (CR) of all users.*

Definition 2.5.4 (Depression Category Hit Rate (DCHR))**.** *It computes the fraction of users where the model prediction and ground truth score of the user fall in the same severity level according to BDI-II [192]. Although DCHR is used only for depression, it can be generally applied to all disorders with questionnaires.*

Definition 2.5.5 (Average Difference of Overall Depression Levels (ADODL))**.** *The Difference between Overall Depression Level (DODL) is computed as* $DODL = \frac{63-ad}{63}$*, where* ad *is the absolute difference between model prediction and ground truth value, and 63 is the maximum possible absolute difference* ($21 \times 3 = 63$*, for a total of* 21 *questions, each with a maximum possible difference of* 3 *points) [192]. The ADODL (Average DODL) is computed as the average of DODL of all users. ADODL can also be applied to other disorders with questionnaires.*

Definition 2.5.6 (Ordinal Regression)**.** *For multi-class classification, the classes are assumed to be independent. However, the classes may have an order sometimes. For example, in case of a mental disorder severity classification with classes - severe, moderate, mild, and minimum - severe is closer to moderate than mild in terms of intensity. To model this ordinal relationship, Sawhney et al. [215] introduced the idea of ordinal regression (also known as ordinal classification) that captures dependencies between the severity classes. Let* $Y = \{r_i\}_{i=0}^{n}$ *represent the* n *ordinal severity levels. For given actual severity levels* $r_t \in Y$*, soft labels are calculated as probability distributions,* $y = [y_0, y_1, ..., y_n]$*. The probability* y_i *of each severity level* r_i *is determined using Equation 2.14, where* $\phi(r_t r_i)$ *is a cost function that penalises the difference between the actual severity level* r_t *and the predicted severity level* $r_i \in y$*. As the difference between* r_t *and* r_i *increases, there is a decrease in the probability* y_i *of the associated prediction* r_i*.*

$$y_i = \frac{exp^{-\phi(r_t r_i)}}{\sum_{k=1}^{\lambda} exp^{-\phi(r_t r_k)}} \tag{2.14}$$

$$FN = \frac{\sum_{i=1}^{N_T} I(k_i^a > k_i^p)}{N_T}, FP = \frac{\sum_{i=1}^{N_T} I(k_i^a < k_i^p)}{N_T} \tag{2.15}$$

To evaluate the effectiveness of a severity estimation model, the measures of False Negative (FN) and False Positive (FP) are modified using Equation 2.15, where N_T is the size of the test data, k^a is the actual severity level, and k^p is the predicted severity level over the test data.

2.5.b Experimental datasets

Datasets serve as fundamental components for data-driven decision-making systems. The dataset curated by Shen et al. [225] from 𝕏 stands out as a frequently utilised resource in research on depression detection. Another notable data source is the eRISK series [143], collected from Reddit, extensively used by researchers for detecting various mental disorders, including depression, EDs, self-harm, and pathological gambling. Dreaddit, compiled by Turcan et al. [249], is a Reddit-based dataset designed for binary depression classification, later enhanced by Naseem et al. [166] for multi-class severity classification. PsycheNet

[195] is a social-contagion-driven dataset constructed from 𝕏, primarily for depression detection. Mihov et al. [156] improved this dataset, creating PsycheNet-G, by incorporating additional features like bidirectional replies, mentions, and quote-tweets to enhance the robustness of social network data. Further details about these and other relevant datasets are available in Table 2.5.

Table 2.5: Notable experimental datasets (✓: publicly available, ⊙: available on request).

Dataset	Level	Avl?	Platform	Disorder	Classes	Statistics	Period
[225]	User	✓	𝕏	Depression	Binary	Depressed Users: 1,402, Non-Depressed Users: $>$ 300 million, Depressed Tweets : 292,564, Non Depressed Tweets $>$ 10 billion	2009-2016
[249]	Post	✓	Reddit	Depression	Binary	Depressed Posts: 1,857, Non-Depressed Posts: 1,698	2017-2018
[166]	Post	✓	Reddit	Depression severity	Multi-class	Minimum depression level: 2,587, Mild depression level: 290, Moderate depression level: 394, Severe depression level: 282	2017-2018
[145]	User	⊙	Reddit	Self-harm and depression	Binary and Multi-class	Submissions for Self-harm: 18,618, Submissions for non-self harm case: 254,642; Depression statistics: Not available, answers of BDI-II questionnaire	2017-2019

Continued on next page

Dataset	Level	Avl?	Platform	Disorder	Classes	Statistics	Period
[182]	User	⊙	Reddit	Pathological Gambling, Self-harm and Depression	Binary and Multi-class	Gambling submissions: 54,674, Non-Gambling submissions: 1,073,883; Self-harm submissions: 69,722, Non self-harm submissions: 943,465; Depression statistics: Not available, answers of BDI-II questionnaire	2019-2021
[181]	User	⊙	Reddit	Pathological Gambling, Self-harm and depression	Binary and Multi-class	Gambling submissions: 69,301, Non-Gambling submissions: 2,087,210; Depressed submissions: 35,332, Non depressed submissions: 687,228; Eating Disorders statistics: Not available, answers of EDE-Q questionnaire	2021-2023
[195]	User	⊙	𝕏	Depression	Binary	Depressed users: 372, Non-Depressed users: 445	Not specified

Continued on next page

Dataset	Level	Avl?	Platform	Disorder	Classes	Statistics	Period
[156]	User	⊙	𝕏	Depression	Binary	Depressed users: 242, Non-Depressed users: 349	Not specified
[81]	User	✓	Reddit	Suicide Ideation	Multi-class	500 total users; Attempt: 45; behaviour: 77; Ideation: 171; Indicator: 99; Supportive: 108.	2005 and 2016
[229]	User	✓	𝕏	ADHD, Bipolar, Anxiety, Depression, PTSD, OCD	Binary and Multi-class	43269 tweets; 27003 users.	2017-2021
[96]	Post	✓	Reddit	Suicide and Depression	Binary	1894 total posts, with 915 control and 980 diagnosed positive.	Not specified
[276]	User	⊙	Reddit	Depression	Binary	9210 depressed users and 107274 control users.	2006 and 2016

Continued on next page

Dataset	Level	Avl?	Platform	Disorder	Classes	Statistics	Period
[51]	User	⊙	Reddit	Multiple (inc. Depression, Anxiety, Bipolar, OCD, Schizophrenia, ED)	Binary	Depression: 14,139 users and 1,272 posts; Anxiety: 8,783 users and 795K posts; Bipolar: 6,434 users and 575K posts; EDs: 598 users and 53K posts; OCD: 2,336 users and 203K posts; Schizophrenia: 1,331 users and 123K posts; Control: 335,952 users and 116M posts.	2006 and 2017
[172]	Post	⊙	𝕏	Suicide Ideation	Multi-class	534 safe to ignore posts; 1029 Possibly concerning posts; 258 strongly concerning posts.	2014
[161]	User	⊙	𝕏	Schizophrenia	Binary	174 positively diagnosed users; 3200 posts.	2008 and 2015

Continued on next page

Dataset	Level	Avl?	Platform	Disorder	Classes	Statistics	Period
[55]	User	⊙	𝕏	Multiple (inc. ADHD, Depression, Anxiety, OCD, ED, PTSD, Bipolar, Schizophrenia)	Binary	ADHD: 102 users and 384k posts; Anxiety: 216 users and 1591k posts; Bipolar 188 users and 720k posts; Depression: 393 users and 546k posts; EDs: 238 users and 724k posts; OCD: 100 users and 314k posts; PTSD: 403 users and 1251k posts; Schizophrenia: 172 users and 493k posts.	2008 and 2015
[44]	User	⊙	Tumblr	Recovery from Anorexia	Binary	18,923 users and 55,334 posts.	2008 and 2013
[223]	User	⊙	Reddit	Bipolar Disorder	Binary	3488 positively diagnosed users and 3931 control users.	2005-2018

The quality of the social media data used to study mental disorders via data-driven methods is crucial [158], but often poor, as they are collected and annotated via crowdsourcing and sometimes annotated through automatic or semi-supervised methods, and they mainly suffer from weak and insufficient labels [68]. Careful reannotation processes conducted by clinical psychologist reveal substantial mislabelling in existing datasets and advocate for the need for an accurate and standardised approach to data annotation, reinforcing the need for greater involvement of field experts in the annotation process [158]. Moreover, many datasets often include anchor posts[7] with explicit hints and patterns of mental disorders, potentially leading to data leakage. To address this, there is a need for datasets that contain detailed historical information of users including their posts (excluding the anchor post), online activities, and social network. It is crucial to highlight that the current volume of annotated data is insufficient for robust data modelling in the context of MDD.

2.6 Open issues and challenges

The development of accurate and XAI models for MDD faces several important issues and challenges. These obstacles impede progress in the field and demand careful consideration for the advancement of robust, reliable, and transparent mental health analytics systems.

Dataset. *Issue:* A critical issue in the domain of MDD is the scarcity and limitations of benchmark datasets. There are very few datasets annotated by clinical psychologists or psychiatrists, as a result, many existing datasets fail to comprehensively represent the complexity of mental health conditions [284, 166]. They also lack crucial associated data such as metadata and users' social network information. Furthermore, the diverse spectrum of mental disorders is not captured in these datasets. *Challenge:* Collaborative initiatives with mental health professionals are essential to ensure datasets are not only larger but also carefully annotated with a focus on severity levels and temporal dynamics. The datasets need to include comprehensive user metadata, social network information, and fine-grained severity labels. The limited access to OSM APIs makes data scraping challenging, which necessitates strategies to overcome these limitations.

User-level vs post-level MDD. *Issue:* The ultimate goal of MDD is to identify disorders experienced by users (user-level MDD). However, a significant portion of existing research focuses on a more straightforward approach, detecting disorders expressed within individual OSM posts (post-level MDD) [156,

[7] An anchor post is the post of a user having clear indication of a disorder, based on which the user is labelled with the disorder. For example, "I have been diagnosed with depression".

166, 215]. While post-level MDD provides valuable insights into the mental states conveyed in isolated posts, the user-level MDD represents a more holistic perspective, offering a comprehensive understanding of an individual's mental health condition [2]. *Challenge:* Addressing user-level MDD necessitates a holistic examination of users' historical data. The challenge lies in developing methodologies that seamlessly integrate various types of historical data, including textual content, behavioural patterns, and social interactions, to construct a rich profile of an individual's mental health journey .

Binary classification vs severity estimation. *Issue:* The predominant focus on binary classification (indicative of a disorder or not) of a user or a post in existing studies on MDD limits the granularity of gained insights [225, 131]. It lacks the diverse spectrum of mental health conditions and their varying degrees of severity. *Challenge:* The primary challenge lies in transitioning from binary classification towards more sophisticated severity estimation models. While some recent studies have ventured into severity detection [166, 83], they remain limited in scope and lack comprehensive coverage across various mental disorders. Effectively addressing this challenge requires leveraging diverse data sources, establishing standardised severity scales for OSM extending the traditional severity scales, and the development of robust severity estimation models capable of discerning subtle variations in severity across different disorders. Most importantly, standard severity scales such as BDI-II [25] rely on the intensity and the accumulation of the symptoms of the mental disorders. However, typical social media posts rarely express more than one core DSM-5 symptom at a time [164]. Therefore, a single social media post generally does not provide sufficient clinical evidence for severity estimation. More reliable estimation requires information regarding the post history.

Multimodality and social graph. *Issue:* Most existing studies exclusively rely on the textual content of user posts within their machine learning models for MDD [166, 116, 221, 95]. While text-based approaches offer valuable insights, they fall short of capturing the richness of multimodal data prevalent in OSM, such as behavioural patterns, contextual information, and the social network. They neglect a comprehensive understanding of mental health expressions in the online space. *Challenge:* The challenge lies in effectively capturing and representing multimodal and social graph data within ML models for MDD. To adequately address this challenge, future research should explore the integration of behavioural patterns, visual content, and metadata associated with user posts [2, 290]. Additionally, harnessing the power of social network structures with the help of GNNs can provide valuable context and relationships [156, 131].

Volume of OSM data. *Issue:* The volume of OSM data (whether lengthy individ-

ual posts or large number of historical posts) sometimes becomes significantly large. In such cases, reasonable strategies need to be employed for properly utilising them. *Challenge:* Implementing strategies like summarisation and other advanced techniques becomes imperative to condense the data, maintaining a balance between compression and retaining crucial information [169].

Explainability. *Issue:* A critical concern in the current status of MDD models is the pervasive neglect of explainability. A majority of existing studies in this domain overlook the need of ensuring transparency and interpretability in their models [156, 131]. Particularly in the context of healthcare, where decisions based on these models can have acute implications, the lack of emphasis on explainability compromises the trustworthiness of these models [232, 292]. Hence, it limits their acceptance in real-world applications and clinical practices. *Challenge:* The main challenge lies in the dual pursuit of achieving high accuracy while concurrently ensuring explainability. The prevalent use of complex DL architectures, while effective in achieving accuracy, poses a barrier to interpretability. The intricate internal workings of these models make it challenging to uncover how they arrive at specific decisions, rendering them akin to "black boxes" .

Temporal evolution of disorders over time. *Issue:* A critical gap in the current status of research on MDD lies in its static nature. Existing studies generally treat it as a snapshot problem, neglecting the inherent temporal fluctuations that individuals experience [131]. In reality, mental health is a dynamic continuum, where individuals may transition through varying states of well-being and distress over different time intervals. *Challenge:* The challenge lies in developing models that transcend the static paradigm and incorporate temporal information to capture the evolving patterns of mental disorders over time. Addressing this challenge involves exploring innovative approaches to temporal modelling accommodating the transient and dynamic nature of these disorders, possibly with the help of dynamic DNNs.

Simultaneous consideration of multiple disorders. *Issue:* Existing studies predominantly focus on the detection of a single mental disorder [2], neglecting the reality where individuals often experience multiple interrelated disorders simultaneously [116]. This myopic approach fails to capture the complexity of mental health conditions, limiting the practical utility of these models in real-world scenarios. *Challenge:* The challenge lies in developing models that can simultaneously consider multiple interrelated disorders with the help of mental health domain knowledge. The dynamic nature of these disorders over time adds another layer of complexity. The challenge is not merely to expand the scope of detection but to devise models that can adeptly capture the intricate

link between multiple disorders and their evolving manifestations.

Consideration of domain-specific knowledge. *Issue:* The integration of domain-specific knowledge in MDD models is pivotal for enhancing interpretability, contextuality, and transparency in decision-making. Some existing studies have demonstrated the efficacy of incorporating domain knowledge from various sources, such as domain lexicons, psychological features, and clinical questionnaires, to enrich the understanding of mental health states within OSM platforms [232, 15, 193, 192]. However, a significant issue persists in their limited application. *Challenge:* The foremost challenge lies in fostering interdisciplinary collaborations between domain experts such as clinical psychologists and psychiatrists, data scientists and AI researchers to develop robust methodologies. This requires establishing effective communication channels, bridging the gap between mental health professionals and data/AI practitioners, and developing methodologies for extracting relevant domain information and insights. Moreover, creating transparent and interpretable models that leverage domain knowledge poses a unique challenge, especially when dealing with inherently complex DL architectures.

2.7 Summary

Mental health challenges affect millions of people worldwide and demand innovative solutions for early detection and intervention. The advent of deep learning models, such as DepressionNet and EDNet, brings powerful tools to the field, yet their black-box nature raises ethical concerns in healthcare applications. The recognition of this gap has led to the rise of XAI models, aiming to illuminate the decision-making processes of complex AI systems. In this chapter, we explored and presented a summary of the traditional mental disorder diagnostic methods, the SOTA research on data- and XAI-driven mental disorder detection, and the rapidly developing field of XAI. It emphasises the need for XAI models to shed light on complex AI decision-making processes, especially in the context of mental health analytics. The chapter calls for a balanced approach by aligning technological advancements with transparency and ethics. It is crucial to bridge the gap between deep learning efficacy and interpretability, in order to obtain meaningful insights that benefit users and healthcare professionals. The outlined research directions provide a roadmap for future endeavours, emphasising real-time analytics for enhanced mental health AI. Ultimately, the chapter envisions a future where mental health AI not only detects disorders but also contributes positively to societal well-being.

3 Knowledge Infused Prediction of Depression Risk and Severity

3.1 Introduction

Depression affects over 280 million people globally, with severe outcomes, including approximately 700,000 suicides annually [261]. Despite available treatments, barriers such as stigma and limited access to healthcare treatments leave over 70% of affected individuals untreated [175]. The COVID-19 pandemic has intensified this crisis, highlighting the urgent need for effective and scalable methods for early symptom identification [224]. Social media platforms such as Facebook, 𝕏, and Reddit provide a rich source of user-generated content reflecting mental states, offering opportunities for automated depression risk identification [83, 16]. Traditional interview- and questionnaire-based approaches, while informative, are resource-intensive and may lack scalability [184, 185]. Recent research has explored social media-based models for depression detection [109, 108]. For example, Shen et al. [225] analysed a labelled 𝕏 dataset for multimodal depression detection, while Sampath and Durairaj [210] employed Reddit data for severity-based classification using machine learning. Advanced deep learning approaches, such as the attention-based model by Naseem et al. [166] and DepressionNet [290], which integrates text summarisation, further improve detection performance.

Despite progress, two major gaps remain. First, most studies formulate depression detection as a binary classification problem [225, 209, 290], overlooking clinically relevant severity levels [109]. Some recent models attempt regression or multi-class classification, but they often lack integration of domain-specific knowledge, limiting their clinical interpretability [166]. Second, the opacity of deep learning models complicates their adoption in healthcare settings, where

transparency and explainability are essential for building trust among clinicians and users [109].

To address these gaps, we propose AttentionDep, a domain-aware attention model for explainable prediction of depression risk and intensity (severity) from social media posts. It is mandatory to mention that our primary goal is the risk and severity prediction, not diagnosis or final clinical assessment. AttentionDep follows a structured, multi-step approach. First, contextual representation learning encodes posts hierarchically using unigrams and bigrams. Attention mechanisms highlight clinically salient tokens, while cross-attention integrates domain knowledge from a Wikipedia-derived knowledge graph to enhance feature relevance. The knowledge graph captures clinical relations and provides embeddings that enrich the contextual representations. Finally, in the depression severity estimation stage, the infused representation is used to predict severity levels through an ordinal regression framework. It respects the inherent ordering of severity and improves clinical relevance. AttentionDep is inherently explainable, allowing insight into the decision-making process and providing transparency in predictions. Overall, we make the following contributions in this chapter.

1. We introduce AttentionDep, a domain-aware attention model that integrates contextual text features with domain knowledge for explainable depression severity prediction.
2. We develop a knowledge graph representation framework that captures post-specific clinical relations from Wikipedia, enhancing the semantic depth of input features.
3. We conduct extensive experiments and analyses to demonstrate the model's effectiveness in accurately identifying depression severity levels.
4. We provide explainability mechanisms to interpret model predictions, ensuring transparency and trustworthiness in clinical contexts.

The remainder of this chapter is structured as follows. Section 3.2 covers problem statement, followed by Section 3.3 presents the proposed model AttentionDep, and experimental results are presented in Section 3.4. Finally, Section 3.5 concludes the chapter.

3.2 Problem statement

The objective of this research is to estimate the severity of depression from social media posts in accordance with established clinical standards, specifically the Depressive Disorder Annotation (DDA) scheme [164] and Beck's Depression Inventory (BDI) [25]. Depression severity is defined as an ordered

spectrum of four distinct levels: $C = \{minimum, mild, moderate, severe\}$. Let $P = \{p_1, p_2, \ldots, p_N\}$ denote a set of social media posts. Each post $p_i \in P$, authored by a user $u_i \in U$, is associated with a ground-truth severity label $y_i \in C$. The research problem is thus formulated as learning a function

$$f : P \to C, \quad f(p_i) = y_i,$$

that automatically classifies each post p_i into one of the four ordered severity levels.

3.3 Proposed model

We propose AttentionDep, a domain-aware attention model for explainable prediction of depression severity from social media posts. The model is designed in three main stages. First, user posts are encoded using unigram and bigram representations, with attention mechanisms to highlight clinically relevant tokens. Second, domain knowledge from a curated depression-specific knowledge graph is incorporated to enrich these representations. Cross-attention then integrates contextual and domain-informed features to create a knowledge-aware post representation. Finally, depression severity is predicted through an ordinal regression framework, respecting the natural ordering of severity levels (minimal, mild, moderate, severe). This architecture not only improves predictive accuracy but also provides interpretable insights into which textual and domain-specific features influenced each prediction. The overall model design is illustrated in Figure 3.1. A social media post is first tokenised into unigrams, which are then converted into FastText vectors to obtain unigram embeddings. These embeddings are passed through a Bidirectional LSTM layer to generate contextual representations, followed by a multi-head attention mechanism that yields unigram-level importance scores. This entire processing flow corresponds to the UR (Unigram Representation) component in Figure 3.1. The resulting contextual unigram representations are further fed into a Conv1D layer to form bigram level embeddings. These bigram representations are processed by a multi-head attention layer to estimate bigram level importance scores. This part of the architecture corresponds to the BR (Bigram Representation) component of Figure 3.1. In parallel, the KR (Knowledge Representation) component retrieves a post-specific subgraph from the Mental Health Knowledge Graph (MHKG) and generates knowledge embeddings using a graph neural network. The MHKG construction process is summarised in the KGB section of Figure 3.1. The retrieved knowledge graph embeddings are integrated into the bigram attention mechanism within the BR module through a cross-attention layer. Finally, the fused representation

is aggregated using masked mean pooling and passed through a multilayer perceptron an ordinal regression layer to predict the depression severity level. This final prediction stage corresponds to the OR (Output Representation) pipeline.

3.3.a Context and domain knowledge modelling

AttentionDep models textual representations of social media posts in a way that captures both the semantic meaning and depression-related expressions, while incorporating clinically relevant domain knowledge. The modelling process is hierarchical. It begins with unigram-level encoding, progresses through contextualisation and attention mechanisms that highlight linguistically and clinically salient tokens, and culminates in bigram-level representations enhanced via knowledge graph infusion. Together, these components enable the model to capture both the linguistic context and symptom-related indicators of depression present in user posts.

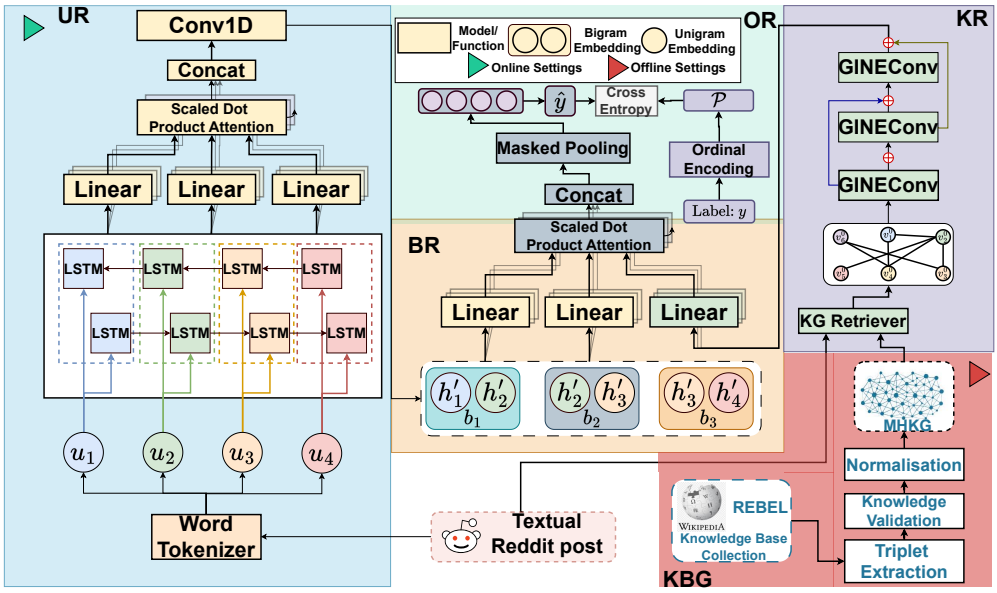


Figure 3.1: Proposed model AttentionDep. It predicts depression severity of a post and generates explanation with unigram- and bigram-level attentions and a knowledge graph.

Contextual unigram encoding

As the first step, social media posts are tokenised into unigrams using the NLTK library[1]. For unigram-level encoding, we adopt FastText [28], a computationally efficient embedding model that integrates character-level n-grams with Continuous Bag of Words (CBOW) and skip-gram training. FastText represents words by composing subword information, making it effective for handling out-of-vocabulary (OOV) and misspelt tokens: $\mathbf{u_w} \leftarrow \frac{1}{|G_w|}\sum_{g\in G_w}\mathbf{z_g}$, where $\mathbf{u_w}$ is the embedding for token w, G_w is the set of its subword n-grams, and $\mathbf{z_g}$ is the embedding of subword g. The ability to model OOV words is particularly critical in our setting, since social media posts about depression often contain informal language, creative spellings, or deliberately obfuscated terms (e.g., to bypass moderation or stigma). We use 300-dimensional FastText embeddings pre-trained on 2 million words with subword information from the Common Crawl corpus (600B tokens)[2]. Hence, for each post p_i, token embeddings are denoted $\mathbf{u}_{i,j} = \text{FastText}(w_{i,j}) \in \mathbb{R}^{300}$ for $j = 1, \ldots, L$ tokens.

Understanding context is particularly crucial in depression-related text, since the same word can signal very different meanings depending on the author's mental state and situational framing. For example, *"I'm **tired** from a productive lecture"* expresses ordinary fatigue, whereas *"I'm so **tired** of living"* conveys potential signs of hopelessness and emotional exhaustion. Although FastText generates a single static embedding for each word, it does not account for such contextual variation. To capture these nuances, we further encode the sequence of tokens using Bidirectional Long Short-Term Memory (BiLSTM) network, a recurrent neural architecture that models the context of each token based on its surrounding tokens in both forward and backward directions. The contextual unigram encodings are thus defined in Equation 3.1, where d is the embedding dimension.

$$\mathbf{H}_i \leftarrow \{\mathbf{h}_{i,j}\}_{j=1}^{L} = \text{BiLSTM}(\{\mathbf{u}_{i,j}\}_{j=1}^{L}) \in \mathbb{R}^{L\times d} \tag{3.1}$$

Attention to clinically salient unigrams

In social media posts, not all words contribute equally to understanding a user's mental health state. Certain tokens (e.g., *"worthless"*, *"empty"*, or *"tired"*) often appear in posts indicative of moderate or severe depression, while many others (e.g., stopwords or neutral terms) carry little diagnostic value. To capture this variability, we incorporate an attention mechanism that learns to assign higher

[1] https://www.nltk.org/api/nltk.tokenize.html

[2] https://fasttext.cc/docs/en/english-vectors.html

weights to words that are more predictive of depression severity labels during training [77]. In this way, the model automatically learns which words are most informative for distinguishing between *minimal*, *mild*, *moderate*, and *severe* classes, without requiring explicit manual annotation of clinical keywords. We adopt a multi-head attention mechanism [252], which computes weighted representations over the sequence of contextualised unigram embeddings. Each attention head captures different patterns of association between tokens and depression severity, thereby improving robustness and interpretability. The attention scores for head k are computed as:

$$\mathbf{A}_i^k \leftarrow \sigma\left(\frac{(\mathbf{H}_i\mathbf{W}_\mathbf{Q}^k)(\mathbf{H}_i\mathbf{W}_\mathbf{K}^k)^T}{\sqrt{d_k}}\right)(\mathbf{H}_i\mathbf{W}_\mathbf{V}^k) \tag{3.2}$$

$$\mathbf{H}_i' \leftarrow \bigoplus_{k=1}^{\kappa_u}\left[\mathbf{A}_i^k\right]\mathbf{W}_O \tag{3.3}$$

where $\mathbf{W}_\mathbf{Q}^k$, $\mathbf{W}_\mathbf{K}^k$, and $\mathbf{W}_\mathbf{V}^k$ are the query, key, and value projection matrices for head k, κ_u is the number of attention heads, σ is the softmax function, and $\mathbf{W}_O$ is a trainable projection matrix. The resulting representation $\mathbf{H}_i'$ captures unigram-level signals that are most predictive of depression severity. Importantly, the learned attention weights provide interpretable insights into which words the model considers depression-relevant, offering a bridge between statistical learning and clinical interpretability.

Enhanced context with bigrams

Individual words in social media posts may not always sufficiently reflect the author's mental health state. However, when combined with other words, they can reveal depression-related linguistic indicators. For example, the words *pretending* and *happy* are neutral or positive individually, but together as *pretending happy*, they convey a negative tone. Similarly, *feel* and *nothing* are vague individually, but their combination suggests loss of pleasure, a common symptom of depression [25]. To capture such patterns, we define bigram embeddings using a one-dimensional convolution operation:

$$\mathbf{b}_{i,j} \leftarrow \text{GELU}\left(\mathbf{W}\cdot\begin{bmatrix}\mathbf{h}_{i,j}'\\ \mathbf{h}_{i,j+1}'\end{bmatrix}+\mathbf{b}\right), \quad j=1,\ldots,L-1 \tag{3.4}$$

where $\mathbf{h}_{i,j}' \in \mathbf{H}_i'$ is the contextual unigram embedding at position j of post p_i, and $\mathbf{W}$ and $\mathbf{b}$ are trainable convolution parameters. The overall bigram

representation of post p_i is then:

$$\mathbf{B}_i \leftarrow \{\mathbf{b}_{i,j}\}_{j=1}^{L-1}. \tag{3.5}$$

Cross-attention for clinical relevance

While unigram embeddings capture individual token information, certain bigrams often convey richer depression-related meaning. To model this, we apply a cross-attention mechanism over bigram representations $\mathbf{B}_i$, guided by domain knowledge from a depression-specific knowledge graph $\mathbf{G}_i$ (detailed later in Section 3.3.b). Unlike standard self-attention, cross-attention allows the model to modulate bigram representations using external domain knowledge. Specifically, queries ($\mathbf{Q}$) are the bigram embeddings $\mathbf{B}_i$, representing the content of the social media post; keys ($\mathbf{K}$) are the knowledge graph embeddings $\mathbf{G}_i$, representing clinically relevant concepts; and values ($\mathbf{V}$) are the bigram embeddings $\mathbf{B}_i$, which are weighted according to their relevance to the knowledge graph. The cross-attention with multiple heads is computed as:

$$\mathbf{\Omega}_i^k \leftarrow \sigma\left(\frac{(\mathbf{B}_i\mathbf{\Theta}_{\mathbf{Q}}^k)(\mathbf{G}_i\mathbf{\Theta}_{\mathbf{K}}^k)^T}{\sqrt{d_k}}\right)(\mathbf{B}_i\mathbf{\Theta}_{\mathbf{V}}^k) \tag{3.6}$$

$$\mathbf{B}_i' \leftarrow \bigoplus_{k=1}^{\kappa_b}\left[\mathbf{\Omega}_i^k\right]\mathbf{\Theta}_O \tag{3.7}$$

where $\mathbf{\Theta}_{\mathbf{Q}}^k$, $\mathbf{\Theta}_{\mathbf{K}}^k$, $\mathbf{\Theta}_{\mathbf{V}}^k$, and $\mathbf{\Theta}_O$ are trainable parameters, and κ_b is the number of attention heads.

By attending to domain knowledge, this mechanism produces bigram representations $\mathbf{B}_i'$ that integrate both data-driven cues and clinically validated depression knowledge, allowing the model to focus on word pairs most indicative of depression severity. The resulting representations are both predictive and interpretable, highlighting clinically relevant phrases in social media posts.

3.3.b Domain knowledge representation

Constructing domain knowledge graph

We construct a Mental Health Knowledge Graph (MHKG), denoted as $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{C}, \mathcal{F}, \mathcal{A})$, to capture contextually appropriate mental health concepts and their relationships. Here, $\mathcal{V}$ represents nodes corresponding to mental health–related entities, $\mathcal{E}$ represents directed edges (relations), and $\mathcal{C}$ contains mental health context sentences. Each edge $e_{s,t} \in \mathcal{E}$ is represented as a triplet $\langle v_s, r_{s,t}, v_t\rangle$, where v_s and v_t are the source and target nodes, and $r_{s,t}$ denotes the relation type. $\mathcal{F} \in \mathbb{R}^{|\mathcal{V}|\times 300}$

denotes the node embeddings of MHKG. Each node embedding $\mathbf{f}_k \in \mathcal{F}$ is computed by applying a pre-trained `FastText` model to the textual attribute of node v_k, i.e., $\mathbf{f}_k = \texttt{FastText}(v_k)$. Similarly, $\mathcal{A} \in \mathbb{R}^{|\mathcal{E}| \times 300}$ represents edge embeddings, where each edge embedding $\mathbf{a}_{s,t} \in \mathcal{A}$ is obtained by applying `FastText` to the textual attribute of the relation $r_{s,t}$, i.e., $\mathbf{a}_{s,t} = \texttt{FastText}(r_{s,t})$. These embeddings allow the model to encode both nodes and relations in a continuous vector space, preserving semantic and contextual alignment for downstream attention and representation learning. Figure 3.2 shows a snapshot of our MHKG, illustrating mental health–related entities and their relationships. Its construction begins with generating a relevant knowledge base, which is followed by its validation, normalisation, and specific MHKG retrieval.

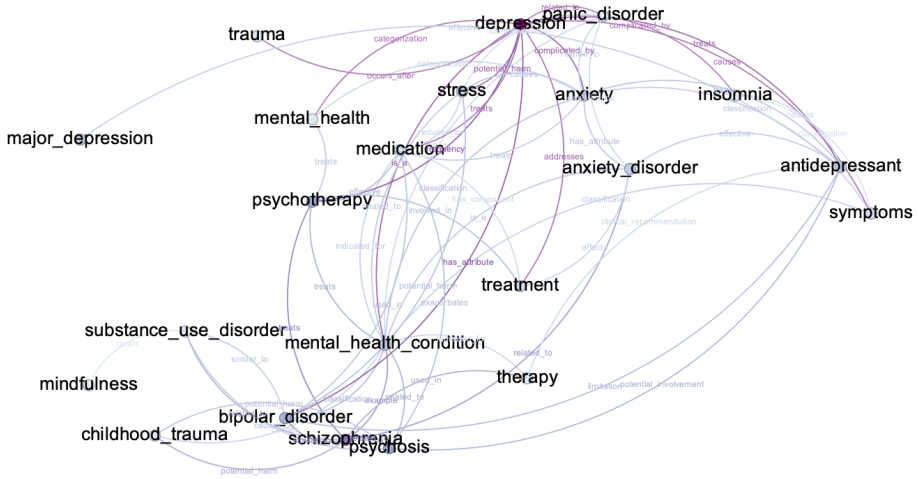


Figure 3.2: Sample snapshot of the MHKG

Generating a knowledge base. In our study, the knowledge base is represented as a set of mental health–related contexts and their associated triplets. We first examine the REBEL dataset[3] [36], a benchmark Wikipedia-sourced knowledge base containing over 3M contexts and triplets. However, as a general-purpose resource, REBEL lacks sufficient coverage of mental health–related contexts. To address this gap, we augment the dataset by extracting potential mental health–specific contexts and triplets leveraging Wikipedia and Large Language Model (LLM). For context collection, we employ the Wikipedia Application Programming Interface (API)[4] to scrape mental health–related content. While the DSM-5 could theoretically serve as a more reliable source of knowledge, we prioritised Wikipedia due to the potential authorship violations of the DSM-V,

[3]https://huggingface.co/datasets/Babelscape/rebel-dataset

[4]https://www.mediawiki.org/wiki/API:Main_page

as it is copyrighted and it would restrict the accessibility of our knowledge base. Moreover, processing the DSM-5 requires complex refinement of unstructured formats (e.g., PDF), whereas Wikipedia provides a structured API that supports organised and efficient data extraction.

In total, we collect information from 183 Wikipedia pages, divided into two categories: (i) 30 *core topics*, consisting of names of mental disorders and related clinical protocols, and (ii) 153 *secondary topics*, including subtypes, causes, and symptoms of these disorders. This ensures the contexts remain focused and clinically relevant while reducing sparsity. The full text of the core topic pages and summaries of the secondary topic pages are concatenated to form the final context corpus, consisting of 10,559 sentences. Tables 3.2 and 3.3 list the core and secondary topics, respectively. Next, we apply a few-shot prompting strategy with $\texttt{GPT}-\texttt{4o}-\texttt{mini}$[5], inspired by [281], to extract triplets from the collected contexts. The exact prompt used is shown in Table 3.1. This process yields 10,559 samples, producing a total of 41,676 triplets. Adopting language models for extracting triplets is a scalable approach as it accelerates the processing of unstructured texts, however, there is a risks of hallucination and noise. The potential risks are mitigated in the next phases during normalisation and validation.

Validating clinical relevance. A triplet $\tau = \langle s, r, t \rangle \in \mathcal{E}$ is considered clinically relevant if it contains entities that are specific, meaningful, and clearly related to mental health (e.g., symptoms, disorders, emotions, behaviours, treatments, or psychological mechanisms). The LLM-based filtering and validation is followed by a manual validation step. The purpose of the filtering and validation is to ensure that the posts are related to mental health, for example, distinguishing between administrative titles and actual clinical disorders. Furthermore, the relation r between entities must represent a causal, plausible, and interpretable connection within the mental health domain. To operationalise this, we curate $\mathcal{M}$ as a predefined set of eligible entities derived from Tables 3.2 and 3.3. A triplet τ is labelled clinically relevant if the following condition holds:

$$\max\left(\max_{m\in\mathcal{M}} \varsigma(s,m),\ \max_{m\in\mathcal{M}} \varsigma(t,m)\right) > 0.9\ \wedge\ \phi(\tau) \tag{3.8}$$

where $\varsigma(\cdot,\cdot) \in [0,1]$ denotes a semantic similarity function, and $\phi(\tau) \in \{\text{True}, \text{False}\}$ is a triplet validity function implemented using an LLM (see Table 3.1). This validation process ensures that each triplet is both semantically and contextually aligned with the mental health domain by verifying that the triplets express clear, interpretable connections (e.g., causal, descriptive, or therapeutic) rather

[5]https://platform.openai.com/docs/models/gpt-4o-mini

Table 3.1: LLM Prompt for Semantic Graph Construction

Purpose	**Prompt**
Triplet Extraction	`Your task is to transform the given text into a mental health related semantic graph in the form of a list of triples. The triples must be in the form of [Entity1, Relationship, Entity2]. In your answer, please strictly only include the triples and do not include any explanation or apologies. Keep the entities and relations as simple and short as possible, and do not make them long, if it is not necessary.` `Here are some examples:` `{few_shot_examples}` `Now please extract triplets from the following text.` `Text: {input_text}`
Triplet Validation	`You are given a knowledge triplet in the form [subject, relation, object].` `Determine if this triplet expresses a valid and meaningful relationship in the context of mental health. A valid relation is one that expresses a clear, interpretable, and contextually appropriate connection-such as causal, descriptive, indicative, or therapeutic-between two mental health-relevant concepts. Evaluate the triplet based on:` `The subject and object must both be specific, meaningful, and clearly related to mental health (e.g., symptoms, disorders, emotions, behaviours, treatments, or psychological mechanisms).` `Avoid vague, overly broad, or overly technical terms unless they directly contribute to mental health understanding.` `The relation must express a plausible and interpretable connection within the mental health domain.` `Return only: yes or no.` `Example (invalid due to vague object):` `["models", "explain", "link between altered brain function and schizophrenia"] → no`

than vague associations. Applying this validation step yields 2,481 contexts and 6,433 validated triplets in total.

Table 3.2: Core mental health topics used in the knowledge graph

Mental health	Mental disorder	Psychology	Major depressive disorder
Depression (mood)	Bipolar disorder	Schizophrenia	Anxiety
Anxiety disorder	Personality disorder	DSM-5	Diagnostic and Statistical Manual of Mental Disorders
Classification of mental disorders	Causes of mental disorders	Antidepressant	Psychotherapy
Cognitive behavioral therapy	Psychosis	Post-traumatic stress disorder	Eating disorder
Dysthymia	Panic attack	Suicide	Social psychology
Personality	Generalized anxiety disorder	Borderline personality disorder	Stress
Mood swing	Cognitive psychology		

Human-assisted normalisation. As the validated knowledge base contents are sourced from two different origins (REBEL and Wikipedia), this introduces heterogeneity. In addition, the newly generated triplets using LLMs may deviate in syntax, style, and phrasing, thereby increasing sparsity and inconsistency. For example, relation phrases such as `causes`, `causeTo`, `cause_to`, and `causeto` all share the same meaning but differ syntactically. They can therefore be interpreted as distinct relations. To address this, we perform triplet normalisation with the assistance of human annotators, who are graduate students trained by academic staff with PhDs in psychology. Entities show similar inconsistencies as well. Therefore, we first normalise entities by identifying those with similar meanings using the semantic similarity of their embedding vectors. If entities with high similarity have the same or close meanings, a unique representative entity is selected. The same logic is applied to relation phrases. However, an additional sanity check is required, as lower cardinality of relation sets is generally maintained in benchmark knowledge graphs due to their higher predictability.

For this purpose, we apply agglomerative clustering to the principal component vectors of their FastText embedding vectors, obtained via principal component analysis (PCA). The resulting clusters are then reviewed by human annotators (trained graduate students), and 177 unique normalised relation types are finalised, as reported in Table A.1. The final MHKG generated after normalisation contains 2,461 contexts and 6,371 triplets comprising 177 unique relations and 4,098 unique entities.

Retrieving post-specific MHKG. As the overall MHKG contains a large amount of information, using it in its entirety for machine learning is both computationally expensive and unnecessary. Many triplets may also be irrelevant for a given post. To address this, we retrieve only a post-specific subgraph. For each post $p_i \in \mathcal{P}$, the retrieved list of triplets is defined as:

$$\mathcal{T}_i \leftarrow \bigcup_{c \in \text{TopK}_{\mathcal{C}}[\varsigma(p_i, \mathcal{C})]} \mathcal{T}(c) \tag{3.9}$$

where $\mathcal{C}$ denotes the set of all available context phrases, $\varsigma(\cdot, \cdot)$ is a semantic similarity function, and $\mathcal{T}(c)$ returns the list of triplets associated with context c. For computing semantic similarity, we adopt a pre-trained cross-encoder model from the `sentence_transformers` library. The retrieval process then selects the top K most similar contexts to the input post p_i and collects all associated triplets. As a result, each post is paired with its own specific set of knowledge triplets, which are subsequently used to construct the corresponding MHKG embeddings.

Knowledge graph representation learning

To capture structural properties in our MHKG and extract clinically relevant indicators for depression analysis, we employ residual sequential layers of the Graph Isomorphism Network with Edge features (GINE) [267]. GINE is a graph neural network designed to capture complex structural patterns via edge features alongside node features, allowing it to distinguish graphs even with similar node attributes and generates informative node features based on the neighbourhood nodes and edges by adopting Multi Layer Perceptron (MLP). AttentionDep aggregates information up to n-hop neighbours by stacking n GINE layers, leveraging GINE's strong representational capacity. We process the retrieved MHKG for post p_i, denoted as $\mathcal{G}_i(\mathcal{V}_i, \mathcal{E}_i, \mathcal{C}_i, \mathcal{F}_i, \mathcal{A}_i)$, using n consecutive GINE layers as defined in Equation 3.10. The embedding $\theta_k^{(l)}$ is learned for each node $v_k \in \mathcal{V}$ at layer l, where $\theta_k^{(0)} = \mathbf{f}_k$ is the initial node feature vector, ϵ denotes the relative importance of the target node compared to its neighbours, $\mathcal{N}(v_k)$

Table 3.3: Extended set of secondary mental health topics

Psychiatric hospital	Mental distress	Mental toughness	Mental state
Philosophy of mind	Mental chronometry	Mental mapping	Mental Health Act
Orientation (mental)	Insanity defense	Mental age	Mental disability
Mental therapy	Mini–mental state examination	Positive mental attitude	Mental health inequality
Insanity	Creativity and mental health	Mental Hygiene	Mental lexicon
Mental reservation	Mental health triage	Mental illness in media	Mental health service
Mental health nursing	Narcissistic personality disorder	Mental status examination	Menstruation and mental health
Mental environment	Mental operations	Abortion and mental health	Mental health law
Rethink Mental Illness	Postpartum depression	Telephone phobia	Anxiolytic
Social anxiety	Castration anxiety	Death anxiety	Hypochondriasis
Social anxiety disorder	Anxiety dream	Stimulant psychosis	Substance-induced psychosis
Postpartum psychosis	Tardive psychosis	Caffeine-induced psychosis	Schizoaffective disorder
Folie à deux	Unitary psychosis	Antipsychotic	Brief psychotic disorder
List of antidepressants	Antidepressant discontinuation syndrome	SSRI	TCA
TeCA	SNRI	Atypical antidepressant	Pharmacology of antidepressants
Antidepressants and suicide risk	Hydrazine (antidepressant)	Second-gen antidepressant	Trazodone
Countries by antidepressant use	Fluoxetine	Tachyphylaxis	Mirtazapine
Antidepressants in Japan	Bupropion	Gestalt psychology	Analytical psychology
Shadow (psychology)	Educational psychology	Evolutionary psychology	Positive psychology
Filipino psychology	Association (psychology)	Developmental psychology	International psychology
Physiological psychology	Manipulation (psychology)	Doctor of Psychology	Narrative psychology
Transpersonal psychology	Individual psychology	Psychopathology	Biological psychopathology
Developmental psychopathology	HiTOP	Child psychopathology	RCAP
Development and Psychopathology	Avolition	Diathesis–stress model	MMPI
PANSS	Stress (biology)	Stressor	Psychological stress
Stress management	Stress hormone	Chronic stress	Anorexia nervosa
Anorexia (symptom)	Anorexia mirabilis	Anorexia athletica	Pro-ana
Sexual anorexia	Atypical anorexia nervosa	History of anorexia nervosa	People with anorexia
Deaths from anorexia	Cachexia	Infection-induced anorexia	Bulimia nervosa
Appetite	Disorganized schizophrenia	Sluggish schizophrenia	Childhood schizophrenia
Risk factors of schizophrenia	Evolution of schizophrenia	Religion and schizophrenia	Origin of influencing machine
People with schizophrenia	Anhedonia	Thought disorder	History of schizophrenia
Bipolar I disorder	Bipolar II disorder	People with bipolar disorder	Cyclothymia
Mood stabilizer	Lamotrigine	Sleep in bipolar disorder	Mood disorder
Epigenetics of bipolar disorder	Mood (psychology)	Mood congruence	Mood tracking
Euphoria	Mood management theory	Psychopathy	Big Five traits
Personality psychology	Antisocial personality disorder	Dissociative identity disorder	Schizoid personality disorder
Personality change	Enneagram of Personality	OCPD	Avoidant personality disorder
Histrionic personality disorder			

represents the neighbourhood of node v_k, and $\mathbf{a}_{k,u}$ represents the edge attributes between nodes v_k and v_u.

$$\theta_k^{(l+1)} \leftarrow \theta_k^{(l)} + \text{MLP}\Bigg((1+\epsilon^{(l+1)})\cdot\theta_k^{(l)} + \sum_{v_u \in \mathcal{N}(v_k)} \left(\theta_u^{(l)} + \mathbf{a}_{k,u}\right)\Bigg) \tag{3.10}$$

The final MHKG representation for post p_i, which serves as the domain knowledge in Equation 3.6, is computed as:

$$\mathbf{G}_i \leftarrow \{\theta_v^{(n)}\}_{v=1}^{|\mathcal{V}|}. \tag{3.11}$$

3.3.c Depression severity with ordinal representation

Depression severity classification

After obtaining the domain knowledge–fused representation from Equations 3.6 and 3.7, we proceed to the final step of depression severity estimation. The prediction for post p_i is generated as:

$$\hat{y}_i \leftarrow \sigma\left(\text{MLP}\left(\frac{1}{\|\mathbf{M}_i\|_1}\mathbf{1}^T(\mathbf{M}_i \odot \mathbf{B}_i')\right)\right) \tag{3.12}$$

$$f(p_i) \leftarrow \text{argmax } \hat{y}_i \tag{3.13}$$

where, $\mathbf{M}_i \in \{0,1\}^{L\times 1}$ is a binary masking matrix indicating whether a representation corresponds to a valid token (1) or padding (0). This ensures that padding tokens do not influence the decision-making process. The element-wise multiplication $\mathbf{M}_i \odot \mathbf{B}_i'$ followed by averaging is called masked mean pooling, which aggregates information from only valid tokens, and MLP$(\cdot)$ denotes a feed-forward multilayer perceptron network. Finally, $f(p_i)$ represents the predicted depression severity for post p_i.

Ordinal encoding and backpropagation

Given the ordinal nature of depression severity levels, we adopt ordinal regression inspired by Sawhney et al. [215]. Let $\mathcal{Y} = \{\text{minimum} = 0, \text{mild} = 1, \text{moderate} = 2, \text{severe} = 3\}$ denote the label space. For a post p_i with true severity $y_i \in \mathcal{Y}$, we generate a soft label distribution $\mathcal{P}^i = [\rho_0^i, \rho_1^i, \rho_2^i, \rho_3^i]$ as

follows:

$$\rho_j^i \leftarrow \frac{\exp(-\phi(y_i, y_j))}{\sum_{y_k \in \mathcal{Y}} \exp(-\phi(y_i, y_k))} \tag{3.14}$$

$$\phi(y_i, y_j) \leftarrow \beta |y_i - y_j| \tag{3.15}$$

where, $\phi(y_i, y_j)$ is a cost function capturing the distance between the true severity y_i and each severity level $y_j \in \mathcal{Y}$, and β is a hyperparameter controlling the penalty magnitude for mispredictions. Larger differences between y_i and y_j result in lower probabilities ρ_j^i, reflecting the ordinal structure of the labels. The prediction error for post p_i is measured using cross-entropy loss:

$$\mathcal{L}_i \leftarrow \sum_{j \in \mathcal{Y}} y_j^i \log(\hat{y}_j^i) \tag{3.16}$$

This loss is minimised during training to update the model parameters and learn an ordinal-aware mapping from the input representation to depression severity levels.

3.4 Experiments and Results

3.4.a Datasets

We construct three experimental datasets from Reddit. We employed and reconstructed publicly available datasets [166, 249, 210] that were originally annotated based on Diagnostic and Statistical Manual of Mental Disorders 5 (DSM-5) criteria [19], where they support depression risk and multi-class severity analysis. Although our analysis focuses on Reddit, the embedding and modelling framework is adaptable to other platforms and languages.

To build these datasets, we start from two publicly available Reddit datasets. The first [166] is an augmented version of the Dreaddit dataset [249] refined according to DSM-5 standards. The second [210] dataset is collected via the Reddit API from mental health–related subreddits. While both provide valuable content and labels, they are not directly used in experiments due to limitations in class coverage and balance. We combine and refine posts from these sources to create our primary experimental dataset, $\mathbf{D_4}$, supporting four-class severity classification. As the *'mild'* class is underrepresented (fewer than 300 posts), we apply augmentation techniques to increase its size. We used semantic paraphrasing to generate linguistically diverse variations of the original posts while preserving their severities. For experiments with a reduced-class setting, we derive a three-class subset, $\mathbf{D_3}$, by excluding mild posts. Binary depression

detection is performed using $\mathbf{D_2}$ [196].

Table 3.4: Summary of datasets used for evaluation.

Dataset	Classes	Class distribution	Source
$\mathbf{D_4}$	4	Minimum: 1500, Mild: 580, Moderate: 2000, Severe: 1000	Reddit
$\mathbf{D_3}$	3	Minimum: 1500, Moderate: 2000, Severe: 1000	Reddit
$\mathbf{D_2}$	2	1293 depressed, 548 control	Reddit

3.4.b Experimental Setup

We select optimal hyperparameters based on the highest graded F_1 score using the Optuna framework[6] with the Tree-structured Parzen Estimator (TPE) sampler. The chosen hyperparameters are: learning rates of 9.8×10^{-5} and 5.4×10^{-5} for **D3** and **D4**, respectively; 200 training epochs; 4 and 2 unigram attention heads for **D3** and **D4**; 4 bigram attention heads for both datasets; a dropout rate of 0.3; batch size of 128; severity scale of 4.5; hidden size of 128; 2 Long Short-Term Memory (LSTM) layers; 3 and 1 Graph Neural Network (GNN) layers for **D3** and **D4**; and a maximum input length of 256 tokens. We fine-tune language models using the Huggingface Transformers library and implement[7] all experiments in PyTorch 2.1 with the Adam optimiser. Experiments run on the Viking HPC Cluster at the University of York, equipped with 12,864 CPU cores, 512 GB RAM, and NVIDIA H100 GPUs.

3.4.c Evaluation Metrics

Standard classification metrics such as Precision, Recall, and F_1 score treat classes as independent and ignore the ordinal nature of labels, making them less informative for severity-level prediction. To address this, we employ Graded Precision (GP), Graded Recall (GR), and Graded F_1 (GF) scores [81, 215], which account for the ordinal structure of depression severity levels. These metrics redefine False Negatives (FN) and False Positives (FP) to reflect whether a prediction

[6] https://optuna.org

[7] The code and data will be publicly available upon acceptance.

underestimates or overestimates the true severity:

$$\text{FN} = \frac{1}{N_T}\sum_{i=1}^{N_T}\mathbb{I}(k_a^i > k_p^i), \quad \text{FP} = \frac{1}{N_T}\sum_{i=1}^{N_T}\mathbb{I}(k_p^i > k_a^i) \tag{3.17}$$

where, k_p^i and k_a^i denote the predicted and actual severity levels, respectively. A prediction is considered a false negative if it underestimates the true severity ($k_p < k_a$) and a false positive if it overestimates it ($k_p > k_a$). True positives occur when the predicted and actual severity levels match exactly.

3.4.d Baseline Models

We compare AttentionDep against twelve baselines, including six generic neural architectures and six state-of-the-art (SOTA) depression detection models. For the State-Of-The-Art (SOTA) baselines where official code was unavailable, we implemented the models from scratch in `PyTorch`, following their original architectures. To ensure a fair comparison, we report the maximum score achieved by each model across: i) the original hyperparameters reported in the respective research papers, ii) the optimal hyperparameters of AttentionDep, and iii) supplementary manual tuning in cases where the first two conditions yielded suboptimal performance.
Generic neural baselines.

- **LSTM, Gated Recurrent Unit (GRU), BiLSTM, Bidirectional Gated Recurrent Unit (BiGRU)**: Recurrent networks for sequential data modelling.
- **Convolutional Neural Network (CNN)**: Extracts salient local features through convolution, widely used in NLP tasks.
- **Bidirectional Encoder Representations from Transformers (BERT)**: Pretrained transformer model for text representation.

SOTA models.

- **DepressionNet** [290]: Performs binary depression classification using text summarisation, BERT, BiGRU, attention, and social media features.
- **Naseem et al.** [166]: Utilises TextGCN, BiLSTM, and attention for ordinal regression of depression severity.
- **Hierarchical Attention Network (HAN)** [274]: Hierarchical attention network that models word- and sentence-level structures.
- **Hierarchical Convolutional Network (HCN)** and **HCN+** [291]: Hierarchical convolutional attention networks for word- and tweet-level feature extraction; **HCN+** includes an additional MLP layer.

- **DepressionX** [108]: Our previous model with a fixed knowledge graph and concatenation-based for knowledge infusion.

3.4.e Performance Comparison

Table 3.5 shows results across all datasets. All experiments are run five times and the average evaluation measures and their standard deviations are reported. Our model, AttentionDep, consistently outperforms all compared models, achieving substantial improvements in both severity-level and binary depression detection.

Among the generic neural baselines, Recurrent Neural Network (RNN)-based models (LSTM, BiLSTM, GRU, BiGRU) perform reasonably well due to their ability to capture sequential dependencies, but are limited by vanishing gradients. CNNs perform slightly better by detecting salient local patterns, while BERT achieves stronger results by leveraging transformer-based contextual representations, particularly on **D3**, where ordinal distinctions are more prominent.

The specialised models - DepressionNet [290], HAN [274], HCN [291], HCN+ [291], Naseem et al. [166] and DepressionX [108] - fall short of expectations despite their architectural complexity. Their weaker performance is likely due to insufficient integration of mental-health-specific linguistic cues, which limits their capacity to model fine-grained severity distinctions. Our previous model, DepressionX, also underperforms because of its simplistic knowledge integration strategy and reliance on evaluation settings that are not well aligned with severity classification.

In contrast, AttentionDep achieves over 5% higher graded F_1 than the best baseline on both severity datasets, reaching 80.5% on **D3** (with 77.3% precision and 84.0% recall) and 78.5% on **D4** (78.0% precision, 81.1% recall). Performance is slightly stronger on **D3**, likely because the *mild* class in **D4** overlaps semantically with neighbouring categories, making classification harder. Finally, evaluation on the binary dataset **D2** further confirms the robustness of our approach, where AttentionDep outperforms all baselines by a clear margin.

3.4.f Ablation Study

We conduct an ablation study to evaluate the contribution of each component in AttentionDep, as shown in Table 3.6. All experiments are run five times and the average evaluation measures and their standard deviations are reported. The six configurations are:

- **A0**: Unigram-only features;

Table 3.5: Performance comparison of the proposed AttentionDep across datasets D4, D3, and D2. Results are reported as mean ± standard deviation for GF, GP, and GR.

Model	GF	GP	GR
	D4		
CNN	0.7419 ± 0.0104	0.7826 ± 0.0126	0.7052 ± 0.0114
LSTM	0.7225 ± 0.0140	0.7477 ± 0.0193	0.6992 ± 0.0149
GRU	0.7346 ± 0.0150	0.7654 ±0.0172	0.7064 ± 0.0185
BiLSTM	0.7287 ± 0.0161	0.7483 ± 0.0229	0.7113 ± 0.0304
BiGRU	0.7373 ± 0.0116	0.7686 ± 0.0129	0.7093 ± 0.0268
BERT	0.7241 ± 0.0134	0.7521 ± 0.0236	0.6993 ± 0.0261
DepressionNet [290]	0.6994 ± 0.0149	0.7257 ± 0.0329	0.6760 ± 0.0137
Naseem et al. [166]	0.5663 ± 0.0323	0.4927 ± 0.0124	0.6657 ± 0.0345
HAN [274]	0.7250 ± 0.0118	0.7693 ± 0.0337	0.6892 ± 0.0443
HCN [291]	0.6898 ± 0.0352	0.7223 ± 0.0351	0.6628 ± 0.0374
HCN+ [291]	0.7032 ± 0.0097	0.7269 ± 0.0337	0.6825 ± 0.0174
DepressionX [108]	0.4838 ± 0.1475	0.4332 ± 0.0862	0.5677 ± 0.2101
AttentionDep	**0.7952 ± 0.0133**	**0.7805 ± 0.0185**	**0.8107 ± 0.0120**
	D3		
CNN	0.7525 ± 0.0126	0.7985 ± 0.0170	0.7116 ± 0.0129
LSTM	0.7342 ± 0.0203	0.7555 ± 0.0380	0.7179 ± 0.0479
GRU	0.7400 ± 0.0218	0.7501 ± 0.0612	0.7404 ± 0.0676
BiGRU	0.7369 ± 0.0185	0.7693 ± 0.0468	0.7141 ± 0.0602
BiLSTM	0.7369 ± 0.0145	0.7571 ± 0.0385	0.7205 ± 0.0348
BERT	0.7461 ± 0.0117	0.7846 ± 0.0172	0.7125 ± 0.0311
DepressionNet [290]	0.7223 ± 0.0151	0.7551 ± 0.0392	0.6940 ± 0.0211
Naseem et al. [166]	0.6198 ± 0.0081	0.5838 ± 0.0220	0.6614 ± 0.0086
HAN [274]	0.7040 ± 0.0303	0.7682 ± 0.0636	0.6544 ± 0.0485
HCN [291]	0.6597 ± 0.0040	0.7589 ± 0.0325	0.5843 ± 0.0162
HCN+ [291]	0.6817 ± 0.0273	0.7925 ± 0.0313	0.5986 ± 0.0295
DepressionX [108]	0.5301 ± 0.0934	0.5808 ± 0.2444	0.6552 ± 0.2214
AttentionDep	**0.8052 ± 0.0095**	**0.7732 ± 0.0195**	**0.8404 ± 0.0239**
	D2		
CNN	0.8745 ± 0.0185	0.8755 ± 0.0178	0.8772 ± 0.0171
LSTM	0.8830 ± 0.0132	0.8856 ± 0.0132	0.8821 ± 0.0129
GRU	0.8906 ± 0.0117	0.8939 ± 0.0116	0.8897 ± 0.0124
BiGRU	0.8825 ± 0.0120	0.8853 ± 0.0100	0.8821 ± 0.0135
BiLSTM	0.8751 ± 0.0245	0.8787 ± 0.0229	0.8740 ± 0.0255
BERT	0.8508 ± 0.0222	0.8533 ± 0.0209	0.8561 ± 0.0201
DepressionNet [290]	0.8183 ± 0.0193	0.8190 ± 0.0196	0.8191 ± 0.0196
Naseem et al. [166]	0.8251 ± 0.0009	0.7823 ± 0.0012	0.8601 ± 0.0032
HAN [274]	0.8620 ± 0.0122	0.8633 ± 0.0135	0.8613 ± 0.0123
HCN [291]	0.8253 ± 0.0331	0.8311 ± 0.0300	0.8208 ± 0.0374
HCN+ [291]	0.8265 ± 0.0275	0.8258 ± 0.0277	0.8274 ± 0.0280
DepressionX [108]	0.8015 ± 0.0407	0.7505 ± 0.1094	0.9117 ± 0.1599
AttentionDep	**0.9187 ± 0.0140**	**0.9195 ± 0.0140**	**0.9185 ± 0.0140**

- **A1**: A0 with concatenated KG representation;
- **A2**: A0 with cross-attention-based KG integration;
- **A3**: Unigram + bigram hierarchical features;
- **A4**: A3 with concatenated KG;
- **A5**: Full model (unigram + bigram + cross-attention-based KG).

Table 3.6: Ablation study results showing graded evaluation measures (GF, GP, and GR; Mean ± Std) across D4, D3, and D2 for different model configurations.

Dataset	Configuration	GF	GP	GR
D4	**A0**: Unigrams only	0.7408 ± 0.0164	0.7493 ± 0.0081	0.7334 ± 0.0329
	A1: Unigarms + KG (concat)	0.7637 ± 0.0143	0.7710 ± 0.0196	0.7566 ± 0.0117
	A2: Unigrams + KG (cross att.)	0.7679 ± 0.0056	0.7801 ± 0.0196	0.7566 ± 0.0134
	A3: Unigrams + Bigrams	0.7595 ± 0.0137	0.7548 ± 0.0226	0.7646 ± 0.0129
	A4: Unigrams + Bigrams + KG (concat)	0.7642 ± 0.0085	0.7500 ± 0.0074	0.7794 ± 0.0200
	A5: **Full:** Unigrams + Bigrams + KG (cross att.)	**0.7952 ± 0.0133**	**0.7805 ± 0.0185**	**0.8107 ± 0.0120**
D3	**A0**: Unigrams only	0.7730 ± 0.0131	0.7622 ± 0.0083	0.7927 ± 0.0226
	A1: Unigarms + KG (concat)	0.7810 ± 0.0138	0.7698 ± 0.0226	0.7930 ± 0.0138
	A2: Unigrams + KG (cross att.)	0.7901 ± 0.0219	0.7783 ± 0.0116	0.8025 ± 0.0332
	A3: Unigrams + Bigrams	0.7878 ± 0.0115	0.7559 ± 0.0197	0.8229 ± 0.0142
	A4: Unigrams + Bigrams + KG (concat)	0.7934 ± 0.0122	0.7623 ± 0.0195	0.8277 ± 0.0189
	A5: **Full:** Unigrams + Bigrams + KG (cross att.)	**0.8052 ± 0.0095**	**0.7732 ± 0.0195**	**0.8404 ± 0.0239**
D2	**A0**: Unigrams only	0.8800 ± 0.0096	0.8719 ± 0.0090	0.8900 ± 0.0125
	A1: Unigarms + KG (concat)	0.8990 ± 0.0114	0.9013 ± 0.0121	0.8984 ± 0.0115
	A2: Unigrams + KG (cross att.)	0.9009 ± 0.0236	0.9039 ± 0.0226	0.9001 ± 0.0239
	A3: Unigrams + Bigrams	0.8866 ± 0.0089	0.8819 ± 0.0153	0.8938 ± 0.0038
	A4: Unigrams + Bigrams + KG (concat)	0.8946 ± 0.0223	0.8865 ± 0.0242	0.9054 ± 0.0179
	A5: **Full:** Unigrams + Bigrams + KG (cross att.)	**0.9187 ± 0.0140**	**0.9195 ± 0.0140**	**0.9185 ± 0.0140**

The objectives of this ablation study are mainly to assess the impact of cross-attention-based knowledge infusion and bigram-level hierarchical context modelling. Unigram features alone provide notable performance for depression severity classification. Although adding bigram features very slightly enhances performance on severity datasets (**D4** and **D3**), their main value is explanatory rather than predictive as bigrams help the model highlight meaningful bigrams that are not identifiable through unigrams alone (see Section 3.3.a).

Incorporating domain knowledge via MHKG consistently improves results. The cross-attention-based integration outperforms concatenation because it dynamically aligns textual and knowledge features. The full model achieves the best performance across both multi-class (**D4**, **D3**) and binary (**D2**) settings. The consistently low standard deviations demonstrate stability and robust convergence across configurations.

3.4.g Parametric Analysis

We perform a parametric analysis using 100 hyperparameter tuning trials. To identify the most influential continuous parameters, we apply Spearman's rank correlation coefficient (ρ). For categorical parameters, we use the Kruskal–Wallis (KW) test. The results are reported in Table 3.7. At the 0.05 significance level, the statistical tests indicate that the learning rate, number of LSTM layers, number of attention heads for both unigram and bigram representations, and the number of consecutive GNN layers are the most influential hyperparameters during tuning. Figure 3.3 shows the effects of varying hyperparameters on classification

Table 3.7: Statistical tests for hyperparameter tuning. Significant p-values ($<$0.05) are in **bold**.

Hyperparameter	**Test**	**D3 (stat, p)**	**D4 (stat, p)**
learning rate: η	S	0.228, **0.0225**	-0.485, **0.0000**
dropout rate: d_p	S	-0.062, 0.5410	-0.020, 0.8458
pen. scale: β	S	0.025, 0.8063	0.188, 0.0613
hidden size: h	KW	0.021, 0.8845	3.406, 0.0649
# uni. att. heads: κ_u	KW	17.398, **0.0002**	29.152, **0.0000**
# LSTM layers: δ	KW	9.645, **0.0080**	15.504, **0.0004**
# bi. att. heads: κ_b	KW	6.848, **0.0326**	14.294, **0.0008**
# GNN layers: n	KW	9.877, **0.0072**	20.441, **0.0000**

performance for the severity datasets.

Learning rate. It regulates the step size of parameter updates. Low values lead to slow convergence or getting trapped in local minima, while excessively high values overshoot optimal minimum, causing unstable learning. On **D4**, F1 decreases as the learning rate increases, indicating that high rates may skip information required to distinguish closely related categories (e.g., *minimum*, *mild*, *moderate*). On **D3**, which excludes the mild class, a slight improvement occurs with higher rates, suggesting that reduced semantic overlap allows faster convergence.

Number of LSTM layers. It controls the model's ability to capture long-term dependencies. Two layers yield the best performance on both datasets. In contrast, the number of GNN layers, which governs how deeply the model integrates structural information from the knowledge graph, minimally impacts **D3** but strongly affects **D4**. Optimal settings are a single layer for **D4** and three layers for **D3**.

Number of attention heads. For unigram and bigram layers, this number is highly influential, enhancing the model's ability to capture diverse patterns such

as emotional cues, symptoms, or potential causes simultaneously. For unigram representations, two heads are optimal for D4 and four for D3. For bigram representations, four heads are optimal for both datasets, highlighting the increased capacity of bigram encoding to capture depression-related patterns.

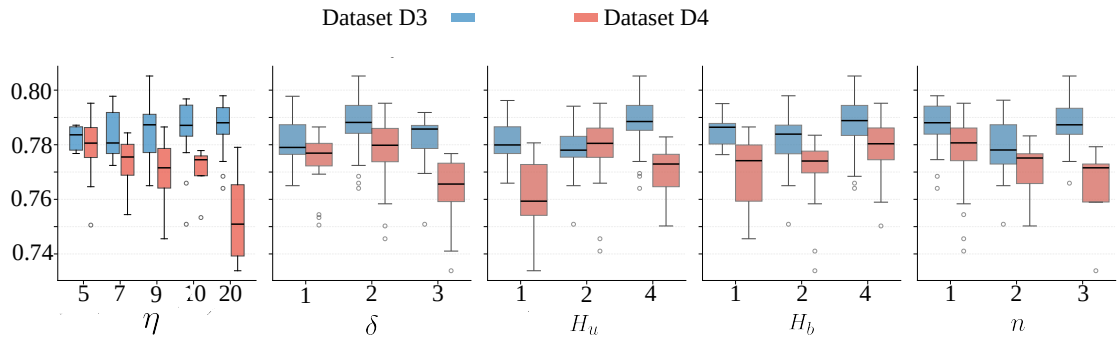


Figure 3.3: Effects of hyperparameters on graded F1 score for severity datasets D4 and D3.

3.4.h Explainability Analysis

Prediction: **Minimum**

actually skyped new year mostly sake grandmother father ' mother positive relations hip somehow friend anyways mother mother selfproclaimed jewish mother subscrib es stereotype upholds disgustingly proud

actually skyped new year mostly sake grandmother father ' mother positive relationship somehow friend anyways mother mother selfproclaimed jewish mother subscribes stereotype upholds disgustingly proud

Figure 3.4: Uni- and Bi-gram level importance visualisation example for the post with minimum depression. Hierarchical attention highlights unigrams and bigrams most influential in the model's predictions, providing insight into the decision-making process.

Attention mechanisms can be associated with model explanation in generating responsible outputs, particularly in healthcare applications where the black-box nature of models raises ethical concerns. Feature importance is one of the most widely used model explanation techniques. Although the consensus is weak among scholars regarding whether attention weights represent feature importance, some studies have empirically validated this association [263]. Mathematically, the partial derivative of the model output y with respect to a feature

x_j is scaled by that feature's attention weight α_j, providing a basis to link attention to feature importance. We employ a hierarchical attention mechanism over

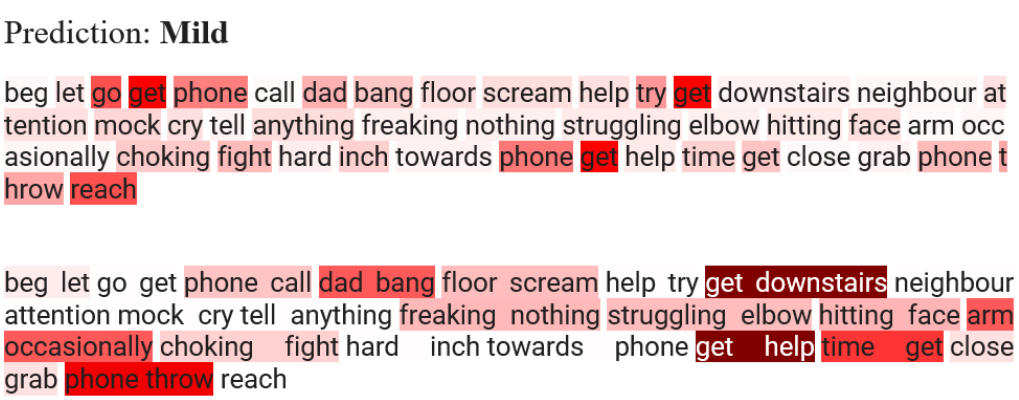


Figure 3.5: Uni- and Bi-gram level importance visualisation example for the post with mild depression.

unigram and bigram features to highlight salient terms, enabling us to analyse how the model makes decisions regarding depression severity. Figures 3.4–3.7 present representative input samples along with their most influential unigrams and bigrams.

For the *minimum* level of depression, the model predominantly focuses on non-depressive words; however, certain negative terms such as *stereotype* or *disgustingly* still receive attention. A similar pattern occurs in the *mild* depression level, supporting the observation that the model struggles to distinguish mild depressive cues from non-depressive content. For *moderate* and *severe* depression, the utility of bigram-level attention becomes more pronounced. In the moderate case, individual words such as *abusive* and *relationship* hold moderate importance, but their combination as the phrase *abusive relationship* significantly increases their saliency, exerting a stronger influence on the final decision. A similar phenomenon appears in the severe class, where phrases like *diagnosed anxiety* receive greater attention than the individual words, highlighting critical patterns indicative of higher depression severity.

3.5 Summary

In this chapter, we present AttentionDep, a novel knowledge-infused and explainable model for predicting depression severity from social media text. AttentionDep integrates hierarchical attention mechanisms over unigram and bigram representations with external domain knowledge from a curated mental health knowledge graph (MHKG). This enables the identification of salient tex-

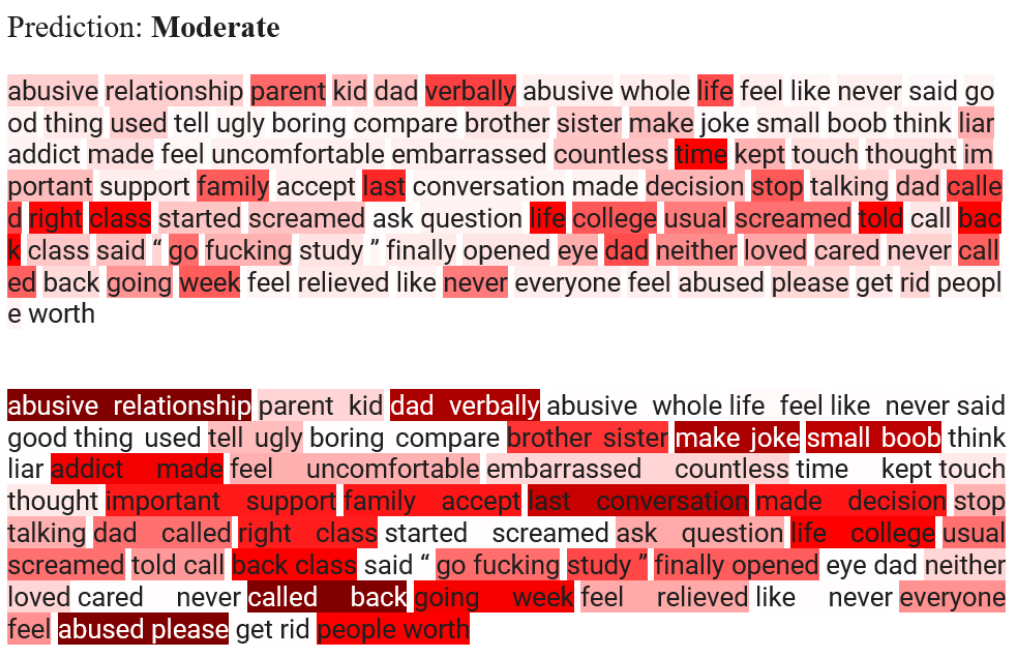


Figure 3.6: Uni- and Bi-gram level importance visualisation example for the post with moderate depression.

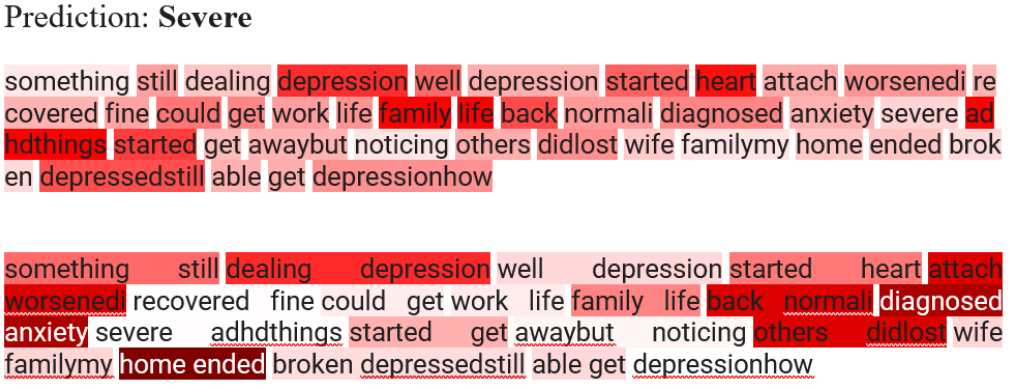


Figure 3.7: Uni- and Bi-gram level importance visualisation example for the post with severe depression.

tual patterns relevant to different depression severity levels. Posts are encoded using FastText embeddings and a BiLSTM to capture contextual dependencies, while convolutional operations extract bigram-level features. In parallel, knowledge from Wikipedia articles is used to construct an MHKG, which is processed via the GINE model to generate enriched knowledge representations. Cross-attention is then applied to fuse textual and MHKG-based features, producing

comprehensive post representations for depression severity classification.

We evaluate AttentionDep on three datasets: **D4** and **D3** (multi-class severity) and **D2** (binary classification). Across all datasets, AttentionDep consistently outperforms baseline methods, achieving graded F_1 scores of 79.52% on **D4** and 80.52% on **D3**, exceeding all compared models by over 5%. On the binary dataset **D2**, the model achieves a graded F_1 score of 91.87%. These results highlight the effectiveness of combining domain knowledge with explainable deep learning for mental health prediction. However, graded scores should not be interpreted in the same way as an accuracy, as they do not reflect the proportion of correctly predicted cases. Instead, these metrics represent how well the model preserves ordinal closeness by penalising deviations between the annotated and predicted severity levels. In the literature [215], effective models typically achieve graded F-scores higher than 70%, and higher values indicate stronger ordinal consistency rather than a direct false-positive or false-negative rates. Therefore, high graded scores reflects that the model's predictions remain close to the ground-truth severity level, even when not exact, which is the intended behaviour for ordinal mental disorder severity estimation tasks.

Despite these promising results, several limitations remain. While attention mechanisms provide interpretability and the model integrates clinically validated domain knowledge, future work could explore additional strategies to further quantify how attention weights correspond to specific depression-relevant features. Moreover, the current model focuses exclusively on textual data, omitting other potentially informative multimodal signals, such as behavioural features or social interactions. Future work could enhance interpretability and predictive performance by incorporating such signals and modelling user connections within social networks. Overall, AttentionDep demonstrates a robust framework for explainable, knowledge-driven depression detection from social media, offering valuable insights for Artificial Intelligence (AI)-assisted mental health applications.

4 Learning Contributors to Mental Health Challenges with Causality Inspired Contextual Attention

4.1 Introduction

Mental Health (MH) conditions, including depression, anxiety, eating disorders, and suicidal thoughts, are major global public health concerns, affecting approximately 12.5% of the population worldwide[1]. Mental health challenges are complex phenomena that significantly influence individuals' mental well-being, affective states, and behavioural patterns. These disorders are closely associated with suicide, leading to nearly 800,000 deaths annually [109]. Stress, a widespread issue, is strongly linked to mental disorders [70], affecting 60% of 18–24-year-olds in 2018 [2]. Early diagnosis and timely intervention are critical for mitigating the long-term effects of these conditions [83]. A significant portion of our society (59.9% approx.[3]) actively engages in Online Social Media (OSM) and Conversational Systems (CS) [109, 16]. OSM and CS enable users to express their emotions and thoughts naturally in a safe environment, offering valuable insights into their mental health and stress-related conditions [2, 153, 112]. However, manual examination of such data is costly and time consuming, especially considering its volume.

Machine learning models have demonstrated potential for automating the identification of possible indicators of the presence of mental health issues using OSM and CS based data, offering a scalable approach to early detection [196, 215, 156, 1]. However, analysing these data is challenging due to its informal nature, brevity, noise, variability, and the diverse backgrounds of users. These

[1] www.who.int/news-room/fact-sheets/detail/mental-disorders

[2] www.mentalhealth.org.uk/explore-mental-health/statistics/stress-statistics

[3] www.datareportal.com/social-media-users

challenges limit the effectiveness of existing methods, which primarily focus on detecting mental disorders and stress without addressing their underlying contributors. Focusing solely on detection can miss the broader context necessary to understand the underlying factors, consequently limiting the development of precise and personalised interventions. Addressing these limitations is critical for improving the reliability, interpretability and explainability of results, as well as the efficacy of mental health support systems [156, 166, 2]. Therefore, identifying and understanding the contributors to mental health conditions is key to improving screening methods and designing individualised intervention methods [49].

Some early studies have investigated the factors contributing to mental health issues. Mauriello et al. [153] created a dataset SAD by annotating everyday stressors in SMS-like conversations and analysed it with traditional machine learning techniques. Similarly, Garg et al. [79] created a multi-class annotated dataset CAMS for identifying potential causes of mental disorders on OSM and applied various machine learning models. However, these datasets and approaches have significant limitations. The label space in CAMS does not fully represent realistic scenarios and needs expansion. The SAD dataset offers a more detailed categorisation of daily stressors, with nine categories compared to CAMS' six. However, it contains short texts in SMS format, complicating detailed analysis. Furthermore, both the SAD and CAMS datasets are prone to biased annotations. In particular, the identified cause may not always directly relate to the author's emotional state but may instead involve third parties mentioned in the post. For instance, the post, *Hello, what's stressing me out right now is my son's homework, he got too much homework*, is labelled as *School* in the SAD dataset, even though the real cause is the author's emotional stress regarding their son's workload, not the school context itself. A similar post in the same dataset is labelled as *Family issues*. Similarly, the post, *I would like to say that I was shook, but I knew she has the habit of consuming a lot of sleeping pills just for not having to deal with the daily problems...* from the CAMS dataset is labelled as *Medication (Health Issue)*. However, this label overlooks the true cause of the author's stress, which is their guilt and helplessness regarding their friend's situation. With rapid development of Large Language Model (LLM)s, Yang et al. [272] fine-tuned LLMs with an aim to explain the reasoning behind mental health issues. Despite their potential, LLMs often produce inconsistent results due to the inherent instability and dataset limitations.

To this end, we conduct a comprehensive study on identifying perceived contributors to mental health challenges, including mental disorders and stress, from textual expressions in OSM and CS. To facilitate this, we introduce RED-

CoM, a new Reddit dataset for studying Contributors to Mental health challenges. dataset aggregates textual social media posts from publicly available corpora explicitly flagged for mental health risk indicators. These posts were further annotated with potential contributors through an automated pipeline, refined by a human-in-the-loop approach. This manual refinement was conducted by graduate students, following annotation guidelines established with academics in psychology. Reddit is primarily selected due to its prevalence in current studies and existing datasets. Unlike character limited platforms, it allows for long, detailed textual posts which provide deeper insight into the emotional states of users. Furthermore, the official Python Reddit API Wrapper (PRAW) offers a free and robust framework for collecting open social media data. The annotations were carried out by trained graduate students under the guidance of academics in psychology, ensuring high-quality labels. To advance detection capabilities, we propose MDCNet, multi-class detection model that leverages transformer-based text representations alongside a causality inspired contextual attention mechanism. This enables the model to estimate latent task relevant features, reduce bias, and improve detection performance. Extensive experiments are conducted to evaluate the model's effectiveness. The results are compared against several baselines and state-of-the-art models using benchmark datasets from diverse sources. In summary, this study makes the following key contributions.

1. We formulate the problem of conducting an extensive analysis of contributors to mental health issues on OSM and CS.
2. We present RedCoM, a new Reddit dataset for analysis of Contributors to Mental health challenges.
3. We introduce MDCNet - Mental Disorder Contributors Network, a novel multi-class detection model designed to identify contributors to mental health challenges. By combining transformer-based text representations with a causality inspired contextual attention mechanism, MDCNet estimates latent task relevant and spurious features, reducing bias and improving detection performance.
4. We perform comprehensive experiments to evaluate the effectiveness of the proposed model. Our results are compared to state-of-the-art methods on benchmark datasets, demonstrating significant improvements.

This chapter is organised as follows. Section 4.2 demonstrates the theoretical background knowledge behind our methodological motivation. Section 4.3 details the annotation process and characteristics of the RedCoM dataset. Section 4.5 presents the proposed model, MDCNet, including a mathematical formulation, an overview, and the architectural details. The experimental study is presented in Section 4.6, and finally we conclude the chapter in Section 4.7.

4.2 Background Knowledge

4.2.a Attention for feature importance

Since their emergence, attention mechanisms have led to significant advances in AI and have become a core component of the architectures of many state-of-the-art models, including LLMs. Although attention is frequently used to estimate feature importance and interpret model behaviour, its effectiveness as a tool for explainability remains actively debated in the literature [111, 263]. To address this question, we begin by presenting a formal theoretical analysis of the relationship between attention and feature importance.

Lemma 4.2.1 (Role of Attention in the Output Decision). *Let $X = [x_1, x_2, \ldots, x_n]^\top \in \mathbb{R}^{n\times d}$, where n is the number of features and d is the embedding dimension of each feature. Let $\Lambda = [\alpha_1, \alpha_2, \ldots, \alpha_n]^\top \in \mathbb{R}^n$ be the attention weights, satisfying $\alpha_j \geq 0$ and $\sum_{j=1}^n \alpha_j = 1$. Define the aggregated representation:*

$$z = \sum_{j=1}^{n} \alpha_j x_j \in \mathbb{R}^d, \tag{4.1}$$

and let the output decision be given by:

$$y = f_\theta(z), \quad \text{where } f_\theta : \mathbb{R}^d \to \mathbb{R} \tag{4.2}$$

is a differentiable decision function parameterised by θ. Then, each attention weight α_j directly influences the output y, and furthermore, α_j reflects the relative contribution (importance) of the corresponding feature x_j to the final decision.

Proof. We prove both claims via gradient analysis.

(1) Direct influence of attention on the output:
For each attention weight α_j, the derivative of y with respect to α_j is given by:

$$\frac{\partial y}{\partial \alpha_j} = \frac{\partial f_\theta(z)}{\partial \alpha_j} = \frac{\partial f_\theta(z)}{\partial z} \times \frac{\partial z}{\partial \alpha_j} \tag{4.3}$$

$$= \langle \nabla_z f_\theta(z), x_j \rangle \tag{4.4}$$

Since x_j and $\nabla_z f_\theta(z)$ are generally non-zero, the attention value α_j has a direct and non-trivial influence on the output y.

(2) Attention encodes feature importance:

We compute the derivative of y with respect to the feature vector x_j:

$$\frac{\partial y}{\partial x_j} = \frac{\partial f_\theta(z)}{\partial z} \times \frac{\partial z}{\partial x_j} \tag{4.5}$$

$$= \alpha_j \cdot \nabla_z f_\theta(z) \tag{4.6}$$

Taking norms, we obtain:

$$\left\| \frac{\partial y}{\partial x_j} \right\| = \alpha_j \cdot \|\nabla_z f_\theta(z)\| \implies \left\| \frac{\partial y}{\partial x_j} \right\| \propto \alpha_j \tag{4.7}$$

which implies that the sensitivity of the output to feature x_j is scaled by α_j. Hence, the magnitude of α_j determines the relative importance of x_j in influencing the decision y.

We have shown that attention weights α_j both (1) directly influence the output decision, and (2) reflect the relative importance of features in determining that decision, under any differentiable decision function f_θ. □

We have simply justified the possibility of interpreting attention weights as indicators of feature relevance. The current formulation assumes the basic form of the attention mechanism attached to a linear aggregation step. In more complex architectures (e.g., transformers), attention weights themselves may be functions of X, leading to a recursive dependency.

4.3 The RedCoM dataset

As discussed in the introduction, the existing datasets (SAD and CAMS), used to identify factors contributing to mental health challenges, suffer from several limitations. In SAD, the textual inputs are very short, reflecting the concise nature of the SMS-based entries. In CAMS, the predefined set of categories is limited, restricting the depth of analysis. Furthermore, the assigned labels in both datasets are prone to bias. To address these shortcomings, we introduce a new Reddit dataset for analysing contributors to mental health challenges, called RedCoM.

4.3.a Overview of Categories

Mental disorders are influenced by a wide range of adverse life events. Experiences such as abuse, violence, discrimination, and cyberbullying are often associated with the onset of mental health problems [37, 99, 200, 260]. Financial hardships, such as low income, limited savings, and lack of homeownership,

particularly affect younger adults and contribute substantially to mental health challenges [71, 74, 288]. Material hardships and career-related setbacks, including debt, unemployment, and workplace conflicts, are closely linked to the symptoms of psychiatric disorders [10, 33, 31, 78, 133, 198, 98]. Furthermore, grief, an emotional response to death, plays a critical role in elevating the risk of mental illness, with vulnerability peaking during the first year following the loss of loved ones [289, 101]. Similarly, physical health problems, such as chronic illnesses and physical impairments, worsen mental health difficulties by reducing quality of life [12, 257, 171]. Substance addiction often forms a supplementary relationship with mental disorders, acting as a contributor to mental health challenges and stress [52]. Moreover, social factors such as relationship breakdowns, family conflicts, and social isolation are significant contributors to psychological disorders, by amplifying the risk of mental illness [136, 241, 254, 178, 152, 82]. Guided by these insights and supported by prior mental health research [79, 153], we identified six potential contributors to mental disorders and stress in consultation with field experts who are academics holding PhD in psychology.

- **Bias/Abuse**: Encompasses emotional or psychological distress resulting from experiences of abuse, bias, violence, cyberbullying, or discrimination. Example: *All my friend's do is harass and bully me so they're no help. I just wish life would go on without me*
- **Relationship Issues**: Involves difficulties in romantic, familial, or social relationships, including conflicts, social isolation, loneliness, and strained interpersonal connections. Example: *Is our relationship worth it? TL;DR My boyfriend of over two years has a huge drug problems and continuously lies to my face about it and it is effecting our relationship in a negative way.*
- **Finance/Career Concerns**: Refers to hardships associated with financial instability, debt, career challenges, and workplace-related pressures. Example: *Can't find a f*****g job, I am lost and hopeless. I need money. I feel lost that I can't fucking find a job. All the interviews that I got, I was rejected.*
- **Loss and Grief**: Covers the psychological impact of bereavement or anticipated loss of close relations, often resulting in prolonged grief and depression. Example: *My dad already killed himself 4 years ago. I don't know how to deal with this pain anymore but I don't want my family and girlfriend to go through a tragic death.*
- **Health Concerns**: Includes the mental health effects of physical illness, injury, chronic pain, physical impairments, and substance addiction. Example: *Cigarettes used to help but they don't anymore. Alcohol helps, but then it comes back worse the next day.*
- **None**: Applies when no identifiable contributor aligns with the above categories or when alternative factors are suggested. Examples: *Wish I*

*could f*****g cry tbh. I have never felt less like a human being*

4.3.b Annotation Mechanism

Data Collection

The candidate OSM posts for annotating mental disorder contributors were sourced from publicly available datasets. These datasets include textual OSM posts flagged by field professionals based on the indications of suicide risk [81, 96], depression [196] and stress [249]. To minimise potential risks and biases in the annotation process, we employed a systematic annotation and validation strategy. Considering the resource-intensive nature of manual annotation, we integrated automated methods using the LLMs with a human-in-the-loop framework. In the first stage, all candidate labels for each post were jointly generated by LLMs and four trained graduate students as annotators following the guidelines established with mental health professionals, who are academics in psychology. In the second phase, the final label for each post was identified through complete consensus between the LLMs and the annotators. In the final phase, data augmentation was conducted by including posts from the CAMS dataset and publicly available Reddit mental health subreddits. Each stage was independently evaluated by human validators to ensure accuracy and consistency.

First Phase of Annotation

Initially, 500 labelled OSM posts (Set_1), where all annotators achieved full agreement, were subsequently labelled by six benchmark LLMs including Generative Pre-trained Transformer (GPT)-4o-mini [4], Qwen2.5-14B-Instruct [5], Llama-3.1-8B-Instruct [6], Mistral-7B-Instruct-v0.3 [7], MentalLlaMA-chat-7B [8], and Phi-3.5-mini-instruct [9]. Selection criteria of the LLMs were their strong performance across multiple leaderboards. For each LLM i, a priority score (α_i) was computed by normalising their F_1-scores, which were then used to derive a confidence score (S_j) for labelling each post j, as defined in Eq. 4.8:

[4] www.openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[5] www.huggingface.co/Qwen/Qwen2.5-14B-Instruct-1M

[6] www.huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[7] www.huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[8] www.huggingface.co/klyang/MentaLLaMA-chat-7B

[9] www.huggingface.co/microsoft/Phi-3.5-mini-instruct

$$S_j = \sum_i \alpha_i \cdot y_{i,j}, \quad \forall j \in \{1, 2, \ldots, N\}, \forall i \in \{1, 2, \ldots, 6\} \tag{4.8}$$

where N denotes the number of samples, and $y_{i,j}$ indicates the prediction for post j generated by LLM i. For each label, confidence scores close to 0 were mapped to label 0, while scores close to 1 were mapped to label 1. Thresholds were empirically optimised to maximise predictive performance on Set_1. We further used LLMs to annotate the next batch of the social media posts. Posts with intermediate confidence scores were flagged as ambiguous and re-annotated by human annotators. This procedure enabled the time and resource efficient annotation of 2,000 additional posts. The prompt employed in this phase is reported in Table A.2. Placeholders within square brackets were adapted to each category.

Second Phase of Annotation

Although substantial agreement among annotators was achieved in the first phase, as measured by Cohen's Kappa (Section 4.3.c), a second round of annotation was conducted to reduce labelling noise. In this phase, LLMs were further employed to determine the final label from the previously assigned multi-labels for each post. The prompt used for final label determination is provided in Table A.2. This step focused on posts where all LLMs reached consensus on the same label, then reviewed under human supervision. Following the human validation process, 701 OSM posts were confirmed to have reliable labels.

Data Augmentation and Third Phase of Annotation

To address the limited number of annotated posts from Phase 2, we augmented the RedCoM dataset using two complementary strategies. First, relevant posts from the CAMS dataset were incorporated, with labels re-verified by two independent graduate students following the training guidelines established with academics in psychology. This integration contributed 230 posts labelled as Relationship Issues, 200 as Finance/Career, 164 as Health Concerns, 81 as Bias/Abuse, and 200 without a specific contributor. To further mitigate class imbalance in underrepresented categories, the dataset was supplemented with posts from mental health related subreddits (Table **??**). Each newly collected post was comprehensively annotated, achieving 100% agreement among annotators. This process yielded 64 posts in Bias/Abuse, 18 in Health Concerns, 110 in Loss and Grief, 5 in Finance/Career, 3 in Relationship Issues, and 18 without a specific contributor. Overall, newly collected posts constitute approximately 12% of the

complete dataset. After removing duplicates, the final dataset comprises 1,598 samples. The resulting class distribution is shown in Figure 4.3.

4.3.c Dataset Characteristics

Annotation Validation

To ensure the quality and reliability of RedCoM, validation was conducted at the end of each annotation phase. During validation, 10% of the dataset was reviewed by an independent, trained graduate student annotator who was not involved in the initial labelling process.

Table 4.1: Cohen's Kappa scores by category

Category	Phase 1 (%)	Phase 2(%)	Phase 3(%)
Bias/Abuse	61%	87%	91%
Finance/Career	64%	92%	97%
Relationship Issues	63%	91%	92%
Health Concerns	62%	97%	92%
Loss/Grief	74%	97%	98%
None of them	70%	93%	90%

Annotation reliability was determined using Cohen's Kappa (C.K.). The results, summarised in Table 4.1, indicate that the overall substantial agreement in Phase 1 (C.K. > 60%) increased to strong agreement in Phase 2 (C.K. > 80%) and reached very strong agreement in Phase 3 (C.K. > 90%) after final decisions and noise reduction. This steady improvement across phases highlights the effectiveness of the three-phase annotation process in reducing bias and annotation noise.

Dataset Analysis

A couple of annotated data samples can be found in Table 4.2. To further examine linguistic patterns, we analysed the LIWC scores across different labels. LIWC features can capture and reflect cognitive, emotional, and social aspects of language, and can help identify potential contributors to mental health issues. The radar plots in Figure 4.1 illustrate how these features are distributed across the categories. Direct emotional signals including anger and anxiety, appear across all groups, indicating that users often express their concerns in emotionally charged ways. Health related posts commonly contain negative health terms and emotional language. In the Bias/Abuse category, the content strongly highlights past events, conflict, risk, and social interaction, likely

Table 4.2: Example of annotated posts across the six labels

Reddit Post	Ground Truth
You ever feel like you're not meant to exist?. Im about to graduate and I feel 0 desire to go to college and find a career or anything. What's the point? To while my time away because everyone else is doing it? Lol. Thanks, but no thanks. I'd much rather just not exist. This shits boring and dumb af.	**Finance/Career**
I pray that nobody else has to endure the terror and horrible horrible physical, sexual, and emotional abuse you have put me through. I am riding myself of all my shame associated with what has happened. I know now, it was all you. I did not do anything wrong to deserve the horrible treatment I received from you.	**Bias/Abuse**
Anyone have PTSD from a traumatic experience of grief of a loved one?. My dad died 3 years ago in Feb and he died in a traumatic way at home, just wanting to see if anyone else has similar experiences?	**Loss and Grief**
Bad family, just got out of a bad relationship, have always been bad at making friends (serious social anxiety). This ain't abnormal but I stay in contact with toxic people because I'm just desperate for that human warmth.	**Relationship**
Cigarettes used to help but they don't anymore. Alcohol helps, but then it comes back worse the next day. In the months that it's not happening, I'm well-adjusted. I thrive. But then my luck changes, someone starts listening to music next door or revving an old engine in their yard, and all my progress collapses.	**Health Concerns**
Fuck i just cant do it im stuck in this nightmare. Here i am in bed holding a knife its been like an hour that im trying to get to courage just to stab myself in the carotid artery already wrote a suicide note and everything but im such a coward i just cant do it fuck fuck fuck i dont wanna live in this nightmare anymore i dont wanna wake for tomorrow everyday day is the same shitty day	**None of them**

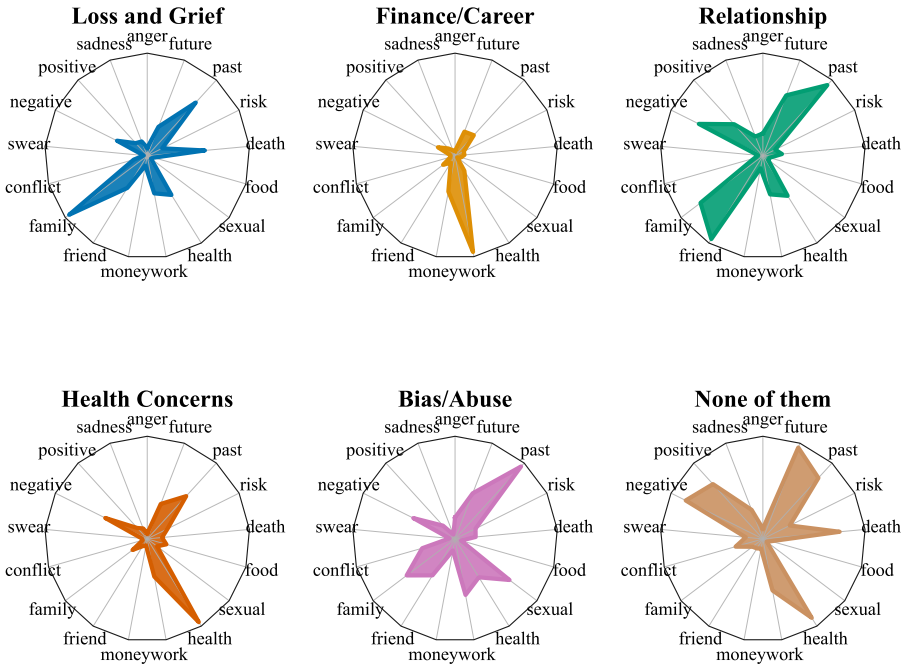


Figure 4.1: Radar plot of Linguistic Inquiry and Word Count (LIWC) feature distributions

reflecting users' traumatic experiences. Family references are salient, possibly the harm originates from a family member or from a lack of family support. Mentions of sexual content is also noticeable, often linked to abuse and boundary violations. Particularly, some posts contain traces of positive tone, which may reflect sarcasm or emotionally repressed trauma. Posts about finance and career are predominantly future oriented, highlighting concerns about stability, expectations, and professional challenges. Although these posts convey mixed emotions, they are occasionally optimistic, reflecting both effort and perseverance. Loss and Grief posts demonstrate a strong focus on the past, family, and death, which aligns with the emotional impact of losing close relatives. These posts also demonstrate a blend of negative and positive emotional tones, suggesting bittersweet memories. Relationship related posts frequently mention friends, conflict, and negative emotions. They are often past-focused, indicating signs of rumination. Frequent mentions of both familial and friendship related cues highlight that these concerns are rooted in social dynamics. Overall, these patterns suggest that LIWC features effectively capture the emotional and contextual differences across mental health related concerns, providing valuable signals for understanding their contributing role in mental health challenges.

4.4 Mathematical Formulation of the Problem

We formally define the task of identifying perceived contributors to mental disorders and stress as a classification problem. Let $\mathcal{X}$ denote the input feature embedding space derived from textual expressions in OSM and CS, and let $\mathcal{Y}$ represent the label space, where $\mathcal{Y} = \{c_1, c_2, \ldots, c_m\}$ and may vary across datasets. The classification model f_θ is defined as:

$$f_\theta : X \subseteq \mathcal{X} \mapsto Y \in \mathcal{Y} \tag{4.9}$$

Let $X = \{x_1, x_2, \ldots, x_N\} \in \mathcal{X}$ denote the set of features (tokens) for a textual input, which may include both task relevant (C) and spurious (U) features. Distinguishing between these two types is essential for debiased classification. Given the complex and latent nature of these features, we adopt a probabilistic approach to separate task relevant and spurious signals. Therefore, we define a measure $\mu : \mathcal{X} \to [0, 1]$ that quantifies the likelihood of a feature belonging to the task relevant set, subject to the normalisation constraint $\sum_{i=1}^{N} \mu(x_i) = 1$ for any feature set X. By definition, if $\mu(x_i) = 1$, then x_i is fully task relevant, and if $\mu(x_i) = 0$, it is fully spurious. Since C and U are disjoint, $\mu(x_i)$ indicates the likelihood of $x_i \in C$, while $1 - \mu(x_i)$ indicates the likelihood of $x_i \in U$. We implement this measure μ using an attention mechanism in our neural architecture, inspired by [236, 47]. The attention mechanism is appropriate for this role, as it can reflect feature importance and directly influence the model output, as shown in Lemma 4.2.1. In the mental health context, developing explicit causal associations between input words and outcomes is challenging and often impossible. The contribution of words is cumulative and context dependent, meaning that tokens can not definitively be labelled as purely causal or purely spurious. Their roles should therefore be remarked as fuzzy (probabilistic) rather than explicit. Consequently, our attention mechanism does not claim to identify true causal features in the sense of causal inference theory. Instead, it should be considered as a debiasing framework for classification inspired by SCM, to identify the **task-relevant features**. Moreover, the softmax function acts as a probability mass function, aligning with the behaviour of μ and reflecting the role of each x_i in the decision-making process. Given a measure μ, we define:

$$C_X = \{x_i | \mu_i \approx 1\} \approx \text{top}_k(X, \boldsymbol{\mu}), \quad U_X = \{x_i | \mu_i \approx 0\} \approx \text{top}_k(X, 1 - \boldsymbol{\mu}) \tag{4.10}$$

The backdoor adjustment equation (Equation B.5) enables the estimation of the actual causal effect by marginalising over all possible causal and confounding (spurious) partitions. In this study, this formula is adapted by weighting the

features using the probabilities obtained via attention mechanism. To optimise this framework, we define the following objective function:

$$\arg\max_{\theta,\mu} \mathbb{E}[\log P(Y|C_X;\theta)] \tag{4.11}$$

where θ represents the model parameters and $\mu = (\mu_1, \mu_2, ..., \mu_d)$ represents the probability distribution over task relevant features. This objective maximises the expected log-likelihood of the true predictions based on the task relevant features, by learning both the model parameters θ and the probabilities μ simultaneously. To ensure that $\mu(x_i)$ accurately reflects the likelihood of a feature being task relevant, we encourage task relevant features to approximate the true output, while spurious features are forced to return incorrect outputs. We further assume that task relevant and spurious features are independent, since they are disjoint by their nature. The resulting optimisation objective is defined as:

$$\arg\min_{\theta,\mu} \left(\mathbb{E} \left[\underbrace{\mathcal{L}_1\left(f_\theta(C_X), y\right)}_{\text{task relevant Prediction}} - \underbrace{\mathcal{L}_2\left(f_\theta(U_X), y\right)}_{\text{spurious Suppression}} + \underbrace{\mathcal{L}_3(C_X, U_X)}_{\text{Independence}} \right] \right) \tag{4.12}$$

where $\mathcal{L}_1$ forces task relevant features to improve prediction, $\mathcal{L}_2$ suppresses the influence of spurious features on the decision, and $\mathcal{L}_3$ enforces disjointness between task relevant and spurious features.

4.5 Proposed Framework: MDCNet

4.5.a Model Overview

Motivated by the mathematical formulation, we propose a novel multiclass classification model, MDCNet, for identifying perceived contributors to mental disorders and stress (Figure 4.2). MDCNet extends advanced transformer based architectures by incorporating a causality inspired contextual attention mechanism to capture implicit task relevant relationships while mitigating the influence of spurious features in textual data. The architecture consists of three primary components:

1. *Input Representation Layer*: leverages a pre-trained transformer encoder to represent textual inputs.
2. *Decomposition Layer*: employs causality inspired contextual attention to compute masked attention to identify task relevant and spurious features.
3. Adjustment Layer: distinguishes between task relevant and spurious features using specialised loss functions.

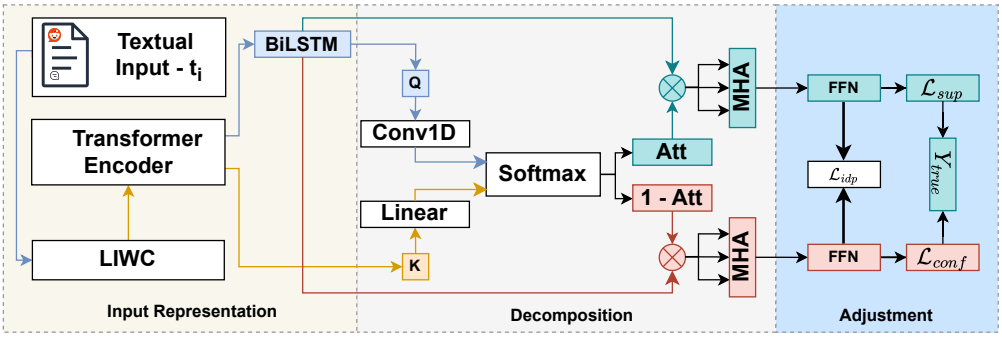


Figure 4.2: Architecture of the proposed MDCNET model

4.5.b Input Representation

Textual content in OSM and CS often lacks clear grammatical structure and contains abbreviations and slang, making it challenging to analyse. Understanding context is particularly important for mental health analysis, as the same word can possess different meanings depending on the author's emotional state and the situation. For instance, both sentences *I'm tired from a productive day* and *I'm so tired of everything* contain the word *tired*, however, in the second sentence, the word reflects signs of exhaustion, potentially hopelessness or emotional fatigue. Contextual embeddings capture word meanings based on their surrounding context, in contrast to traditional methods that assign a single vector to each word. Accordingly, each textual input t_i is first passed through the model $\Phi_{\text{enc}}(\cdot)$, a placeholder that can be replaced by any pre-trained transformer encoder capable of generating contextual embeddings. This model generates a last hidden state tensor of shape $\mathbb{R}^{L\times d}$, where L is the maximum sequence length, and d the embedding dimension. To further enhance sequential contextual understanding, these embeddings are processed by a Bidirectional Long Short-Term Memory (BiLSTM) network [102], described as $\mathbf{E}_{t,i} = \textbf{BiLSTM}(\Phi_{\text{enc}}(t_i))$, where $\mathbf{E}_{t,i} \in \mathbb{R}^{L\times d}$ represents the contextualised embeddings derived from t_i. The BiLSTM captures complex sequential linguistic patterns and contextual relationships by processing both forward and backward dependencies in OSM and CS content. This capability is particularly beneficial for analysing contributors to mental health challenges, where latent linguistic cues may significantly influence the identification of underlying factors. In addition to textual embeddings from the pre-trained transformer, we extract psycholinguistic features using LIWC [190], a quantitative method that categorises words into psychologically meaningful dimensions, such as emotion, cognition, and social concerns. The re-

sulting LIWC features are subsequently processed by a pre-trained transformer encoder as follows:

$$\mathbf{V}_i, \xi_i = \text{LIWC}(t_i); \qquad \mathbf{E}_{k,i} = \mathbf{V}_i \odot \Phi_{\text{enc}}(\xi_i) \tag{4.13}$$

where, $\mathbf{V}_i$ represents the values of the LIWC features associated with the categories ξ_i, while $\mathbf{E}_{k,i} \in \mathbb{R}^{L \times d}$ denotes the final LIWC based embeddings, and $\odot$ indicates the Hadamard product.

4.5.c Decomposition

As discussed in Section 4.4, the attention mechanism quantifies the likelihood μ that tokens contribute genuinely to mental disorders and stress. To ensure that μ reflects psycholinguistic significance, we employ contextual cross-attention, which captures domain-specific token importance. Following [252], the attention values are computed as:

$$\mu_i = \sigma\left(\frac{\text{Conv1D}(\mathbf{E}_{t,i}) \times (\mathbf{E}_{k,i}\mathbf{W}_{\mathbf{K}_\mu})^T}{\sqrt{d}}\right) = \tag{4.14}$$

$$= \sigma\left(\frac{\left\{\mathbf{W}_{\mathbf{C}_\mu} \cdot \begin{bmatrix}\mathbf{E}_{t,i,j} \\ \mathbf{E}_{t,i,j+1}\end{bmatrix} + \mathbf{b}\right\}_{j=0}^{L-1} \times (\mathbf{E}_{k,i}\mathbf{W}_{\mathbf{K}_\mu})^T}{\sqrt{d}}\right) \tag{4.15}$$

where σ denotes the softmax function, $\mathbf{W}_{\mathbf{C}_\mu}$ and $\mathbf{W}_{\mathbf{K}_\mu}$ are trainable parameter matrices, and μ_i captures the psycholinguistic importance of tokens in the textual input t_i. This indicates that the importance and likelihood of a token being task relevant are scaled according to its psycholinguistic significance and potential impact on contributors to mental health issues. As expressions and phrases can be as informative as individual words in analysing mental health challenges, we incorporate n-gram feature analysis, using the Conv1D output of textual social media data as the query for computing attention scores. LIWC represents features in tabular form, therefore, a linear model is employed during key and value computation to appropriately process these different feature types. task relevant and spurious feature extraction is defined by selecting top-K features on the embedding space $\mathbf{E}_{t,i}$ using μ_i, as follows:

$$\mathbf{E}^C_{t,i} = \text{topK}(\mathbf{E}_{t,i}, \mu_i, K) = \{\mathbf{e}_{t,i,j} \in \mathbf{E}_{t,i} \mid \mu_{i,j} \in \text{top-K}(\mu_i)\} \tag{4.16}$$

$$\mathbf{E}^U_{t,i} = \text{minK}(\mathbf{E}_{t,i}, \mu_i, K) = \{\mathbf{e}_{t,i,j} \in \mathbf{E}_{t,i} \mid \mu_{i,j} \in \text{top-K}(\mathbf{1} - \mu_\mathbf{i})\} \tag{4.17}$$

where top-K$(\cdot)$ returns the K largest values in the attention matrix, and the top$K$ (or min$K$) operation selects the corresponding elements from $\mathbf{E}_{t,i}$ according to these values. Finally, task relevant and spurious features are disentangled via a modified multi-head attention (MHA) mechanism [252]:

$$\hat{\mathbf{y}}_i^C = \text{MHA}_C(\mathbf{E}_{t,i}^C) = \tag{4.18}$$
$$= \text{FFN}_C\Bigg(\underbrace{\text{Max}\left[\bigoplus_{k=1}^{H}\sigma\left(\frac{(\mathbf{E}_{t,i}^C\mathbf{W}_{\mathbf{Q}_C}^k)(\mathbf{E}_{t,i}^C\mathbf{W}_{\mathbf{K}_C}^k)^T}{\sqrt{d_k}}\right)(\mathbf{E}_{t,i}^C\mathbf{W}_{\mathbf{V}_C}^k)\right]\mathbf{W}_C}_{\mathbf{H}_C}\Bigg)$$

$$\hat{\mathbf{y}}_i^U = \text{MHA}_U(\mathbf{E}_{t,i}^U) = \tag{4.19}$$
$$= \text{FFN}_U\Bigg(\underbrace{\text{Max}\left[\bigoplus_{k=1}^{H}\sigma\left(\frac{(\mathbf{E}_{t,i}^U\mathbf{W}_{\mathbf{Q}_U}^k)(\mathbf{E}_{t,i}^U\mathbf{W}_{\mathbf{K}_U}^k)^T}{\sqrt{d_k}}\right)(\mathbf{E}_{t,i}^U\mathbf{W}_{\mathbf{V}_U}^k)\right]\mathbf{W}_U}_{\mathbf{H}_U}\Bigg)$$

where MHA$_C$ and MHA$_U$ denote multi-head attention mechanisms responsible for disentangling task relevant and spurious features, respectively. $\mathbf{W}_{\mathbf{Q}_C}^k$, $\mathbf{W}_{\mathbf{K}_C}^k$, $\mathbf{W}_{\mathbf{V}_C}^k$, $\mathbf{W}_{\mathbf{Q}_U}^k$, $\mathbf{W}_{\mathbf{K}_U}^k$ and $\mathbf{W}_{\mathbf{V}_U}^k$ represent trainable parameters. $\mathbf{H}_C$ and $\mathbf{H}_U$ are final hidden representations of task relevant and spurious features. The use of MHA in this process is intentional, as mental health related features often exhibit complex and latent characteristics. Moreover, contributors to mental health challenges can vary substantially, requiring analysis from multiple perspectives. MHA is particularly well suited for this task, as it enables the model to attend to diverse aspects of the data and thereby achieve more robust representations. Finally, predictions are obtained through task relevant and spurious feedforward networks (FFN$_C$ and FFN$_U$), where $\hat{\mathbf{y}}_i^C$ and $\hat{\mathbf{y}}_i^U$ denote the outputs based on task relevant and spurious features, respectively, for the input post t_i, and $\hat{\mathbf{Y}}^C$ and $\hat{\mathbf{Y}}^U$ denote their vectorised forms for a batch of textual inputs ($\mathcal{B}$).

4.5.d Adjustment

Adjustment refers to selecting appropriate optimisation strategies for achieving the mathematical objective of MDCNET, as defined in Equation 4.12. This adjustment mechanism is essential, as contamination from spurious features in the final representation may result in misdetecting contributors to mental disorders and stress, thereby posing potential life-threatening risks to users' mental state. The objective function integrates multiple optimisation goals into a unified

learning paradigm, represented via the total loss function:

$$\mathcal{L} = \beta_1 \mathcal{L}_{\text{sup}} + \beta_2 \mathcal{L}_{\text{conf}} + \beta_3 \mathcal{L}_{\text{idp}} \tag{4.20}$$

where $\mathcal{L}_{\text{sup}}$ encourages correct predictions based on task relevant features, $\mathcal{L}$conf enforces incorrect predictions from confounder features, and $\mathcal{L}_{\text{idp}}$ preserves the disjointedness between task relevant and spurious features. The total loss function is directly derived from the objective function in Equation 4.12. The weights β_1, β_2, and β_3 are hyperparameters that balance the contributions of the individual loss terms, subject to $\beta_1 + \beta_2 + \beta_3 = 1$. $\mathcal{L}_{\text{sup}}$ denotes the cross-entropy loss between the task relevant prediction $\hat{\mathbf{Y}}^C$ and the ground-truth labels $\mathbf{Y}_{true}$, which ensures that task relevant features of the input tokens contribute directly to identifying true contributors to mental health challenges. In contrast, $\mathcal{L}_{\text{conf}}$ corresponds to the negative error rate between the predictions based on confounder features ($\hat{\mathbf{Y}}^U$) and ground truth labels ($\mathbf{Y}_{true}$). Using the negative sign with the loss obliges the optimiser to maximise error for spurious features during the total loss minimisation, thereby preventing them from contributing to accurate predictions of mental health contributors. However, directly employing a negative sign may lead to training instability and divergence of the total loss toward negative infinity. To mitigate this risk, we regulate $\mathcal{L}_{\text{conf}}$ as defined in Equation 4.21:

$$\mathcal{L}_{\text{conf}} = \max\left(0, \rho - \left\|\hat{\mathbf{Y}}^U - \mathbf{Y}_{true}\right\|\right), \tag{4.21}$$

where ρ is a hyperparameter that regulates the limit of loss, and $\|\cdot\|$ denotes a vector norm (e.g., ℓ_1 or ℓ_2). Furthermore, to enforce independence between task relevant and confounder features, we employ the Hilbert–Schmidt Independence Criterion (HSIC) [88] as a loss function, which enables the detection of dependencies that may simultaneously arise between task relevant and spurious contexts:

$$\mathcal{L}_{\text{idp}} = \text{HSIC}(\mathbf{H}_C, \mathbf{H}_U) = \frac{1}{\left(|\mathcal{B}| - 1\right)^2} \text{tr}(\mathbf{K}_{\mathbf{H}_C} \mathbf{H} \mathbf{K}_{\mathbf{H}_U} \mathbf{H}) \tag{4.22}$$

where $|\mathcal{B}|$ denotes the batch size, $\text{tr}(\cdot)$ represents the trace of a matrix, and $\mathbf{K}_{\mathbf{H}_C}$ and $\mathbf{K}_{\mathbf{H}_S}$ correspond to the kernel matrices (e.g., Gaussian kernels) of the task relevant and confounder feature matrices, $\mathbf{H}_C$ and $\mathbf{H}_S$, respectively. The centring matrix is defined as $\mathbf{H} = \mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^T$, where $\mathbf{I}$ denotes the identity matrix and $\mathbf{1}$ is a column vector of ones. Consequently, irrelevant features to true contributors are effectively separated, thereby reducing the risk of predictions being distorted by spurious or biased inputs.

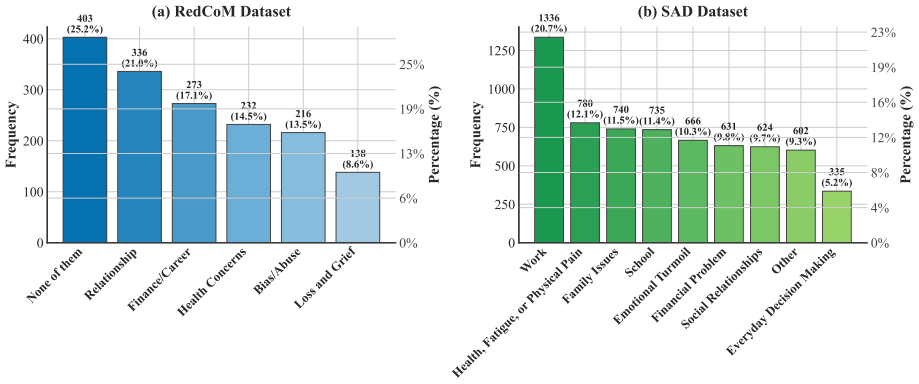


Figure 4.3: Label distributions of datasets

4.6 Experimental Results

4.6.a Dataset

In this study, we conducted and validated experiments using two benchmark datasets: SAD [153] and a newly constructed dataset, RedCoM. The SAD dataset consists of 6,850 SMS-style sentences annotated with nine categories of daily stressors. The categories derived from stress management literature, chatbot interactions, crowdsourcing, and web scraping, were validated by human annotators to determine the final label. Each sentence was assigned a primary label through a majority voting procedure.

The RedCoM dataset includes publicly available OSM posts associated with mental disorders, labelled with potential contributors to these disorders. The label distributions of both datasets are shown in Figure 4.3.

4.6.b Experimental Settings

The MDCNet model was implemented using PyTorch 2.4.0 with CUDA 12.1 support. Training was performed with 5-fold cross-validation to ensure the reliability of the results. All experiments were conducted on the *Viking HPC cluster* provided by the University of York. Optimal hyperparameters were selected through Bayesian optimisation using Optuna to maximise the validation F_1-score. To ensure fairness, the maximum score is reported achieved by each model across: i) the original hyperparameters reported in the respective research papers, ii) the optimal hyperparameters of MDCNet, and iii) supplementary manual tuning in cases where the first two conditions yielded suboptimal performance. All models were trained with the Adam optimiser (default parameters) and a linear learning-rate scheduler. The detailed hyperparameter settings

are reported in Table 4.3.

Table 4.3: Experimental settings for MDCNET model

Parameter	Description	SAD	REDCOM
Φ_{enc}	Pre-trained encoder	`mental-roberta-base`	`mental-bert-base-uncased`
β_1	Weight controls $\mathcal{L}_{\text{sup}}$	0.35	0.52
β_2	Weight controls $\mathcal{L}_{\text{conf}}$	0.38	0.30
β_3	Weight controls $\mathcal{L}_{\text{idp}}$	0.27	0.30
$hidden\_size$	Hidden size of the model	64	128
$\lvert\mathcal{B}\rvert$	Batch size	16	16
d_p	Dropout rate	0.40	0.40
η	Learning rate	7.0×10^{-4}	7.5×10^{-4}
K_p	Fraction of features selected by topK	0.21	0.20
ϕ	Conv1D kernel size	2	2
ρ	Limit controller for $\mathcal{L}_{\text{conf}}$	0.4	0.2

4.6.c Evaluation Metrics

To evaluate model performance, we employ standard macro averaged metrics: *macro-precision*(P), *macro-recall*(R), and *macro-F_1 score* (F_1). These metrics are particularly appropriate for evaluating performance fairly in class imbalanced scenarios, as they assign equal significance to all classes regardless of their frequency. Macro-averaged metrics are formally defined as follows:

$$\text{P} = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FP_i} \quad \text{R} = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FN_i} \quad \text{F}_1 = \frac{1}{C}\sum_{i=1}^{C}\frac{2\cdot \text{P}_i \cdot \text{R}_i}{\text{P}_i + \text{R}_i} \tag{4.23}$$

where C denotes the number of classes, and TP_i, FP_i, and FN_i represent the true positives, false positives, and false negatives for class i, respectively. We adopt one-vs-all approach for multi-class classification in order to define the TP_i, FP_i, and FN_i values.

Moreover, we introduce the **Disentanglement Score (DS)** to assess the effectiveness of our causality inspired attention mechanism that differentiates between task relevant and spurious features. DS quantifies the relative performance difference between predictions based on task relevant and spurious features:

$$\text{DS} = \frac{\text{F}_{1,tr} - \text{F}_{1,sp}}{\text{F}_{1,tr}} \tag{4.24}$$

where $\text{F}_{1,tr}$ and $\text{F}_{1,sp}$ denote the F1-scores obtained using solely task-relevant and spurious features, respectively. A higher DS reflects a more effective disentanglement of task relevant and spurious features by the causality inspired contextual attention mechanism.

4.6.d Performance Evaluation

Table 4.4: Performance comparison of MDCNET against baseline models across datasets

	RedCoM			SAD		
Model	**F_1**	**P**	**R**	**F_1**	**P**	**R**
Traditional Machine Learning						
LR	0.427	0.689	0.418	0.628	0.646	0.630
SVM[153]	0.501	0.703	0.497	0.643	0.683	0.641
Neural Networks						
LSTM	0.673	0.692	0.664	0.647	0.670	0.656
BiLSTM	0.669	0.680	0.662	0.656	0.658	0.657
GRU	0.668	0.698	0.654	0.655	0.660	0.655
BiGRU	0.656	0.677	0.645	0.669	0.670	0.667
CNN	0.706	0.728	0.693	0.683	0.683	0.685
CNN-LSTM[79]	0.639	0.680	0.622	0.649	0.657	0.646
CNN-GRU	0.655	0.680	0.639	0.662	0.673	0.661
Pre-trained Language Models						
BERT-base[64]	0.503	0.524	0.534	0.655	0.663	0.696
RoBERTa-base[287]	0.533	0.649	0.575	0.689	0.705	0.716
Mental-BERT[113]	0.675	0.692	0.684	0.677	0.688	0.716
Mental-RoBERTa[113]	0.625	0.645	0.647	0.669	0.690	0.710
XLNet[273, 121]	0.347	0.458	0.409	0.383	0.402	0.426
Mistral-7B[115]	0.203	0.322	0.244	0.607	0.673	0.589
MentalLLaMA-7B[272]	0.563	0.652	0.569	0.471	0.685	0.459
GPT-3.5-Turbo[176]	0.622	0.683	0.702	0.693	0.682	0.693
GPT-4o-Mini[177]	0.628	0.686	0.688	0.695	0.695	0.707
Qwen-14B[239]	0.719	0.750	0.702	0.632	0.681	0.616
MDCNet (Ours)	**0.750**	**0.757**	**0.738**	**0.711**	**0.719**	**0.749**

Note: P = Precision, R = Recall.

We evaluated the performance of the proposed MDCNET model against 19 baseline and state-of-the-art models by classifying input texts into nine stressor categories defined in the SAD dataset and six mental disorder contributors included in the REDCOM dataset.

Compared models involve two classical machine learning models (SVM and LR), seven traditional neural network architectures (CNN, LSTM, GRU, BiLSTM, BiGRU, CNN-LSTM, CNN-GRU), and ten transformer-based pre-trained language models ($\text{BERT}_{\text{BASE}}$ [64], $\text{RoBERTa}_{\text{BASE}}$ [287], $\text{Mental-BERT}_{\text{BASE}}$ [113], $\text{Mental-RoBERTa}_{\text{BASE}}$ [113], XL-Net [273], GPT-3.5 Turbo [176], GPT-4 Mini [177], MentalLLaMA [272], Mistral 7B Instruct [115], Qwen 14B [239]). Particularly,

three of the baseline models (SVM [153], CNN-LSTM [79], and XLNet [121]) have been employed as baseline models because of their strong performance in similar works. For traditional machine learning and deep learning models, Word2Vec embeddings were employed to represent textual expressions in OSM and CS.

Table 4.4 demonstrates a detailed comparison of model performance, highlighting the strong capabilities of MDCNet across both datasets. MDCNet achieves remarkable macro F_1 scores of 0.7110 and 0.7501 on the SAD and RedCoM datasets, respectively, surpassing all baselines. These results underscore the model's ability to capture complex and latent characteristics of the contributors to mental health conditions and stress. While an F_1 score of $\sim$0.70-0.75 might not appear as a promising result in the healthcare domain, it surpasses the baselines in this high cardinality multi-class classification task, significantly exceeding the random baseline of $\sim$9–16%. In Table 4.5 we present some instances of mispredictions, and it appears that the main sources of incorrect predictions are having multiple contributors simultaneously and possessing semantic overlaps. Although outcomes are promising, considering the criticality of the domain, we believe that further expansion of the dataset would be needed to improve the model performance before the model can be used in practice to reduce the workload of mental health practitioners and identify high-risk cases.

Traditional machine learning models show poor performance compared to more advanced methods, achieving macro F_1 scores around 0.5, particularly on the RedCoM dataset. On the SAD dataset, their performance is comparatively more reliable, likely due to the shortness and simplicity of the texts, in contrast to the longer and more complex textual posts in RedCoM. Traditional Recurrent Neural Network (RNN) and CNN based, and hybrid neural network models, outperform traditional models and, surprisingly, transformer encoders such as BERT-base and RoBERTa-base, achieving F_1 scores of 0.7056 and 0.6829 on the RedCoM and SAD datasets, respectively. Nevertheless, hybrid CNN-RNN models do not yield significant improvements in the current experiments, while they demonstrated promising performance in previous studies [79]. Fine-tuned versions of BERT-base and RoBERTa-base (Mental-BERT and Mental-RoBERTa) outperform their base models, as expected due to their specialisation for mental health related datasets. Despite its transformer based nature, XLNet yields relatively poor performance, suggesting that permutation based language models may not effectively capture contributors to mental disorders and stress. Contrary to expectations, generative AI models do not always perform well compared to the traditional models. Particularly, Mistral 7B and Mental-LLaMA 7B perform significantly worse than many traditional models, especially in F_1

Table 4.5: Sample instances of wrong predictions

Input Text	**GT**	**Pred.**
"Every time I hear/read/talk about someone who killed themselves my only thought is "They're lucky. Why couldn't that be me?"."	None of them	Loss and Grief
"... When I relapsed on alcohol everything goes to shit. I could be going to jail. I want to die. I'm 30 and it gets worse and worse and worse. Sure there's small spaces of joy but it's not worth it. I just want to end this suffering. I've been in love and I've had some good times. But I don't want to go through this life any longer. I also have a physical disability. I'm just a little scared if death but we are going to die either way. I'm scared of pain but I just want this life to end."	Health Concerns	None of them
"No friends whatsoever. Living in isolation... Please help me I have severe social anxiety. Every day is almost total isolation. I go to class. I fantasize about working up the courage to talk to someone or say something in class but never do. I go back to my apartment. I am too afraid to talk to them. Day in. Day out. Rinse. Repeat. This has been my life for several years now. I'm incredibly insecure. I have some combination of body dysmorphia and an eating disorder. ... If someone could just be there for me I'd feel better. But I've given up almost all hope...."	Relationship	Bias/Abuse
"Now I am listening to Pink Floyd's The Wall album on opioids and contemplating on what went wrong. Of course I can live without you. But please remember there was a moment in my life when you made me feel like I can't. *sad reacts only* "	Relationship	None of them

and Recall, likely due to their relatively small volume. Although GPT-3.5 Turbo and GPT-4 Mini demonstrate superior performance among generative models, with evaluation scores consistently above 0.6, their adoption poses challenges, including financial costs and ethical concerns, such as closed-source deployment that obscure their internal mechanisms. Qwen 14B surpasses all models except MDCNet on the RedCoM dataset, with the evaluation scores exceeding 0.7, while slightly underperforming compared to GPT models on the SAD dataset. Despite its open-source nature, Qwen 14B's large volume requires substantial computational resources, leading to significant financial costs.

Overall, MDCNet yields consistent and robust performance across both datasets, highlighting its effectiveness in revealing task-relevant features of contributors to mental health challenges.

4.6.e Ablation Study

We conducted an ablation study to evaluate the contributions of key components of MDCNet by selectively omitting them, as summarised in Tables 4.6 and 4.7. The complete MDCNet model consistently achieved superior results across all metrics and datasets, with macro F_1 scores of 0.7501 and 0.7110 for RedCoM and SAD, respectively, validating the effectiveness of the integrated approach. Dis-

Table 4.6: Ablation study of the MDCNet model on the RedCoM dataset

Config.	F_1	DS	P	R
MDCNet	0.7501 ± 0.0181	0.9348 ± 0.0155	0.7570 ± 0.0171	0.7375 ± 0.0154
w/o LIWC	0.7198 ± 0.0210	0.9343 ± 0.0197	0.7424 ± 0.0112	0.7156 ± 0.0202
w/o $\mathcal{L}_{\text{idp}}$	0.7222 ± 0.0326	0.9398 ± 0.0167	0.7333 ± 0.0328	0.7156 ± 0.0325
w/o $\mathcal{L}_{\text{sup}}$	0.6933 ± 0.0259	0.9086 ± 0.0379	0.7533 ± 0.0168	0.7000 ± 0.0190

abling LIWC features led to a substantial performance drop across both datasets (2–4% in macro F_1), underscoring the importance of psycholinguistic features for identifying contributors to mental health conditions. The independence loss component exhibited dataset-specific effects as its removal caused a 6% decrease in DS for the SAD dataset, indicating an impaired capacity to separate task relevant and spurious features. In contrast, spurious feature suppression was critical for the RedCoM dataset, as its removal led to a 6% decline in macro F_1 and a 3% reduction in DS. This effect may be due to the longer and informative nature of Reddit posts that are more prone to noisy or irrelevant information. The results further reflect the unique characteristics of the two datasets, as RedCoM comprises longer social media posts and SAD consists of short, conversational inputs. Despite the performance drop due to component removal, the model maintained

Table 4.7: Ablation study of the MDCNET model on the SAD dataset

Config.	F_1	DS	P	R
MDCNET	0.7110 ± 0.0119	0.9687 ± 0.0494	0.7191 ± 0.0121	0.7488 ± 0.0097
w/o LIWC	0.6931 ± 0.0142	0.9667 ± 0.0272	0.7052 ± 0.0057	0.7403 ± 0.0081
w/o $\mathcal{L}_{\text{idp}}$	0.6940 ± 0.0118	0.9080 ± 0.0326	0.7161 ± 0.0164	0.7372 ± 0.0102
w/o $\mathcal{L}_{\text{sup}}$	0.6710 ± 0.0187	0.9651 ± 0.0152	0.7029 ± 0.0165	0.7263 ± 0.0091

relatively high DS scores (above 0.90) across all configurations, indicating that the attention module effectively identifies the task relevant features. Overall, the findings demonstrate that each component has a unique effect on model's effectiveness, and their integration is essential for achieving high performance across diverse types of mental health challenges. We performed paired t-tests for the significance analysis of the ablation studies, which proves that the MDCNET model significantly outperforms all ablated configurations on the RedCoM dataset (all $p < 0.05$), and significantly outperforms the w/o LIWC and w/o $\mathcal{L}sup$ configurations on the SAD dataset ($p < 0.05$). Although the model outperforms the w/o $\mathcal{L}idp$ configuration on SAD, this difference is not statistically significant.

4.6.f Sensitivity Analysis

Figure 4.4 illustrates the effects of the parameters β_1, β_2, and β_3, corresponding to the weights of $\mathcal{L}$sup, $\mathcal{L}$conf, and $\mathcal{L}_{\text{idp}}$, respectively, varied within the range 0.1–0.5. Steady performance gains are observed across both datasets with

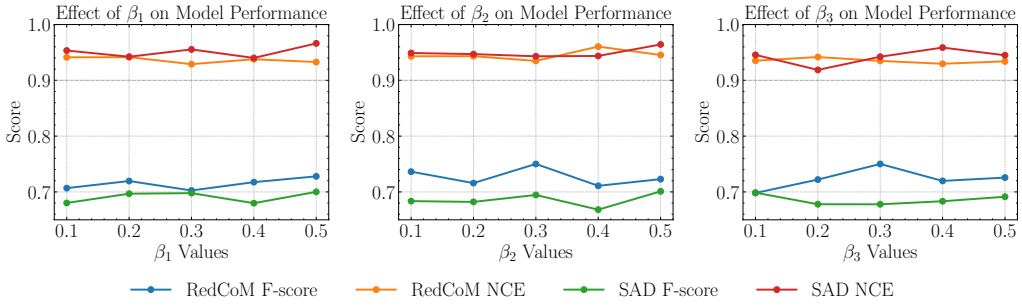


Figure 4.4: Parametric analysis of loss coefficients (β_1, β_2, and β_3) on model performance for REDCOM and SAD datasets

increasing values of the β_1 coefficient. In REDCOM, the F_1-score improves consistently from $\beta_1 = 0.1$ to $\beta_1 = 0.5$, while the DS remains high (> 0.92). This indicates that the supervised learning component contributes effectively to both classification and feature separation. Particularly, REDCOM exhibits robust performance across all values of β_1, with insignificant change in DS, reflecting its

stability during supervision. In contrast, the SAD dataset shows a more noticeable reaction to increasing β_1, as the F_1-score improving from $\beta_1 = 0.1$ and peaking at $\beta_1 = 0.35$, before slightly declining at higher values, likely due to early overfitting caused by short text inputs. Meanwhile, DS steadily increases, indicating that stronger supervision enhances both feature separation and classification.

The β_2 parameter, regulating spurious feature suppression, exhibits dataset specific performance characteristics. Both RedCoM and SAD demonstrate consistent increases in DS as β_2 rises, with optimal values at $\beta_2 = 0.4$ (DS=0.9605) and $\beta_2 = 0.5$ (DS=0.9643), respectively. This pattern indicates that stronger suppression improves the distinction between task relevant and spurious features. However, F_1-scores demonstrate dataset dependent effects. RedCoM achieves peak performance at $\beta_2 = 0.18$ (F_1=0.7501), while SAD gains its optimal F_1-score at $\beta_2 = 0.38$, indicating that intense suppression of spurious features may negatively affect classification, even though DS benefits from stronger suppression.

The independence loss parameter β_3 exhibits dataset specific effects, as well. Both datasets exhibit unsteady performance patterns. In RedCoM, the best F_1-score is achieved at $\beta_3 = 0.3$ (F_1=0.7501), while DS values remain consistently high (>0.95), suggesting that enforcing feature independence enhances classification performance without significantly affecting feature separation. SAD demonstrates similar sensitivity, reaching optimal performance at $\beta_3 = 0.27$ (F_1=0.7110). This sensitivity indicates that feature distinction enforcement should be carefully applied to avoid the distortion of feature learning in simpler datasets such as SAD.

Overall, these results underscore the necessity of domain specific parameter tuning to optimise the balance between supervised learning, spurious suppression, and feature independence.

4.7 Summary

In this chapter, we introduce the RedCoM dataset and the MDCNet model for analysing potential contributors to mental disorders and stress through textual expressions on OSM and CS. This dual contribution allows exploration of contributors to mental health challenges from both data and model based perspectives. The RedCoM dataset covers a broad range of mental health related categories by incorporating OSM posts associated with diverse mental health challenges. The labelling process integrates automated methods with human-in-the-loop strategies, and verified through multiple validation stages. Independent annotators (trained graduate students) verified labelling consistency,

confirming the dataset's reliability, validity, and domain relevance. From the modelling perspective, MDCNet employs a causality inspired contextual attention mechanism to reveal latent task relevant structures linked to mental disorders and stress. It outperforms state-of-the-art benchmark models and demonstrates robust consistency across all evaluation metrics and datasets. MDCNet's contextual attention, domain knowledge integration, and confounder mitigation components are particularly effective for modelling complex and latent patterns. Although interpretability is not directly maintained and not evaluated, the integration of task relevant and attention based mechanisms provides a promising foundation for future research on explainable and trustworthy AI in mental health. Looking ahead, extending MDCNet to a multi-label classification setting shows promise for enabling more fine-grained and reliable inferences. Our long-term objective is to integrate these advancements into thorough mental disorder analysis systems to enhance the accuracy and context-sensitive screening of mental health conditions.

5 Optimising Explainability in Mental Health Risk Identification

5.1 Introduction

Artificial Intelligence (AI) models enable successful data-driven decision making by leveraging large volumes of data [8]. In recent years, there has been rapid progress [282] in developing novel data-driven solutions for mental healthcare, including early diagnostics [166], AI-driven therapy [59], and monitoring through routine screening [48]. Meanwhile, Online Social Media (OSM) provides valuable resources for detecting mental disorders, such as depression [290, 166], eating disorders [2, 1], and suicide ideation [167, 215] using AI. Moreover, complex Deep Learning (DL) models, including Long Short-Term Memory (LSTM)s [102], Gated Recurrent Unit (GRU)s [50], and Convolutional Neural Network (CNN)s [110], have become effective tools for identifying mental health conditions. Although these algorithms often achieve impressive performance, their complex internal mechanisms prevent transparent reasoning about their decision-making process, called a **black-box** nature of AI models [8]. However, adopting black-box models in real-world applications, particularly in healthcare, raises significant ethical concerns regarding transparency, fairness, trustworthiness, and safety [114, 163].

Computational approaches for mental health risk prediction offer benefits such as identifying risky behaviours, defining sensitive groups for prioritization, and flagging who needs human attention, such as therapists or hospitalization [43]. However, since explicit expert systems are not feasible for complex social media data, supervised AI systems are adopted for the risk prediction. We must acknowledge the serious limitations and consequences of using AI based systems with social media data. First, as mentioned earlier, it is often impos-

sible to identify more than one symptom within a single post, which makes any form of diagnosis impossible based on a single social media publication [164]. In addition, the quality of the post labels requires strong and careful inspection [43]. Moreover, the risk of the looping effect is a serious consequence of adopting such models, as these classifications can influence users' behaviour and cognition, further affecting their mental states [92]. Furthermore, exposing important features, such as specific words, may cause users to avoid them, potentially worsening their condition by suppressing their mental states. To alleviate these issues, we suggest using a history of posts within a defined timeframe rather than relying on a single post, as this may increase accuracy. Regarding stakeholders, detailed explanations should be directed to clinicians (especially in high-risk cases) so they can confirm whether human assessment is genuinely needed. These insights should remain strictly between users and clinicians and should not be disclosed to third parties. Therefore, it is essential to develop Explainable Artificial Intelligence (XAI) methods that provide stakeholders (mental health practitioners) with insights into the working mechanisms and decision logic of black-box models [163], improving trustworthiness and mitigating safety concerns [114, 8]. Despite its importance, explainability often remains underexplored in AI-based mental-health assessment. Kelly et al. [123] and Liu et al. [138] developed interpretable machine-learning approaches using logistic regression for predicting mental health treatment and depression. Furthermore, Kerz et al. [124], Saxena et al. [219], and Alghazzawi et al. [6] have developed machine learning models for analysing mental disorders using social media data, employing Local Interpretable Model-agnostic Explanations (LIME) [204] and SHapley Additive exPlanations (SHAP) [147] to explain model predictions. Moreover, attention based methods have also been involved for explainable mental health screening [9, 167, 292]. However, their effectiveness on explainability remains debatable. The explainability performance is rarely evaluated quantitatively, and, importantly, none of these methods incorporated mental health-related domain knowledge during the prediction and generation or approximation of explanations. Further details on XAI methods for Mental Health (MH) can be found in Chapter 2.4.

Our contributions. We introduce MentalXAI, a novel model for explainable mental disorder detection using social media data. It generates mental health aware cues regarding its decision-making process through cross-attention mechanisms coupled with attention adaptors. Our contributions include:

1. We propose XAI concepts specifically tailored for MH, providing them as desiderata that discuss the common explainability challenges by addressing the problematic approaches.

2. We provide a comprehensive theoretical framework with simplified mathematical justifications supporting the suitability of the proposed method for explainability.
3. We develop an advanced deep learning model, MentalXAI, which generates faithful, robust and MH-aware explanations employing knowledge-infused cross attention and the attention adaptor component that can be plugged in any deep learning model.
4. We conduct extensive experiments to evaluate detection and explanation performance against competitive baselines and component configurations.
5. We also analyse the psycholinguistic relevance of the explanations generated by our model, a key requirement for XAI in mental health

The remainder of this chapter is structured as follows. Section 5.2 defines the theoretical framework of XAI for MH. Section 5.3 introduces MentalXAI, Section 5.4 presents the experimental results, and finally, Section 5.5 concludes the chapter.

5.2 Theoretical Background of the XAI for MH

5.2.a What is Explainability and XAI?

According to the Oxford English Dictionary [1], the verb *to explain* is defined as: *to describe or give an account of in order to bring about understanding; to explicate; to provide details or elaborate on specifics*. In the AI domain, explainability is often described as the functional knowledge aimed at describing the model's *black-box behaviour*. The black-box nature of AI systems indicates the difficulty of producing interpretable reasoning for the model's decisions [8].

Although there is no universally accepted definition of explainability and XAI, we attempt to formalise it both verbally and mathematically. The principal objective of explainability is to uncover black-box systems and illuminate the underlying processes of Machine Learning (ML) models [114, 41]. There are several methods for achieving explainability, including causal chains, feature importance measures, and weight distributions [159]. In this study, we define **explainability** as the identification of a subset of key features that play a critical role in a model's decision making process. This is motivated by the fact that the extraction of important features can be easily interpreted by non-technical stakeholders, such as mental health practitioners. Therefore, to the best of our knowledge, a mechanism can be considered explainable if it satisfies the following criteria:

[1] https://www.oed.com/

1. It reflects the influence of individual features on the model's output (**feature importance**). These key features constitute the **explanations**.
2. Explanations must be **faithful**: the model's performance should remain similar when relying solely on these explanations.
3. The rest of the features (unimportant features - non-explanatory) should contribute minimally to the model's performance.
4. Explanations must be **robust** to small perturbations: minor perturbations applied to explanatory features should significantly affect model performance, whereas perturbations to non-explanatory features should not have significant impact.

Let $f : \mathbb{R}^{L\times d} \to \{0, 1\}$ be a neural network based detection model trained for a binary classification task. For an input $\mathbf{x} \in \mathbb{R}^{L\times d}$, where L is the sequence length and d is the feature dimension, $\mathbf{x}_I$ is the important feature matrix (explanations), and $\mathbf{x}_U$ is the matrix of unimportant features. Let $\delta(\cdot)$ be the function that returns perturbations with standard deviation of s that obey Gaussian distribution ($\delta \sim \mathcal{N}(0, s)$). Define $\mathbf{x}'_I = \mathbf{x}_I + \delta(\mathbf{x}_I)$ and $\mathbf{x}'_U = \mathbf{x}_U + \delta(\mathbf{x}_U)$ which indicates the inputs with the perturbed explanations and perturbed unimportant features respectively.

Definition 5.2.1 (Explainability for Mental Health Assessment (MHS))**.** *A model f is explainable for MHS if it satisfies:*

$$
\begin{aligned}
&\|f(\mathbf{x}) - f(\mathbf{x}_I)\|_2 \ll \epsilon_1 \quad \wedge && \text{(Sufficiency)}\\
&\|f(\mathbf{x}) - f(\mathbf{x}_U)\|_2 \gg \rho_1 \quad \wedge && \text{(Comprehensiveness)}\\
&\|f(\mathbf{x}) - f(\mathbf{x}_I \oplus (\mathbf{x}_U + \delta(\mathbf{x}_U)))\|_2 \ll \epsilon_2 \quad \wedge && \text{(Robustness)}\\
&\|f(\mathbf{x}) - f((\mathbf{x}_I + \delta(\mathbf{x}_I)) \oplus \mathbf{x}_U)\|_2 \gg \rho_2 && \text{(Robustness)}
\end{aligned}
\tag{5.1}
$$

where (ϵ_1, ϵ_2) and (ρ_1, ρ_2) are lower and upper boundaries of the tolerance margins respectively.

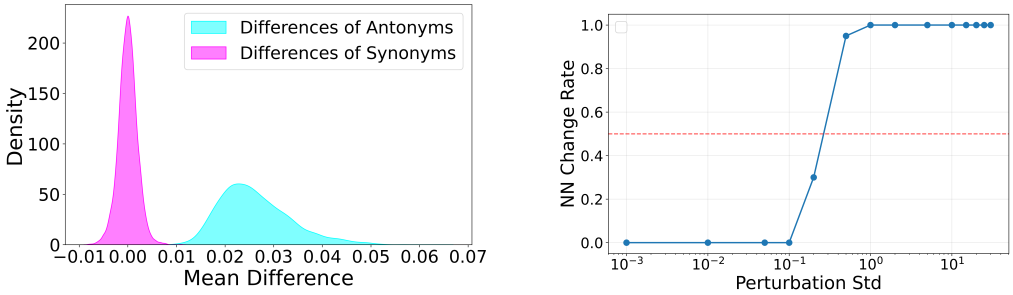
This definition can be applied to any domain. However, in this chapter, we focus on explainability in the MH domain and therefore extend the definition to be specialised for MH.

5.2.b Desiderata of Explainability in MH

Issue of current definitions

In Section 5.2.a, we presented a general definition of explainability and highlighted that, in terms of robustness, explanations should reveal stability under

small perturbations. In Natural Language Processing (NLP) tasks, particularly in mental health analysis based on textual social media posts, it is essential to account for the sensitivity of word embeddings to such perturbations. First,



(a) Distribution of synonyms and antonyms

(b) Word meaning changes across perturbation standard deviation

Figure 5.1: Semantic analysis of the word embedding vectors

Figure 5.1a presents the kernel density estimation plots for the mean of the embedding differences of synonym and antonym pairs. It can be observed that there is no substantial gap between the deviation in differences of synonym and antonym pairs, suggesting that small perturbations may not alter a word's meaning. Consequently, requiring such perturbations to affect the model's decision may raise a logical issue (if the meaning of the word does not change significantly, then why should the outcome change?). To further support this observation, we plotted the proportion of words whose meaning changes against the standard deviation of perturbations drawn from a normal distribution in Figure 5.1b. According to the figure, adding random noise from a Gaussian distribution with a standard deviation of 1 is sufficient to alter the meaning of words. However, values lower than 1 may preserve the semantic meaning of the vector. This result supports that adding small noise to either important (unimportant) words does not alter their semantic meaning, and that requiring such changes to influence (or not influence) the model's decision is inaccurate. Therefore, instead of adding small perturbations to words, we inject significant Gaussian noise to maintain robustness.

Updated Conditions for XAI in MH

We outline the following desiderata to ensure that AI models for MHS are transparent, fair, interpretable, and robust. A candidate model can be considered explainable for MHS if and only if it satisfies the following criteria:

1. **The model is intrinsically explainable:** Black-box models are frequently employed across many fields, including those requiring high ethical stan-

dards such as safety, transparency, and trustworthiness, as in healthcare. To bypass these ethical requirements, post-hoc explainability methods have often been employed. However, the adoption of such models may raise societal risks, as the main requirements for explainability have not yet been fully explored. Moreover, these models often fail to provide true reasoning for their outcomes, instead approximating feature importance using simple surrogate models rather than offering insights into the model's internal decision making [207]. *Therefore, we require the model to generate explanations intrinsically, without the need for additional training or auxiliary post-hoc models.*

2. **Explanations are *faithful* to model behaviour and *robust* to perturbations:** These two properties constitute major requirements for general explainability [265]. The main objective of explainability is to highlight the decision making patterns of the model. Therefore, the generated explanations should produce similar outputs to those from the original inputs, thereby demonstrating their importance in the decision-making process. In contrast, features not identified as explanations should lead to degraded performance and significantly different outputs. Moreover, perturbations to important input features should significantly affect the model's output, whereas perturbations to unimportant features should have minimal impact, indicating the robustness of the explanations. *Therefore, we require the model to demonstrate balanced performance, generating faithful and robust explanations while also achieving high accuracy.*

3. **The model incorporates domain knowledge for decision and explanation:** In real-world scenarios, clinical training and experience are mandatory for the mental health based assessments [19]. However, current methodologies in the literature often rely heavily on statistical learning while underexploring domain knowledge. Although extracted explanations may reflect patterns relevant to the model's decision-making process, it is essential to ensure that they are generated with consideration of domain knowledge in mental health. *Therefore, we require the model to incorporate domain knowledge for both detection and explanation.*

4. **Explanations possess psycholinguistic relevance:** Various linguistic signals are significantly correlated with mental health challenges. Psycholinguistic features can effectively predict mental health issues [173]. Moreover, psycholinguistic patterns possess the advantage of identifying the causes and stressors of mental disorders [90]. *Therefore, explanations should align with psycholinguistic cues, as professional assessment often relies on detecting psychological signals in language.*

Overall, any model that claims to be explainable in MH settings must be intrinsically explainable. The explanations should be faithful and robust, based on domain knowledge for both decision making and explanation, and aligned with psycholinguistic insights.

5.2.c Attentions can explain!

We previously defined explainability as the ability to generate important features, and the set of these features is referred to as explanations. To model explanation generation, we employ the attention mechanism. However, the explainability potential of attention remains a topic of debate [111]. Therefore, we first provide a simple demonstration that the attention mechanism has the potential to explain the model, prior to presenting empirical evidence supporting this claim.

Lemma 5.2.1 (Attentions can explain!)**.** *Let $\mathcal{X} = \{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^d$ be an ordered set of feature vectors, where each $x_i \in \mathbb{R}^d$. We define the input matrix $\mathbf{X} = \Phi(\mathcal{X}) \in \mathbb{R}^{n \times d}$, where $\mathbf{X}_{i,:} = x_i$. Let $\mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_n\} \subset \mathbb{R}$ denote the associated set of attention weights with $\alpha_i \geq 0$, and define $\mathbf{\Lambda} = \Phi(\mathcal{A}) \in \mathbb{R}^n$. Assume the attention function $\alpha : \mathbb{R}^d \to \mathbb{R}$ is linear, i.e., $\alpha_i = \alpha(x_i) = \omega^\top x_i$ for trained $\omega \in \mathbb{R}^d$. We define a simple differential attention based decision function with a trained parameter vector $\theta \in \mathbb{R}^d$:*

$$y = f_{\theta,\mathbf{\Lambda}}(\mathbf{X}) = \sum_{i=1}^{n} \theta^\top(\alpha_i x_i),$$

Then, $\mathbf{\Lambda}$ has the potential to indicate feature importance for explanation purposes.

Proof. Let $\mathcal{X}_I \subset \mathcal{X}$ be the non-empty set of important features (explainabilities) and $\mathcal{X}_U \subset \mathcal{X}$ be the non-empty set of unimportant features, where $\mathcal{X}_I \neq \varnothing$,$\mathcal{X}_U \neq \varnothing$, $\mathcal{X}_I \cap \mathcal{X}_U = \varnothing$ and $\mathcal{X}_I \cup \mathcal{X}_U = \mathcal{X}$:

$$x_i \in \mathcal{X}_I \iff |f_{\theta,\mathbf{\Lambda}}(\Phi(\mathcal{X})) - f_{\theta,\mathbf{\Lambda}}(\Phi(\mathcal{X} \setminus \{x_i\}))| = |\alpha_i(\theta^\top x_i)| > 0 \tag{5.2}$$

and the set of unimportant features:

$$x_j \in \mathcal{X}_U \iff |f_{\theta,\mathbf{\Lambda}}(\Phi(\mathcal{X})) - f_{\theta,\mathbf{\Lambda}}(\Phi(\mathcal{X} \setminus \{x_j\}))| = |\alpha_j(\theta^\top x_j)| = 0 \tag{5.3}$$

Now, let $(x_i, x_l) \in \mathcal{X}_I \times \mathcal{X}_I$, i.e., $|\theta^\top(\alpha_i x_i)| > 0$ and $|\theta^\top(\alpha_l x_l)| > 0$. For any $\lambda \in [0, 1]$, we define $x_m = \lambda x_i + (1 - \lambda)x_l$. Then:

$$\theta^\top(\alpha_m x_m) = \lambda\theta^\top(\alpha_m x_i) + (1 - \lambda)\theta^\top(\alpha_m x_l) \tag{5.4}$$

$$\alpha_m = \lambda\alpha_i + (1 - \lambda)\alpha_l$$

(Linearity of attentions)

$$\theta^\top(\alpha_m x_m) = \theta^\top((\lambda\alpha_i + (1 - \lambda)\alpha_l) \cdot (\lambda x_i + (1 - \lambda)x_l)) \tag{5.5}$$

$$|\theta^\top(\alpha_m x_m)| > 0 \tag{5.6}$$

$$x_m \in \mathcal{X}_I \tag{5.7}$$

Similarly, for $(x_j, x_k) \in \mathcal{X}_U \times \mathcal{X}_U$, where $|\theta^\top(\alpha_j x_j)| = 0$ and $|\theta^\top(\alpha_k x_k)| = 0$, define $x_n = \beta x_j + (1-\beta)x_k$ for $\beta \in [0, 1]$. Following the same derivation:

$$\theta^\top(\alpha_n x_n) = \theta^\top((\beta\alpha_j + (1-\beta)\alpha_k) \cdot (\beta x_j + (1-\beta)x_k)) \tag{5.8}$$

$$|\theta^\top(\alpha_n x_n)| = 0 \tag{5.9}$$

$$x_n \in \mathcal{X}_U \tag{5.10}$$

Thus, in the ideal settings, $\mathcal{X}_I$ and $\mathcal{X}_U$ are convex sets. By the Hyperplane Separation Theorem, there exists a hyperplane v that separates $\mathcal{X}_I$ and $\mathcal{X}_U$ as they are disjoint and non-empty convex sets. Moreover, attention values can approximately learn v as an attention mechanism is generally formulated as neural networks (by the Universal Approximation Theorem). Therefore, attention mechanisms can learn the separation boundary between important and unimportant features which **demonstrates their explanatory potential.** □

The proposed proof is a simple mathematical demonstration illustrating how the attention mechanism can distinguish important features, referred to as explanations, from unimportant features under ideal conditions. However, the model's architectural constraints may prevent the attention mechanism from focusing on these explanations, which requires further actions, such as explanation based optimisation.

5.3 Methodology

MentalXAI (Figure 5.2) is a component of the AttentionDep model (Figure 3.1) and primarily focuses on explainability-driven optimisation. The main objective of this study is not only to achieve high classification performance, but also to enhance explainability. Given a set of social media posts $P = \{p_1, p_2, \ldots, p_N\}$, the objective is to train a model $f_{\theta,\Lambda}(p_j) \rightarrow c_j$ by learning the model parameters θ and attention values Λ, where $p_j \in P$ and $c_j \in \{0, 1\}$.

5.3.a Textual Representation

Our model processes textual posts from social media platforms as input. For representing the textual posts, we employed a mechanism similar to that described in Chapter 3.3.a, performing tokenisation with NLTK and FastText [28] for initial text embeddings. However, understanding context remains essential for this study; therefore, we define the contextual embedding vector at position i using a one-dimensional convolution operation:

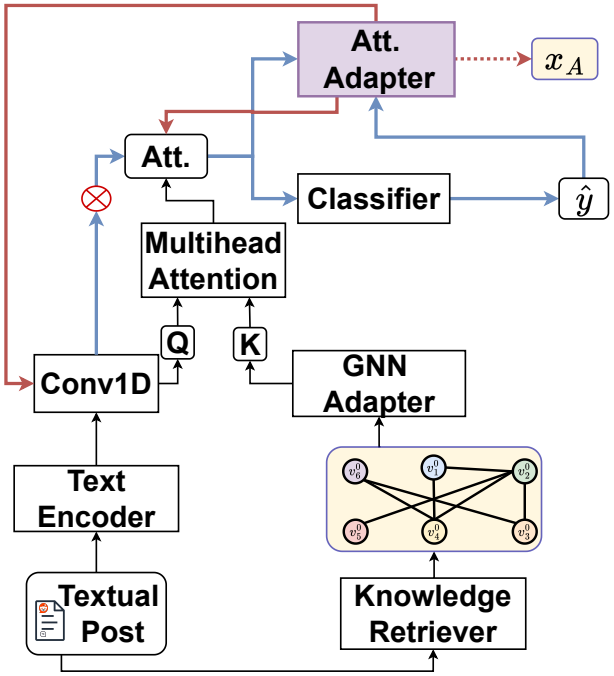


Figure 5.2: Proposed model MENTALXAI

$$\mathbf{h}_{i,j} \leftarrow \left(\mathbf{W} \cdot \begin{bmatrix} \mathbf{u}'_{i,j} \\ \mathbf{u}'_{i+1,j} \end{bmatrix} + \mathbf{b} \right), \quad i = 0, \ldots, n-1 \tag{5.11}$$

where $\mathbf{u}'_{i,j} \in \mathbf{H}'_j$ is the token embedding at position i of post p_j obtained by FastText text embedder, W and b are trainable parameters. The overall contextual text representation of post p_j can be estimated as:

$$\mathbf{B}_j \leftarrow \{\mathbf{b}_{i,j}\}_{i=1}^{L-1} \tag{5.12}$$

5.3.b Knowledge Representation

In this study, we adopt the Mental Health Knowledge Graph (MHKG) as defined in Chapter 3.3.b (Figure 3.2). Given the large number of MHKG triplets (6,433), which results in computationally expensive and complex representations, and to avoid the redundant triplets for each post, we extract a post-specific knowledge subset, as expressed in Equation 3.9. The retrieved MHKG for post p_j, denoted as $\mathcal{G}_j(\mathcal{V}_j, \mathcal{E}_j, \mathcal{C}_j, \mathcal{F}_j, \mathcal{A}_j)$, is first passed through the Graph Neural Network (GNN) adaptor described in Chapter 3.3.b, which consists of n consecutive

Graph Isomorphism Network with Edge features (GINE) layers (Equation 3.10), producing the knowledge matrix $\mathbf{G}_j \in \mathbb{R}^{|\mathcal{V}_j| \times d}$, where h denotes the hidden dimension.

5.3.c Knowledge Infusion

Although the attention mechanism can highlight important segments of the input, it is essential to ensure that attention values align with domain knowledge. For this purpose, we employ a cross attention mechanism:

$$\mathbf{A}_j^k \leftarrow \sigma\left(\frac{(\mathbf{H}_j\mathbf{\Theta}_{\mathbf{Q}}^k)(\mathbf{G}_j\mathbf{\Theta}_{\mathbf{K}}^k)^T}{\sqrt{d_k}}\right)(\mathbf{H}_j\mathbf{\Theta}_{\mathbf{V}}^k) \tag{5.13}$$

$$\mathbf{\Lambda}_j \leftarrow \bigoplus_{k=1}^{K}\left[\mathbf{A}_j^k\right] \times \mathbf{\Theta}_O \tag{5.14}$$

where $\mathbf{\Theta}_{\mathbf{Q}}^k$, $\mathbf{\Theta}_{\mathbf{K}}^k$, $\mathbf{\Theta}_{\mathbf{V}}^k$, and $\mathbf{\Theta}_{\mathbf{O}}$ are trainable parameters, while K indicates the number of heads for attention model. Moreover, $\mathbf{G}_j \in \mathbb{R}^{|\mathcal{V}_j| \times d}$ indicates the learnt embedding vectors of the domain specific knowledge graph with $|\mathcal{V}_j|$ nodes. Final knowledge infused post representation can be defined as:

$$\hat{\mathbf{y}}_j \leftarrow \sigma\left(\text{MLP}\left(\frac{1}{\|\mathbf{M}_j\|_1}\mathbf{1}^T(\mathbf{M}_j \odot \mathbf{B}'_j)\right)\right) \tag{5.15}$$

where $\mathbf{M}_j \in \{0,1\}^{L\times 1}$ represents binary masks indicating whether a given representation corresponds to a real input feature (1) or padding (0) which reduces the influence of padding during the decision making process and enables classification based on valid input features.

5.3.d Attention adaptor

The Attention adaptor is configured to encourage attention values to be explainable, i.e., capable of identifying features that are influential on the model's decisions. Important and unimportant features are defined as follows:

$$\mathbf{x}_I = \Phi(\{x_i \mid \alpha_i > \tau\}), \tag{5.16}$$

$$\mathbf{x}_U = \Phi(\{x_i \mid \alpha_i \leq \tau\}). \tag{5.17}$$

Here, $\mathbf{x}_A$ and $\mathbf{x}_B$ denote the *important features (explanations)* and the *unimportant features* respectively. Explanations are the top features ranked by attention values. However, selecting top-K indices may disrupt the computational graph due to differentiability constraints, therefore, we adopt sigmoid function with attention

threshold to maintain a fully differentiable computational graph. The threshold τ is defined as the $\hat{q}^{th}$ percentile of attention values within each batch, computed at every training iteration:

$$\tau \leftarrow \inf \left\{ \tau \in \mathbb{R} : \frac{1}{L} \sum_{l=1}^{L} \mathbf{1}_{\{a_l \leq \tau\}} \geq \hat{q} \right\} \tag{5.18}$$

where $\hat{q}$ configures the length of the generated explanations. Following the threshold calculation, soft mask values representing feature importance are computed using a sigmoid function with sharpness hyperparameter γ:

$$\mu = \frac{1}{1 + e^{-\gamma(\alpha - \tau)}} \tag{5.19}$$

Using μ, important and unimportant input features can be identified via element-wise multiplication. Perturbed versions of these features are generated by adding Gaussian noise $\delta \sim \mathcal{N}(0, s)$, where s represents the standard deviation of the noise distribution:

$$\mathbf{x}_I = \mathbf{x} \odot \mu, \quad \mathbf{x}_U = \mathbf{x} \odot (1 - \mu) \tag{5.20}$$

$$\mathbf{x}_I^{\delta} = \mathbf{x} + \delta \odot \mu, \quad \mathbf{x}_U^{\delta} = \mathbf{x} + \delta \odot (1 - \mu) \tag{5.21}$$

where $\odot$ denotes element-wise multiplication and s is set as a hyperparameter. Next step indicates the implementation of the loss functions for improving overall detection performance and attention adapting:

$$\mathcal{L}_1 = -\sum_{c=1}^{C} y_c \log \left(\frac{e^{\hat{y}_c}}{\sum_{j=1}^{C} e^{\hat{y}_j}} \right), \tag{5.22}$$

where C is the number of classes, and in our case $C = 2$, and y is the ground truth label. In the next phase, we ensure that the generated explanations $\mathbf{x}_I$ are faithful to the model outputs which means that important tokens (explanations) from the social media post should produce outputs similar to the original, whereas unimportant features should result in outputs that differ significantly from the original output, as represented below:

$$\mathcal{L}_2 = \|f_\theta(\mathbf{x}_I) - \hat{\mathbf{y}}\|_2^2, \; \mathcal{L}_3 = \frac{1}{1 + \exp\left(\|f_\theta(\mathbf{x}_U) - \hat{\mathbf{y}}\|_2^2 \right)}, \tag{5.23}$$

where f_θ shares the same weights with the original model, with attention and knowledge infusion mechanisms are disabled.

In calculating $\mathcal{L}_3$, the negative sign in the mean squared error loss ensures

that the model outputs corresponding to unimportant features differ from the original output. However, explicit minimisation may cause the loss function to diverge toward negative infinity, therefore, we apply a sigmoid function for the final computation of $\mathcal{L}_3$. Furthermore, to maintain robustness, we ensure that perturbations of important features produce outputs that differ significantly from the original, whereas perturbations of unimportant features yield outputs similar to the original:

$$\mathcal{L}_4 = \left\| f_\theta(\mathbf{x}_U^\delta) - \hat{\mathbf{y}} \right\|_2^2, \; \mathcal{L}_5 = \frac{1}{1 + \exp\left(\left\| f_\theta(\mathbf{x}_I^\delta) - \hat{\mathbf{y}} \right\|_2^2 \right)}, \tag{5.24}$$

Finally, the total loss is a weighted average of all loss terms:

$$\mathcal{L}_{\text{total}} = \frac{\sum_{i=1}^{5} \lambda_i \cdot \mathcal{L}_i}{\sum_{i=1}^{5} \lambda_i} \tag{5.25}$$

where each λ_i is the hyperparameter that controls the contribution of its associated loss term. This final objective formulation enforces the model to be accurate and to generate faithful and robust explanations, identifying the components of social media posts that are most influential in detecting mental disorders.

5.4 Experiments and Results

5.4.a Datasets

Three publicly available and fully anonymised datasets across three distinct mental health challenges, including psychological stress, depression, and suicide ideation, have been involved in this study to validate our proposed methodology.

Dreaddit Dataset ($\mathbf{D_R}$): Dreaddit is a labelled textual Reddit dataset to support the screening of psychological stress in social media texts, developed by Turcan et al. [249]. It comprises 3,553 textual social media posts across a variety of domains, including abuse, social relationships, anxiety, Post-Traumatic Stress Disorder (PTSD), and financial difficulties, manually annotated via Amazon Mechanical Turk with binary stress labels (stress vs. not stress). As a result, Dreaddit provides high-quality, human-annotated and labelled social media posts with a rich and expressive nature in terms of psychological distress, which makes it a critical resource for explainable mental disorder risk detection.

Depression Dataset ($\mathbf{D_D}$): To explore the effectiveness of our explainability method on depression assessment, the dataset introduced by Pirina and Çöltekin [196] for binary depression detection has been included in our study. It contains 1,841 textual Reddit posts collected from over 10,000 Reddit communities, including depression-related subreddits (e.g., r/depression) and control groups from subreddits such as r/breastcancer, r/family, and r/friendship. The rich and diverse nature of the $\mathbf{D_D}$ dataset makes it an indispensable source for explainable depression detection.

SDCNL Dataset ($\mathbf{D_S}$): For the assessment of suicide risk detection, we employ the SDCNL dataset by Haque et al. [96], which contains 1,895 textual Reddit posts scraped via the Python Reddit Application Programming Interface (API) (PRAW). These posts are labelled as suicidal or non-suicidal, containing both explicit and implicit mentions of suicide ideation. $\mathbf{D_S}$ serves as a valuable source of information by providing helpful insights into the early stages of suicide risk.

Dataset Statistics and Preprocessing: For each dataset, a standard text cleaning pipeline including HTML tag removal, emoji and URL stripping, lowercasing, lemmatisation, stopword removal, and tokenisation have been followed. Table 5.1 demonstrates the statistical information about the $\mathbf{D_D}$, $\mathbf{D_R}$, and $\mathbf{D_S}$. The

Table 5.1: Basic descriptive statistics for the datasets

Dataset	Samples	Negative (%)	Positive (%)	Avg. Length	Tokens
$\mathbf{D_R}$	3553	1696 (47.7%)	1857 (52.3%)	39.30	139,631
$\mathbf{D_D}$	1841	548 (29.8%)	1293 (70.2%)	98.26	180,904
$\mathbf{D_S}$	1895	915 (48.3%)	980 (51.7%)	81.26	153,984

depression dataset ($\mathbf{D_D}$) is significantly imbalanced, with 70.2% of posts being depressive, compared to the the stress dataset ($\mathbf{D_R}$) and the suicide ideation dataset ($\mathbf{D_S}$) that are relatively balanced. $\mathbf{D_R}$ posts are shorter (39.3 tokens on average) than $\mathbf{D_D}$ and $\mathbf{D_S}$, with average token lengths of 98.26 and 81.26 tokens respectively.

5.4.b Experimental Settings

The MentalXAI model was implemented in PyTorch 2.4.0 within a CUDA 12.1 environment. Model training was conducted with a 5-fold cross-validation procedure to ensure the reliability of results.

Model training and evaluation was conducted on *Viking HPC cluster provided by the University of York*. The optimal hyperparameters were determined through

Table 5.2: Experimental settings for each dataset

Parameter	Description	D_R	D_D	D_S
$hidden\_dim$	Hidden size of the model	32	64	256
gnn_layers	Number of GNN components	2	1	1
$heads$	Number of heads in the attention mechanism	2	8	8
$\mathcal{B}$	Batch size	16	8	8
η	Learning rate	6.63×10^{-5}	8.75×10^{-5}	9.27×10^{-5}
d_p	Dropout rate	0.2090	0.4431	0.3433
α	Sharpness rate	144.558	229.484	221.934
λ_1	Weight controls prediction perf.	0.2955	0.7291	0.7290
λ_2	Controls **Su**	0.3556	0.3899	0.1903
λ_3	Controls **Co**	0.8677	0.5805	0.6768
λ_4	Controls $\mathbf{Ro}^+$	0.1076	0.1226	0.4588
λ_5	Controls $\mathbf{Ro}^-$	0.3896	0.6327	0.6727
$\hat{q}$	Quantile level for attention threshold	0.9	0.9	0.9
s	Standard deviation of the noise	5.0	5.0	5.0

Bayesian optimisation via Optuna with 100 trials to maintain the best balance across metrics. To ensure fairness, the maximum score achieved by each model is reported across: i) the original hyperparameters reported in the respective research papers, ii) the optimal hyperparameters of MentalXAI, and iii) supplementary manual tuning in cases where the first two conditions yielded sub-optimal performance. All models were trained using the Adam optimiser with default parameters and a cosine scheduler. The detailed values for hyperparameters are reported in Table 5.2.

5.4.c Evaluation Metrics

Although the number of studies in XAI research is growing, a consensus regarding XAI model evaluation has not been established [194, 197]. Therefore, based on Definition 5.2.1, we adapt several widely used methods to evaluate the explainability performance of our methodology against baseline models. Figure 5.3 presents the hierarchical structure of the metrics used to assess model explainability. De Young et al. [65] state that explanations generated by candidate models must be **faithful** to the model's original outputs. Faithfulness can be measured in terms of **sufficiency** and **comprehensiveness** [65], where sufficiency indicates whether the explanations can reproduce the model's original outputs, and comprehensiveness assesses whether all necessary features are identified as explanations. Additionally, **robustness** is a key property of XAI [125], as it quantifies model uncertainty and evaluation of how explanations respond to feature perturbations [8, 103, 75].

To measure the explainability performance of our methodology using the aforementioned methods, we employ the following evaluation strategy:

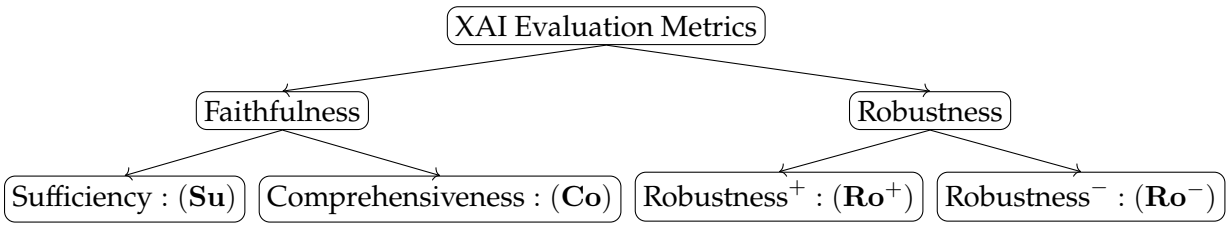


Figure 5.3: Evaluation metrics for the XAI models

- To monitor the overall model performance, we compute $\mathbf{F_1^o} = F_1\left(\arg\max\left(f_{\theta,\Lambda}(\mathbf{x})\right),\ y\right)$, - the F_1 score using the full input which is expected to be **high**.
- Sufficiency (**Su**) measures the model's performance using only the explanations (important features) and is calculated as:

$$\mathbf{Su} = F_1\left(\arg\max\left(f_\theta(\mathbf{x}_I)\right),\ y\right) \tag{5.26}$$

  Higher values of **Su** indicate better sufficiency - expected to be **high**.
- Comprehensiveness (**Co**) measures the model's performance using only the unimportant features and is calculated as:

$$\mathbf{Co} = F_1\left(\arg\max\left(f_\theta(\mathbf{x}_U)\right),\ y\right) \tag{5.27}$$

  The comprehensiveness score is expected to be **low**.
- Robustness is measured by applying perturbations to the important features ($\mathbf{Ro}^+$) and unimportant features ($\mathbf{Ro}^-$). Perturbing important features should drastically affect the output, while perturbing unimportant features should have minimal effect.

$$\mathbf{Ro}^+ = F_1\left(\arg\max\left(f_\theta(\mathbf{x}_I^\delta)\right),\ y\right) \tag{5.28}$$

$$\mathbf{Ro}^- = F_1\left(\arg\max\left(f_\theta(\mathbf{x}_U^\delta)\right),\ y\right) \tag{5.29}$$

  Accordingly, $\mathbf{Ro}^+$ is expected to be **low**, while $\mathbf{Ro}^-$ is expected to be **high**.

5.4.d Evaluation Models

We evaluate MentalXAI against four widely used post-hoc explainability methods for feature importance (LIME, SHAP, Integrated Gradients (IG), Layer-wise Relevance Propagation (LRP)) and four intrinsic XAI models (Vanilla Attention, Hierarchical Attention Network, Self-Attention, and Cross-Attention) across all

three datasets. The intrinsic methods are evaluated under three configurations: No Adaptor (N), Half Adaptor (H), and Full Adaptor (F):

- **No Adaptor - N:** The training process does not incorporate any configuration for enhancing faithfulness and robustness.
- **Half Adaptor - H:** The training process incorporates mechanisms to enhance faithfulness but does not address robustness.
- **Full Adaptor - F:** The training process incorporates mechanisms to enhance both faithfulness and robustness.

The detailed list of models employed in this study is as follows:

- **Local Interpretable Model-agnostic Explanations - LIME** [204] is a post-hoc, local, model agnostic explainability method used to explain any black-box model via the feature importance approximation. LIME identifies feature importance by fitting simple intrinsic models like linear regression to the perturbed features that helps to approximate the most influential portions of the inputs.
- **SHapley Additive exPlanations - SHAP** [147] is also a post-hoc, local, model agnostic explainability method that assigns importance values (Shapley values) for each feature, inspired by the concept of Shapley values from game theory, by generating feature permutations.
- **Integrated Gradients - IG** [237] is a local post-hoc model-specific explainability method. It is model specific, as it can only be applied to differentiable models. IG identifies the importance of the features by computing the gradients of model's output when the baseline inputs (usually zero vectors) are interpolated into the actual input.
- **Layer-wise Relevance Propagation - LRP** [20] is a post-hoc, local, model agnostic explainability method that identifies feature importance based on their contribution to the final decision via backward propagation through the network.
- **Vanilla Attention** [21] is an early version of the Attention mechanism, a neural network based local, model specific, intrinsic explainability method that can be plugged into almost all neural network models. Attention values are attached to the input feature representations that are further learnt during training which indicate the feature importance.
- **Hierarchical Attention Network (HAN)** [274] is a specific form of the vanilla attention, but considers multiple levels of attention based on different components of the input, such as words and sentences. Conventional HAN adopts word and sentence level attention mechanism, therefore represents word and sentence level importance.

- **Self Attention** [252] is a contemporary attention mechanism that is a key idea behind Multihead Attention and Transformer models (therefore LLMs), that calculates the attentions scores with the help of the semantic similarities of the key and query inputs. In Self attention the key and query vectors are derived from the same source of information.
- **Cross Attention** [252] is a special version of self attention, but adopts distinct query and key vectors. The query vector is computed from the original input, whereas the key vector is derived from another source of information, in our case, mental health domain knowledge. As a result, both attention mechanisms returns the feature importance scores.

5.4.e Performance Evaluation

In this section, we evaluate our MentalXAI model against post-hoc and intrinsic baseline explainability methods, analysing the role of the attention adaptor. Table 5.3 presents a thorough comparison of model performance, reporting both prediction F_1 scores and explainability metrics for all three mental health datasets. We foresee trade-offs between classification and explainability performance across all mental health conditions.

Our complete model, MentalXAI, demonstrates the best overall balance across all datasets. On stress detection, it achieves 0.7618 $\mathbf{F_1}$, 0.7244 sufficiency, 0.1362 comprehensiveness, 0.2354 robustness for important features, and 0.6195 for unimportant features. On the depression detection task, MentalXAI achieves 0.9223 $\mathbf{F_1}$, 0.7031 sufficiency, 0.0079 comprehensiveness, 0.4775 robustness for important features, and 0.7980 for unimportant features. Finally, on suicide risk detection, our model yields 0.7184 $\mathbf{F_1}$, 0.7152 sufficiency, 0.0245 comprehensiveness, 0.2215 robustness for important features, and 0.6224 for unimportant features. These results clearly confirm that MentalXAI satisfies all the conditions of explainability stated in Definition 5.2.1, generating faithful and robust explanations for all mental health assessment tasks.

Table 5.3: Performance across models on three datasets

Dataset	Model	$\mathbf{F_1^o} \uparrow$	$\mathbf{Su} \uparrow$	$\mathbf{Co} \downarrow$	$\mathbf{Ro^+} \downarrow$	$\mathbf{Ro^-} \uparrow$
	LIME [204]	0.7337	0.3111	0.7225	0.1133	0.1890
	SHAP [147]	0.7314	0.7135	0.5817	0.3522	0.4380
	IG [237]	0.7294	0.7105	0.5457	0.2358	0.3048

Continued on next page

Table 5.3 – continued from previous page

Dataset	Model	$\mathbf{F_1^o} \uparrow$	$\mathbf{Su} \uparrow$	$\mathbf{Co} \downarrow$	$\mathbf{Ro^+} \downarrow$	$\mathbf{Ro^-} \uparrow$
	LRP [20]	0.7294	0.6470	0.6468	0.2366	0.2375
	Vanilla Attention + N [21]	0.7577	0.7642	0.7586	0.5312	0.5292
	Vanilla Attention + H	0.7473	0.7379	0.6169	0.5820	0.5704
	Vanilla Attention + F	0.7348	0.7211	0.2961	0.4425	0.6528
	HAN + N [274]	0.7346	0.5048	0.6738	0.6718	0.6347
	HAN + H	0.7432	0.7280	0.6916	0.5914	0.4359
	HAN + F	0.7335	0.7117	0.6508	0.5698	0.6796
	Self Attention + N [252]	0.7584	0.5533	0.5538	0.5788	0.5785
	Self Attention + H	0.7507	0.7456	0.5722	0.5188	0.5372
	Self Attention + F	0.7483	0.7223	0.7288	0.7223	0.7223
	Cross Attention + N	0.7513	0.3584	0.3629	0.5255	0.5255
	Cross Attention + H	0.7607	0.7527	0.2775	0.3555	0.3583
	MentalXAI (Ours)	0.7618	0.7244	0.1362	0.2354	0.6195
D_D	LIME [204]	0.9169	0.8511	0.9139	0.4535	0.5457
	SHAP [147]	0.9273	0.8823	0.8431	0.2754	0.5907
	IG [237]	0.9208	0.8358	0.8465	0.4750	0.6454
	LRP [20]	0.9208	0.9197	0.9204	0.6661	0.6641
	Vanilla Attention + N [21]	0.9198	0.9179	0.9113	0.7468	0.7280
	Vanilla Attention + H	0.9321	0.9284	0.8280	0.7637	0.7196
	Vanilla Attention + F	0.9246	0.9205	0.8208	0.8159	0.8140
	HAN + N [274]	0.9110	0.8232	0.8916	0.5849	0.3254
	HAN + H	0.9073	0.8964	0.8252	0.2499	0.1627
	HAN + F	0.9030	0.8852	0.8241	0.1532	0.8556
	Self Attention + N [252]	0.9251	0.7529	0.7523	0.7269	0.7269
	Self Attention + H	0.9263	0.8867	0.5728	0.5553	0.5519
	Self Attention + F	0.9252	0.9051	0.0016	0.3727	0.7972
	Cross Attention + N	0.9243	0.6639	0.7427	0.7473	0.7448
	Cross Attention + H	0.9260	0.9151	0.2482	0.5556	0.5791
	MentalXAI (Ours)	0.9223	0.9031	0.0079	0.4775	0.7980
D_S	LIME [204]	0.7176	0.6849	0.7148	0.5861	0.5361
	SHAP [147]	0.7097	0.7106	0.5122	0.6062	0.5459
	IG [237]	0.7076	0.7087	0.5137	0.6193	0.5421
	LRP [20]	0.7076	0.7024	0.7028	0.5684	0.5684
	Vanilla Attention + N	0.7410	0.7270	0.6407	0.5366	0.5195
	Vanilla Attention + H	0.7386	0.7227	0.3101	0.4963	0.5056

Continued on next page

Table 5.3 – continued from previous page

Dataset	Model	$\mathbf{F_1^o} \uparrow$	$\mathbf{Su} \uparrow$	$\mathbf{Co} \downarrow$	$\mathbf{Ro^+} \downarrow$	$\mathbf{Ro^-} \uparrow$
	Vanilla Attention + F	0.7413	0.7111	0.1893	0.3790	0.5050
	HAN + N [274]	0.6914	0.6782	0.6762	0.5403	0.5489
	HAN + H	0.6801	0.6732	0.5632	0.4361	0.4049
	HAN + F	0.7035	0.6918	0.6330	0.4405	0.4372
	Self Attention + N [252]	0.7161	0.6015	0.6020	0.5931	0.5929
	Self Attention + H	0.7219	0.7094	0.2293	0.5290	0.5257
	Self Attention + F	0.7186	0.7006	0.1170	0.2994	0.5816
	Cross Attention + N	0.7124	0.5771	0.5916	0.4951	0.4820
	Cross Attention + H	0.7289	0.7154	0.1015	0.4186	0.4214
	MentalXAI (Ours)	0.7184	0.7152	0.0245	0.2215	0.6224

Post-hoc explainability methods generally produced sufficient explanations for all mental health conditions. However, LIME failed to do so on the screening of stress, with a sufficiency score of 0.3111. Nevertheless, post-hoc explanations generally lacked comprehensiveness, with scores ranging from 0.54 to 0.92 in all cases, particularly LRP and LIME generated the least comprehensive explanations on all conditions. Furthermore, these methods produce unreliable explanations regarding general robustness metrics. Important features generated by post-hoc methods were relatively robust in depression and stress detection, except for LRP, which showed poor robustness on important features with a score of 0.6661, indicating the general unreliability of post-hoc explainers for robustness evaluation. Nonetheless, LIME, SHAP, and IG perform better on depression detection, likely due to more explicit depressive cues in D_D. Overall, these results indicate that post-hoc explainers fail to approximate true explanations for mental health assessment models.

Vanilla attention, as one of the most widely used intrinsic explainability methods, generated sufficient explanations for all three mental health conditions. However, without any enabled adaptors, the vanilla attention fails to generate comprehensive and robust explanations across all datasets. Importantly, enabling the attention adaptor mechanism (half and full configurations) enhanced vanilla attention's performance as comprehensiveness scores improved to 0.2961 for stress detection and 0.1893 for suicide risk detection.

HAN, as a specific variant of attention mechanisms, was anticipated to yield superior performance. However, it demonstrated weaker performance compared to other baseline models. While HAN generated sufficient explanations for suicide and depression detection, it produced insufficient explanations for stress detection with 0.5048 sufficiency score. Enabling attention adaptors im-

proved HAN's explainability performance on stress data and slightly enhanced robustness for both stress and depression detection. However, HAN with attention adaptors failed to show consistent improvements across all metrics, as the generated explanations persisted as non-comprehensive.

Self Attention, as an advanced form of attention mechanisms, demonstrated strong overall performance. It achieved high prediction performance consistently and generated sufficient explanations for depression and suicide detection, though it underperformed on stress detection with 0.5533 sufficiency. Enabling the half attention adaptor improved faithfulness metrics across all conditions, while the full adaptor enhanced both faithfulness and robustness performance. Particularly, full adaptor configuration of Self Attention achieves performance nearly identical to MentalXAI on depression detection with attaining 0.9252 F_1, 0.9051 sufficiency, 0.0016 comprehensiveness, 0.3727 robustness on important features, and 0.7972 robustness on unimportant features, validating exceptional explainability.

Domain knowledge infusion through The Cross Attention mechanism demonstrates strong classification performance across all conditions and adaptor configurations. However, attention mechanisms, including Cross Attention, require adaptor refinements to achieve true explainability through improved faithfulness and robustness, as previously discussed. The half attention adaptor enhances faithfulness across all conditions, while enabling full adaptor (forming MentalXAI) achieves the best overall balance between explainability and prediction performance. Although our model performs similarly to Self Attention with enabled full attention adaptor on depression detection, it outperforms Self Attention on both stress and suicide risk detection.

In summary, these results confirm that the half attention adaptor improves explanation faithfulness, while the full adaptor enhances both faithfulness and robustness. The identical performance of Self Attention + F and MentalXAI on depression detection validates the effectiveness of our adaptor mechanism in empowering explainability independent of domain knowledge. Furthermore, integrating domain knowledge with the full attention adaptor yields the most consistent performance across all metrics and mental health conditions. Moreover, we applied both the standard Wilcoxon test and the Holm–Bonferroni corrected Wilcoxon test, and the corresponding results are presented in Table A.3. Ideally, achieving the highest scores across all metrics would provide the strongest foundations for explainability. Under the current setting, however, although the model does not obtain the best score for each individual metric, it demonstrates consistently strong and balanced performance across all of them. The Wilcoxon test indicates that MentalXAI achieves statistically significant im-

provements over all baseline models except SelfAttention with t**he full attention adaptor (+F)**. While MentalXAI performs slightly better than this model, the difference is not statistically significant. To address the issue of false positives arising from multiple comparisons, we further applied the Holm–Bonferroni correction. After this adjustment, our model remained significantly better than 15 of the 18 baselines. The three models for which the differences are not statistically significant are: SelfAttention + F, HAN + F, and Vanilla + F. Particularly, **"+F"** denotes the full attention adaptor, which constitutes one of the central components and hypotheses of this chapter. As such, these findings offer additional support for our claims.

5.4.f Sensitivity Analysis

Sensitivity of quantile level values and perturbation standard deviation

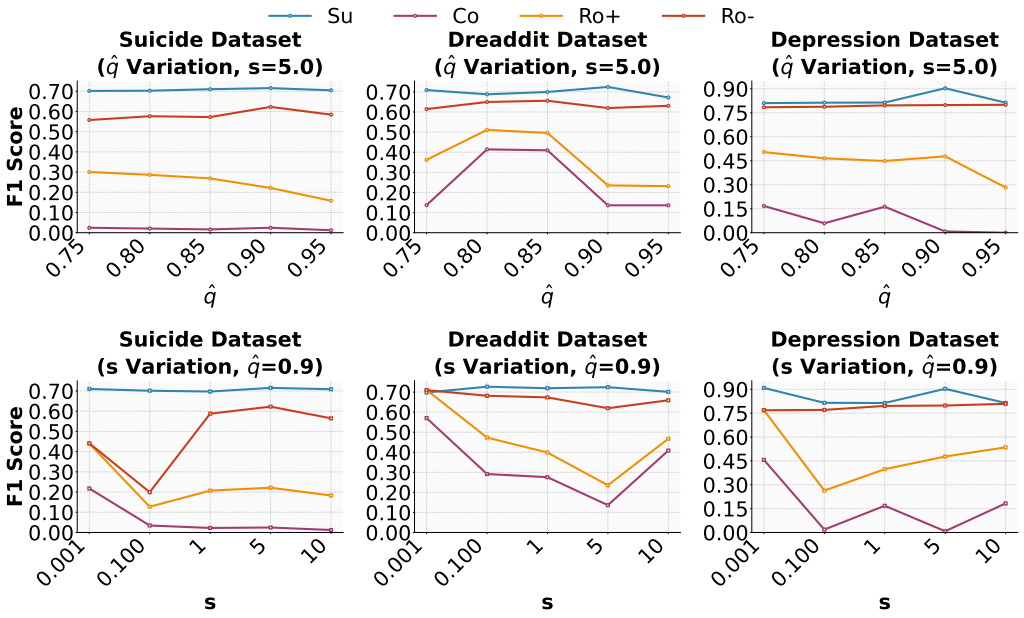


Figure 5.4: Effects of the quantile level value ($\hat{q}$) and delta (s)

Figure 5.4 demonstrates the effectiveness of the quantile level values in the range of $[0.75, 0.95]$ and Gaussian noise standard deviation within $[0.001, 10]$.

The quantile level value ($\hat{q}$) controls the number of important features by regulating the attention threshold, where a higher threshold results in fewer explanations being generated. We observe that sufficiency scores remain consistently high across different $\hat{q}$ configurations for all mental health conditions, reaching

their maximum at $\hat{q} = 0.9$. The behaviour of comprehensiveness varied across datasets as suicide detection model remained at zero (optimal) throughout, while stress and depression classification models fluctuated before achieving optimal values at $\hat{q}$ of 0.9 and 0.95. Robustness of the unimportant features remained stable across different values, while robustness of the important features generally improved. Overall, the most balanced performance across the different explainability metrics is achieved at a $\hat{q}$ of 0.9.

Standard deviation of the added Gaussian noise as perturbation (s) is a critical parameter especially for the explainability, as it directly influences the robustness of the generated explanations. Across different s values, sufficiency demonstrated similar behaviour with the quantile value experiments. Comprehensiveness of the explanations generally improved with increasing s, despite some fluctuations. Robustness scores demonstrated remarkable patterns across different s values, the robustness of unimportant features generally improved for all mental health conditions, while similar improvements for the robustness of the important features were evident particularly in the suicide and stress detection models. Optimal balanced performance across all detection and explainability metrics has been observed at $s = 5.0$.

Sensitivity of Loss function weights

Figure 5.5 summarises the effect of the loss function weights in maintaining explainability conditions.

λ_2 primarily controls the sufficiency criterion across all mental health conditions. Sufficiency scores consistently demonstrate improvement with increasing λ_2 values. Comprehensiveness improved for the suicide and depression datasets, but fluctuated for the stress dataset. Robustness scores did not show a consistent improvement across λ_2 values. Overall, the most balanced scores are obtained at λ_2 values of 0.39 for depression detection, 0.81 for suicide detection, and 0.29 for the stress detection.

λ_3 is responsible for controlling the comprehensiveness of generated explanations. Sufficiency remained stable across λ_3 values, while comprehensiveness improved particularly in the depression dataset. For the suicide and stress datasets, comprehensiveness improved up to a certain point before deteriorating. Robustness scores also demonstrated slight improvements across most mental health conditions. The optimal λ_3 values for maintaining balanced performance are 0.33, 0.86, 0.58 for suicide, stress, and depression detection, respectively.

λ_4 and λ_5 are responsible for controlling the robustness of important and unimportant features, respectively. For all conditions, robustness of the important features improved before declining, indicating the critical need for careful

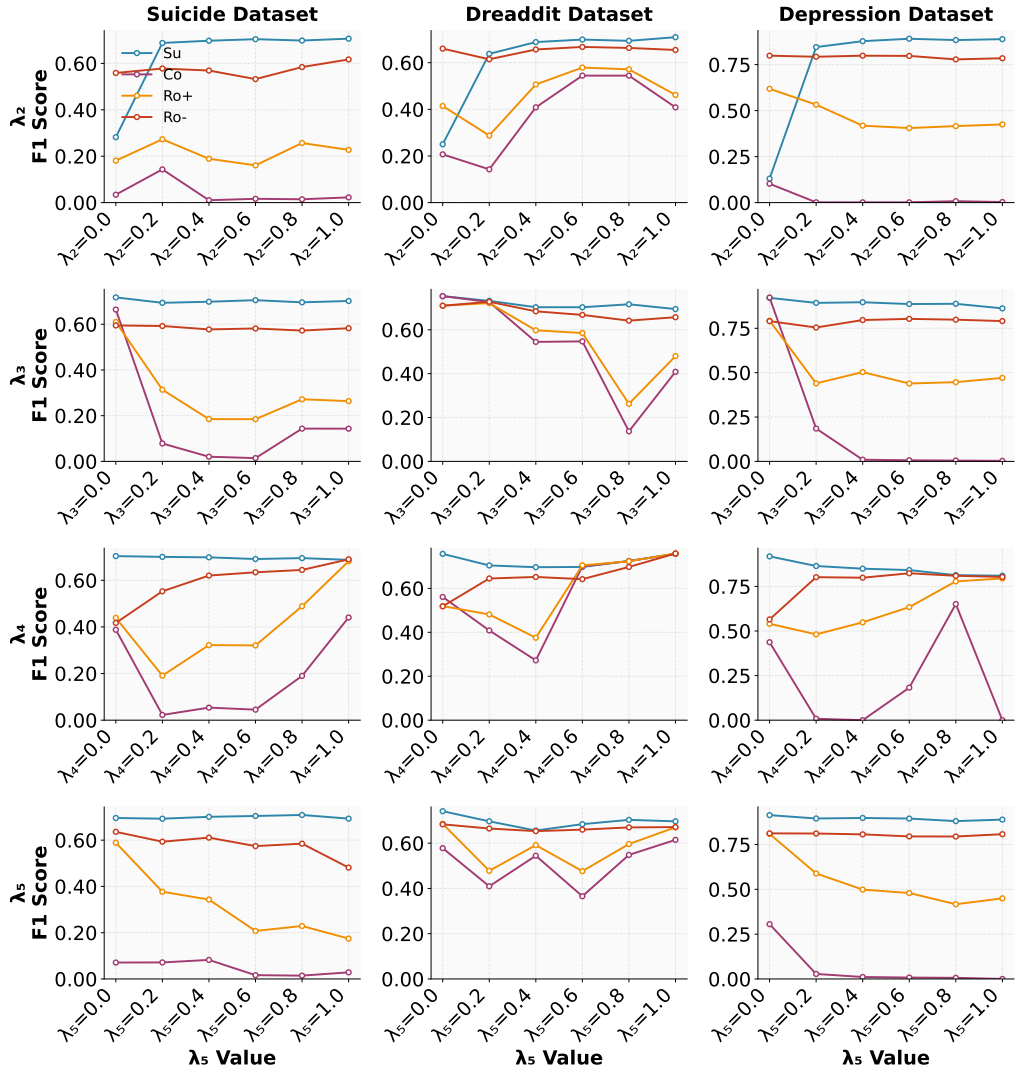


Figure 5.5: Effects of the loss function weights

calibration of λ_4. The optimal values are 0.22, 0.10, 0.12 for suicide, stress, and depression detection. Across λ_5 values, we perceived a steadier enhancement, particularly in the robustness of unimportant features generated by suicide and depression detection models. Despite some fluctuations, an overall improvement in the robustness of important features was observed. The optimal λ_5 values are selected as 0.79 for suicide detection, 0.38 for stress detection, and 0.63 for depression detection models.

In summary, our model does not demonstrate any deterioration in sufficiency, which is a positive outcome. However, sufficiency alone is not sufficient as

comprehensiveness and robustness must also be maintained. Moreover, hyperparameters cannot be selected solely based on the maximum value of their associated metrics, as balanced performance across all metrics is crucial for explainability.

5.4.g Qualitative Analysis of Explanations

Psycholinguistic Relevance Analysis

One of the key desiderata of our XAI framework for MH is that the generated explanations possess psycholinguistic relevance. To evaluate the relevance of explanations from a mental health perspective, we use the Empath [73], a tool for understanding topic signals in large-scale text. Empath generates these topic signals by employing a vector space model (VSM) trained via a neural network on a large text corpus.

Rank	Full (+)	Score	Full (–)	Score	Expl. (+)	Score	Expl. (–)	Score
1	neg_emo	0.0709	neg_emo	0.0553	neg_emo	0.1413	neg_emo	0.0655
2	pain	0.0414	pos_emo	0.0395	death	0.1068	suffering	0.0296
3	pos_emo	0.0392	pain	0.0380	kill	0.0986	death	0.0284
4	death	0.0391	shame	0.0335	violence	0.0680	children	0.0279
5	violence	0.0384	violence	0.0293	suffering	0.0676	pos_emo	0.0271
6	suffering	0.0373	suffering	0.0289	med_emer	0.0517	pain	0.0234
7	shame	0.0346	love	0.0273	pain	0.0500	violence	0.0221
8	sadness	0.0260	nervousness	0.0269	crime	0.0422	healing	0.0213
9	hate	0.0259	sadness	0.0267	hate	0.0345	family	0.0188
10	nervousness	0.0241	contentment	0.0232	weapon	0.0336	home	0.0179

Table 5.4: Top Empath categories with scores for the Suicide dataset and associated explanations

Table 5.4 shows the salient topics of the original social media posts from the dataset used to train the suicide detection model, together with the generated explanations for these posts and associated Empath scores. We divided the posts into two categories including positive (posts indicating suicide risk) and negative (posts without suicide risk) based on the predictions of the model predictions. Original input posts flagged by the model as having the suicide risk mostly contain signals with negative implications, such as *negative_emotion*, *pain*, and *death*. It is noteworthy that, the original posts contain signals related to positive emotions. The explanations generated for the samples with positive predictions, however, reduced the prominence of positive signals, while, the salient signals are directly related to suicide, such as **death**, **kill**, **violence**, and **suffering**. Furthermore, posts with negative predictions contain a mixture of the positive and negative emotions, and generated explanations do not have the suicide related signals to the same extent as those with positive predictions, as expected.

Rank	Full (+)	Score	Full (–)	Score	Expl. (+)	Score	Expl. (–)	Score
1	neg_emo	0.0435	friends	0.0281	neg_emo	0.1025	neg_emo	0.0440
2	nervousness	0.0331	pos_emo	0.0252	nervousness	0.0832	violence	0.0349
3	pain	0.0310	party	0.0249	pain	0.0782	pain	0.0326
4	violence	0.0287	children	0.0249	violence	0.0752	nervousness	0.0317
5	shame	0.0243	communication	0.0240	suffering	0.0640	shame	0.0254
6	pos_emo	0.0239	family	0.0208	shame	0.0573	health	0.0222
7	health	0.0236	home	0.0206	fear	0.0554	fear	0.0213
8	friends	0.0226	speaking	0.0203	hate	0.0455	sadness	0.0213
9	suffering	0.0217	business	0.0186	sadness	0.0455	suffering	0.0190
10	speaking	0.0213	neg_emo	0.0186	health	0.0426	pos_emo	0.0186

Table 5.5: Top Empath categories with scores for the Dreaddit dataset and associated explanations

Table 5.5 shows the salient signals identified in both the inputs and generated explanations generated by the stress detection model. According to the table, we can observe similar patterns seen in the suicide risk detection task. The posts with positive predictions and their associated explanations generated by suicide risk detection model reflect content with negative implications, such as *negative_emotion*, *nervousness*, *shame*, and *violence*. However, prominence of the stress related cues such as **violence**, **shame**, **pain**, **fear**, and **hate** is more notable compared to the full input texts, indicating that the explanations successfully extract these cues. Meanwhile, unlike in suicide detection, social media posts with negative predictions contain more positive signals. Although, their corresponding explanations still contain negative signals such as *negative_emotion*, *violence*, and *pain*, these are less intense compared to explanations for positive predictions.

Rank	Full (+)	Score	Full (–)	Score	Expl. (+)	Score	Expl. (–)	Score
1	neg_emo	0.0520	children	0.0629	sadness	0.0711	children	0.1294
2	pos_emo	0.0385	family	0.0613	death	0.0654	family	0.1235
3	pain	0.0318	home	0.0483	neg_emo	0.0654	friends	0.1142
4	shame	0.0278	friends	0.0458	health	0.0583	home	0.0826
5	violence	0.0270	party	0.0431	suffering	0.0552	party	0.0814
6	suffering	0.0262	pos_emo	0.0402	neglect	0.0452	pos_emo	0.0802
7	sadness	0.0256	neg_emo	0.0393	pos_emo	0.0450	help	0.0755
8	friends	0.0245	speaking	0.0278	friends	0.0411	youth	0.0422
9	party	0.0239	dom_work	0.0275	party	0.0358	death	0.0416
10	nervousness	0.0236	communication	0.0269	children	0.0309	dom_work	0.0381

Table 5.6: Top Empath categories with scores for the depression and associated explanations

Table 5.6 reports the major topics of the inputs and explanations generated by the depression detection model. However, the patterns observed in this case are slightly different from those seen during the stress and suicide risk detection. For positive predictions, no atypical patterns are observed as the original input posts contain primarily negative signals with some minor positive implications,

and the explanations tend to highlight negative signals such as **sadness**, **death**, and **negative_emotion**. However, the original inputs with negative predictions and their associated explanations contain significantly more positive signals, which differs significantly from the patterns observed in the previous cases, particularly regarding the saliency of positive cues in the explanations. This difference can be related to the labelling strategy of the datasets. For instance, labelling could follow one of three strategies:

a Examples with positive predictions incorporate indicative patterns of the mental disorder, while negative examples reveal melancholy but do not present the disorder.

b Examples with positive predictions exhibit patterns of the related disorder, and negative examples have neutral sentiment, clearly lacking the disorder.

c Examples with positive predictions contain clear patterns of the disorder, whereas negative examples display overall optimistic sentiment, with the absence of the disorder.

Among these strategies, a) is the most challenging due to the subtle margin between classes, while c) is the easiest, with the clearest separation between the two labels. This observation can be linked with the prominence of optimistic signals in posts labelled negative (indicating the absence of depression) and associated generated explanations by the depression detection model. We can further support this statement by the results reported in Table 5.3, which indicates that the prediction performance during the depression detection process is very high across all variants, especially compared to the stress and suicide risk detection.

Attention Visualisation

We further analyse the important words that indicate mental disorders using heatmap based attention value visualisation. Figures 5.6–5.8 illustrate the salient words based on the explainability adapted and knowledge aware attention values. For the suicide detection task, the most salient words are generally suicide and desperation related terms, such as *die*, *death*, *suicide*, *kill*, and *end*. This aligns with Table 5.6, as it further confirms the prominence of suicide related topics. Moreover, Figure 5.7 highlights the most influential words in the stress detection task, demonstrating the impact of negatively valenced words such as *hate*, *cruelty*, *injuries*, and *regret*. Finally, Figure 5.8 demonstrates the most important words that indicate depression risk, where salient terms include *depression*, *depressant*, and *anxiety*. Overall, the highlights in Figures 5.6–5.8 align with Tables 5.4–5.6.

I know I'll either die by accident or by my own hand I know im going to kill myself, I've known it for 7 years, Ive been mentally ill since I was 6. I know what the end game is, the only thing unsure is the timeframe.,

Nothing's changed, hope is gone, I guess I got unlucky. Peace out y'all. <3",

Pragmatic suicide This might sound ridiculous, but I want to die, and I'm not severely depressed. I'm approaching this purely philosophically.\n\nThe way I see it, my life is more suffering than joy, and that's probably not going to change. In fact, I would say this is true of most people. We spend most of our time doing things we don't want to do: working, cleaning, driving in traffic, feeling hungry, experiencing physical discomfort, etc, etc. Why would it not be better to not exist at all (I don't believe in an afterlife) rather than experience the totality of consciousness? \n\nNow other people will mourn my death, this much is true. But they're going to mourn my death regardless of when it happens. No mourning is avoided by making it happen now. If anything, I'm at a point in my life when no one is dependent on me and my death will have the smallest impact it probably ever will.\n\nSo why not die?

Figure 5.6: Attention visualisation for suicide data

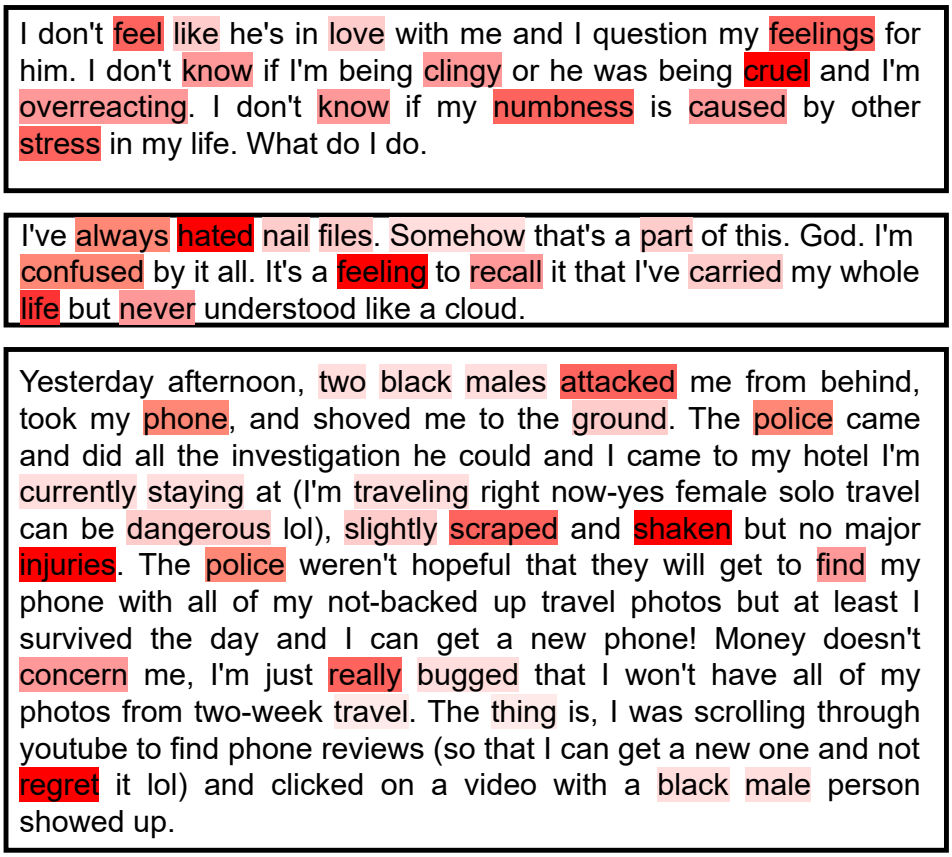


Figure 5.7: Attention visualisation for stress data

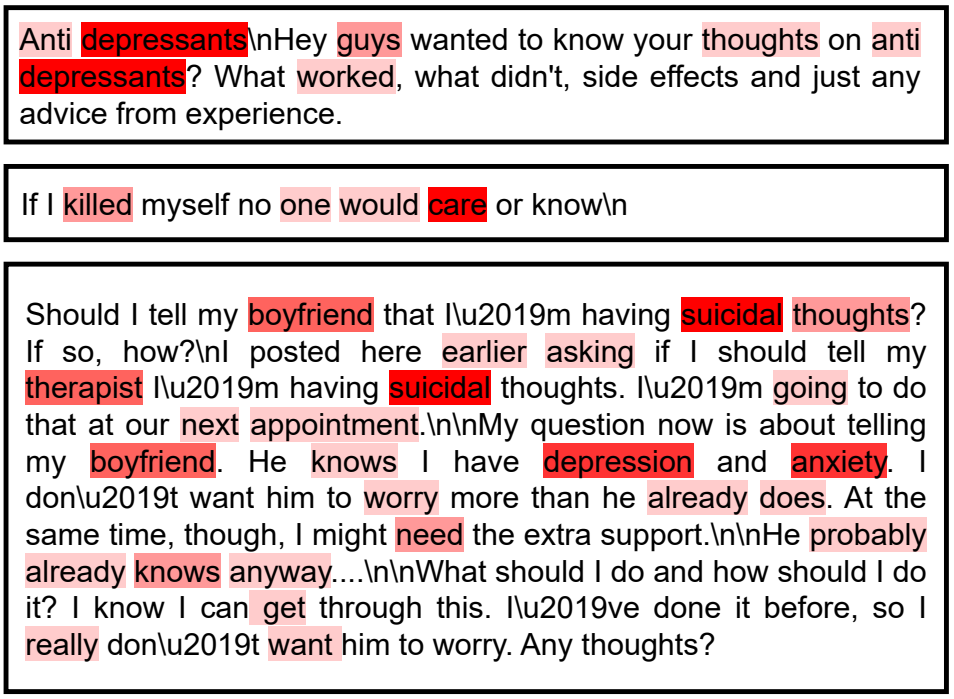


Figure 5.8: Attention visualisation for depression data

5.5 Summary

In this chapter, we propose a theoretical framework for XAI in MH applications in the form of desiderata, given that existing concepts are not fully applicable to NLP-based MH applications. Building upon this theoretical background, we introduce the MentalXAI model for explainable mental health assessment of textual social media data, which generates knowledge-aware explanations through the infusion of a mental health knowledge graph using cross-attention and attention adaptor mechanisms. This contribution enables us to explore the underlying decision mechanisms of the model by generating important features (explanations) and enforcing these explanations to be faithful, comprehensive, and domain-related. MentalXAI has been evaluated on three prevalent mental health conditions including stress, depression, and suicide ideation. The model outperforms state-of-the-art explainability baselines and configurations, demonstrating consistency across all detection and explainability metrics for all evaluated mental health challenges. Moreover, consistent with our proposed desiderata, we analyse the psycholinguistic relevance of explanations generated by MentalXAI, which reflect meaningful and domain-related signals.

6 Conclusion, Limitations and Future Work

This chapter summarises the contributions of the dissertation, investigates its main limitations, and outlines potential directions for future research.

6.1 Conclusion

The primary objective of this dissertation was to develop innovative Explainable Artificial Intelligence (XAI) frameworks for analysing mental disorders using social media data. The proposed approaches in each chapter addressed different components of the XAI trinity for Mental Health (MH) by examining disorder severity, contributing factors, and indicative words in social media posts, while integrating domain knowledge.

First, we provided a detailed and extensive review of mental health analysis via XAI methods based on social media data. The review process consisted of four phases, including traditional diagnostic methods, data-driven detection approaches, explainability of AI-based methods for mental health analysis, and evaluation methods and datasets. Our study revealed a clear growth of interest in data-driven approaches for mental health analysis. However, major ethical aspects such as explainability remained overall underestimated.

Second, we introduced knowledge-infused depression severity screening in Chapter 3. Prior works had generally undervalued the importance of severity assessment and domain knowledge, and the few that considered domain knowledge often lacked generalisability. To address this shortcoming, we investigated the research question **RQ1**. We proposed the AttentionDep model for knowledge-aware depression severity estimation via hierarchical attention over unigram and bigram representations. Textual inputs were vectorised using FastText pre-trained embeddings for computational efficiency, followed by

a Bidirectional Long Short-Term Memory (BiLSTM) to capture contextual dependencies, and a convolution operation to generate bigram-level representations. Our model integrated domain knowledge from Wikipedia to construct a domain-specific knowledge graph, which was processed using graph neural networks to enrich the knowledge embeddings. The final representation of a social media post was obtained by infusing bigram representations and knowledge embeddings through a cross-attention mechanism. We conducted extensive evaluations on three datasets, two with multi-class settings and one with binary labels. Our model consistently outperformed all baselines across these datasets, and ablation studies verified the contribution of each component.

Third, we analysed the potential contributors to mental disorders in Chapter 4. Unfortunately, prior studies did not cover a detailed list of potential contributors, their annotation processes were often biased, and they lacked unbiased detection methods, which is an essential requirement in the mental health domain. Therefore, in this chapter we addressed **RQ2**. We developed the RedCoM dataset for analysing contributors to mental health challenges by annotating Reddit posts through automated methods and human-in-the-loop mechanisms. Moreover, we presented an innovative multi-class learning model, MDCNet, which integrated pre-trained transformer-based text representations with a causality inspired contextual attention mechanism. This enabled the model to distinguish between task relevant and spurious features and to extract task-relevant, fine-grained features for unbiased identification of contributors to mental health challenges. The MDCNet model also employed knowledge infusion. However, since its text representation method (pre-trained transformers) was already complex and computationally expensive, we did not add further complexity with a knowledge graph. Instead, we used a simpler and less expensive method, Linguistic Inquiry and Word Count (LIWC) features. Extensive evaluations showed that MDCNet outperformed baseline and state-of-the-art methods across all evaluation metrics and datasets. These results highlighted that MDCNet enhanced the contextual understanding of contributors to mental disorders and stress.

Finally, in Chapter 5, we investigated explainability in the mental health domain. Prior studies on explainability in mental health had not examined the applicability of existing explainability concepts particularly in the MH domain, had not justified the roles of different tools, and had not evaluated their models via explainability metrics. Based on this gap, we addressed **RQ3**. We introduced the MentalXAI model for explainable assessment of mental disorders through social media data. Our model generated knowledge-aware explanations by infusing post-specific Mental Health Knowledge Graph (MHKG) through cross-

attention and attention adaptor mechanisms. As the core of MentalXAI (excluding the attention adaptor) was inspired by AttentionDep (Chapter 3), text representation was performed using FastText, which was less computationally expensive than pre-trained transformers. Therefore, we employed a knowledge graph as the knowledge source, instead of LIWC, as LIWC was simpler but less detailed. Moreover, in Chapter 4 we used top-K value selection for feature choice, which was simple and interpretable. In Chapter 5, however, we employed a more complex feature selection method, reflecting the technical improvements made throughout the studies. Although the top-K based method yielded strong results, it could suffer from gradient flow limitations. Therefore, to address this potential shortcoming, we employed a sigmoid function with an attention threshold derived from quartiles. MentalXAI was evaluated on stress, depression, and suicide ideation tasks, and it outperformed state-of-the-art baselines and configurations, demonstrating stability across both detection and explainability metrics.

Overall, to the best of our knowledge, the research questions we formulated and the proposed solutions open new horizons for the ethical analysis of mental disorders from different perspectives, including detection, severity estimation, and the investigation of contributing factors. The knowledge graph we developed can supply domain information to any deep learning model, while the dataset can support deeper analyses of contributors. The models we developed can be integrated into any web-based platform, as they do not depend on platform-specific configurations, despite being trained on Reddit data. Moreover, since the text representation methods we adopted are available in multiple languages, and given the translation and transliteration capabilities of Large Language Model (LLM)s, our methods can be adapted to a wide range of languages.

We must acknowledge that the technical contributions of this dissertation do not make clinical diagnostic claims, as they are designed solely for risk prediction. The real-world consequences of errors vary: predicting higher severity than actual, or flagging a non-depressed user as at-risk (false positive), leads to unnecessary clinician review time. In contrast, a false negative (classifying a depressed user as non-depressed or underestimating severity) represents a critical safety failure and requires further analysis/training. Furthermore, regarding mental disorder contributor classification, misclassification could lead to ineffective mental disorder analysis. Therefore, our explainability framework is essential to reveal the decision patterns of the models and mitigate these risks.

Ethics Approval: All analyses are conducted under strict ethical guidelines. This study is approved by the *Physical Sciences Ethics Committee of the University of*

York under application reference *Ibrahimov20230330*. No personally identifiable information was accessed during this research.

6.2 Limitations

While our contributions demonstrate promising outcomes in explainable mental health analysis, several limitations must be mentioned.

In Chapter 3, the number of entities and relations in MHKG is limited. Due to computational constraints, we were unable to cover the full content of the Wikipedia pages. Moreover, we used a simple post-specific knowledge triplet retriever with a cross-encoder for computational efficiency,

In Chapter 4, the number of posts in the RedCoM dataset is limited due to resource constraints for annotation, but the dataset could be further expanded. Moreover, we used the top-K feature selection method within the PyTorch framework, which produced strong results, however, it may raise gradient flow issues.

Chapter 5 focuses on the explainability analysis of mental health models, starting with a proof-based clarification of the usability of attention as an explainability mechanism. However, this clarification assumes ideal settings, thus, a more detailed mathematical proof of the explainability of attention could still be derived. Moreover, this study considered binary classification for detection, however, in more advanced settings such as severity estimation or multi-class classification, single attention values per word may be less effective, since each word can influence the label differently. Finally, we used automated methods to assess the psycholinguistic importance of the explanations, however, additional human evaluation is missing.

6.3 Future research directions

The field of XAI for MH holds considerable promise, with several research directions well-positioned to advance transparency, interpretability, and real-world applicability. Addressing these directions can contribute significantly to the evolution of robust and resilient mental health analytics systems.

Enhanced multimodal and multi-label data collection. To enhance data comprehensiveness, researchers should develop advanced strategies for generating diverse and well annotated multi-modal datasets representing a diverse set of mental health conditions simultaneously. It includes textual as well as multimedia posts, user interactions, behaviours, social networks and contextual information [40, 39]. Collaborative initiatives with mental health professionals are important to ensure the creation of larger datasets that are carefully annotated,

encompassing severity levels and temporal dynamics [15].

Advanced graph representation learning. To detect (the severity of) multiple disorders leveraging social networks and other structural information within Online Social Media (OSM) contents, it is important to further graph representation learning [131, 72]. It involves developing novel approaches that incorporate temporal dynamics and evolving relationships in OSM data. The aim is to create models that effectively represent and analyse the structural components, providing deeper insights into the complex patterns of mental health expressions within online communities .

Realtime mental health analytics. Mental health is a continuous and evolving process, which necessitates a paradigm shift from static analyses to real-time perspectives. This research direction advocates for developing real-time mental health analytics systems capable of tracking and predicting moment-by-moment mental health conditions and conducting longitudinal analysis. It is important to investigate how models can adapt and provide timely insights to address emerging mental health concerns, especially during significant global events such as warfares, economic recessions and pandemics. This requires the development of agile and adaptive models capable of monitoring and responding to shifts in online mental health expressions in real-time [131, 146, 181]. It will pave the way for proactive interventions and personalised support.

Explanation with large language models. Large language models (LLMs), such as ChatGPT, are a recent breakthrough in NLP, prominently built upon transformer-based architectures. Researchers are starting to use LLMs to explain insights of a wide range of things [226]. To enhance the generation of realistic and diverse explanations of Mental Disorder Detection (MDD) models, a promising research direction is to investigate the integration of these generative models with MDD models [271]. This integration has the potential to improve the explainability of MDD models, providing more meaningful insights into mental health predictions, and thereby contribute to a more transparent and understandable decision-making process.

Incorporating cultural variations. To enable an MDD model to accurately understand user expressions on OSM, it is important to be mindful of the inherent diversities in cultural expressions of mental health around the globe. It requires collaboration with experts from diverse cultural backgrounds to ensure the construction of datasets that accurately represent the spectrum of cultural experiences. Additionally, the models should be developed to adapt and generalise across different cultural settings [189, 86].

A Auxiliary Tables

Table A.1: List of relation types used to represent semantic links between mental health concepts

source_or_authority	diagnostic_indication	clinical_recommendation
categorization	context_dependent	background_factors
instance_of	opposite_of	aims_to_help
family_history	helps_manage	unavailable
negation	engagement	potential_cause
used_in	helps_with_process	occurs_in_context
similar_to	prevalence	has_condition
structural_attribute	overlaps_with	helps_develop
comparative_benefit	targets_patients	aims_to_improve
treatment_practice	treatment_line	exclusion
remission_timeline	resolves_with	symptom_similarity
part_of	potential_effectiveness	frequently_co_occurs
causes	illustrates	treatment_preference
side_effect	examples_include	onset_age
has_heritability	addresses	treatment_magnitude
temporal_property	defines	complicated_by
measured_by	comparative_magnitude	created_by
limitation	defined_as	alternative_to
likelihood	enhanced_effectiveness	not_only
diagnosis_related	helps_prevent	communication
adverse_effect_or_outcome	has_severity	inhibits
modifies	modulates	response_rate
frequency_pattern	aims_to_resolve	usage
field_of_work	requires	susceptibility
helps_distinguish	onset	varies
requires_exposure	assists_with	recovery_rate
treatment_access	prescriptive_authority	found_in
has_goal	classification	has_model
redefined_as	helps_identify	shares_mechanism_with
reduced_by	helps_provide	has_role
evidence_based	is_reflection_of	historical_finding
compared_with	different_from	has_mechanism
complementary_to	clinical_expectation	theoretical_framework
topic_related	genetic_association	synonym
induces	has_method	aims_to_change
term_equivalence	potential_history	clinical_practice
indicated_for	belief_claim	helps_treat
occurs_after	affected_group	research_focus
caused_by	differential_diagnosis	educates
conceptual_model	eliminates	potential_harm
negatively_associated_with	explains	has_attribute
evidence_detail	contraindication	evidence_claim
potential_involvement	suggests	occurs_during
has_part	aims_to_reduce	has_characteristic
historical_classification	affects	developed_from
disrupts	manages	common_pattern
prevents	occurs_independently	avoidance_level
effective	combination_therapy	correlates_with
exacerbates	authorization	has_component
frequency	occurs_on_top_of	trained_in
example	treats	involved_in
occurs_before	subtype_of	key_factor_for
related_to	early_onset	has_type
is_a	needs_adaptation	has_diagnosis
not_caused_by	common_occurrence	used_for
cannot_occur_during	origin	follows
helps_change	represents	potential_benefit
avoids	current_status	outcomes
untreated_outcome	shows	delivery_method

Table A.2: List of LLM Prompts Used in the Study

Phase	**LLM Prompt**
First	"You are a digital mental health expert. You HAVE TO to determine whether [Contributor] could be the cause of the mental disorder underlying the post. [Contributor] refers to the [definition of contributor]. Important indicators of [Contributor] include: [List of Indicators] Keywords or phrases that may signal [Contributor] include, but are not limited to: [List of keywords] You may consider the following examples for the [Contributor], that is confirmed by field experts: [List of few examples] When labeling, consider both explicit mentions and implicit contexts where [Contributor] is implied, even if specific keywords are absent. You HAVE to ONLY respond either by YES or NO, whether [Contributor] could be a contributor to the mental health concerns underlying the post, not any explanatory text. "
Second	"You are a digital mental health expert. You HAVE TO to select the major contributor to mental health challenges underlying the post from the list: [Labels]. You HAVE to ONLY choose one major contributor to mental health challenges underlying the post, not any explanatory text. "

Table A.3: Wilcoxon Signed-Rank Test Results With and Without Holm–Bonferroni Correction

Model	**p-value (raw)**	**p-value (Holm)**	**γ (w/o)**	**γ (w/)**
HAN-MCM	0.000031	0.000549	+	+
CrossAttn	0.000061	0.001038	+	+
SelfAttn	0.000092	0.001465	+	+
HAN	0.000092	0.001373	+	+
IG	0.000153	0.002136	+	+
LRP	0.000305	0.003967	+	+
TrEncoder	0.000305	0.003662	+	+
VeriX	0.000763	0.008392	+	+
SelfAttn+H	0.001007	0.010071	+	+
HAN+H	0.001678	0.015106	+	+
SHAP	0.002136	0.017090	+	+
LIME	0.003159	0.022115	+	+
Vanilla+H	0.006226	0.037354	+	+
Vanilla	0.007538	0.037689	+	+
CrossAttn+H	0.007827	0.031307	+	+
Vanilla+F	0.023956	0.071869	+	-
HAN+F	0.027679	0.055359	+	-
SelfAttn+F	0.151398	0.151398	-	-

B Auxiliary Theoretical Information

B.1 Structural Causal Modeling (SCM) and Do-Calculus

Let $X = \{x_1, x_2, \ldots, x_N\} \in \mathbf{X}$ denote the set of features (tokens) for a given textual input. Some of these N features may act as confounders, influencing the model f_θ and potentially leading to biased decisions [188]. More formally, we partition X into two disjoint subsets: causal features C and confounding features U, such that $X = C \oplus U$ (i.e., $X = C \cup U$ and $C \cap U = \emptyset$). This structure can be formally expressed as:

$$\text{Input } X \rightarrow \begin{cases} C \rightarrow R \rightarrow Y & \text{(front-door path)} \\ U \dashrightarrow R \rightarrow Y & \text{(blocked back-door path)} \end{cases} \tag{B.1}$$

A visual representation of this causal model is provided in Figure B.1, which serves as a prototype causal model based on the Structural Causal Model (SCM) framework. Our primary objective is to eliminate or minimise the influence of confounding features (U) on the outcome (Y)—formally, to block the back-door path $U \dashrightarrow R \rightarrow Y$. Judea Pearl's *do-calculus* [188] provides a formal mathematical framework for this, allowing us to adjust for confounders via an intervention on the causal features C:

$$\mathbb{P}_m(Y|C) = \mathbb{P}(Y|\text{do}(C))$$

The following assumptions are essential to derive the backdoor adjustment:

- *Conditional Probability Invariance:* the conditional probability of Y given C

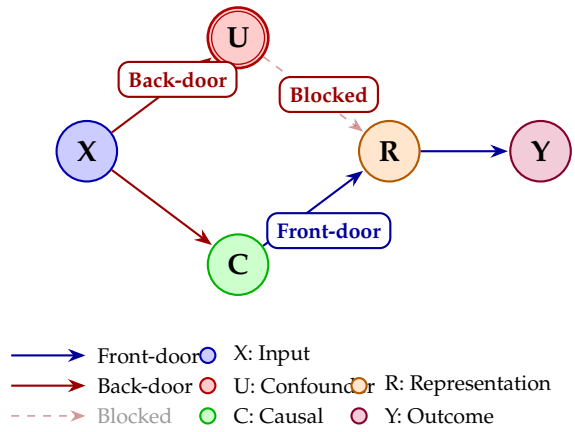


Figure B.1: Structural Causal Model for mental disorder contributor analysis illustrating front-door ($C \to R \to Y$) and back-door ($C \leftarrow X \to U \to R \to Y$) paths. The $U \to R$ path is blocked by design.

and U is invariant under intervention on C:

$$\mathbb{P}_m(Y|C,u) = \mathbb{P}(Y|C,u)$$

- *Marginal Probability Intervention:* The marginal probability of U remains unaffected by the intervention:

$$\mathbb{P}_m(u) = \mathbb{P}(u)$$

- *Independence Under Intervention:* C and U are independent under intervention on C:

$$\mathbb{P}_m(u|C) = \mathbb{P}_m(u) = \mathbb{P}(u)$$

Using these properties, we can deduce the backdoor adjustment as follows:

$$P(Y|\text{do}(C)) = \mathbb{P}_m(Y|C) = \sum_{u \in \mathbf{U}} \mathbb{P}_m(Y|C,u) \cdot \mathbb{P}_m(u|C) \qquad \textit{(Bayes' Rule)} \quad \text{(B.2)}$$

$$= \sum_{u} \mathbb{P}_m(Y|C,u) \cdot \mathbb{P}_m(u) \qquad \textit{(Independence)} \quad \text{(B.3)}$$

$$= \sum_{u} \mathbb{P}(Y|C,u) \cdot \mathbb{P}(u) \qquad \textit{(Invariance)} \quad \text{(B.4)}$$

Equation (B.5), known as the *backdoor adjustment* [188], is a foundational technique for eliminating confounding bias in causal inference.

$$\mathbb{P}(Y|\text{do}(C)) = \sum_{u \in \mathbf{U}} \mathbb{P}(Y|C, U=u) \cdot \mathbb{P}(U=u) \qquad \text{(B.5)}$$

However, in practice, both C and U are typically latent, and identifying them requires evaluating subsets of X, which is computationally infeasible due to its exponential complexity of $\mathcal{O}(2^N)$.

B.2 Auxiliary Theorems

Thorem B.2.1 (Universal Approximation Theorem (UAP) [58, 105]). *Let σ be a continuous sigmoidal function. Then finite sums of the form*

$$G(x) = \sum_{j=1}^{N} \alpha_j \, \sigma(w_j \cdot x + \theta_j), \tag{B.6}$$

where w_j and θ_j are parametric terms, and α_j are coefficients. Given any $f \in \mathcal{C}(I_n)$, and $\epsilon > 0$, there is a sum, $G(x)$, of the above form, for which

$$\left|G(x) - f(x)\right| < \varepsilon, \ \forall x \in I_n \tag{B.7}$$

Alternatively, single hidden layer feedforward networks can approximate any measurable function arbitrarily well, regardless of the continuous nonconstant function used, regardless of the dimension of the input space, and regardless of the input space environment. In this precise and satisfying sense, neural networks are universal approximators.

Thorem B.2.2 (Hyperplane Separation Theorem). *[29][1] Let A and B be two disjoint nonempty convex subsets of $\mathbb{R}^n$. Then, there exist a nonzero vector v and a real number c such that:*

$$\langle x, v\rangle \geq c \quad \textit{and} \quad \langle y, v\rangle \leq c \tag{B.8}$$

$\forall (x, y) \in A \times B$; i.e. the hyperplane $\langle \cdot, v\rangle = c, v$ the normal vector, separates A and B.

[1] www.en.wikipedia.org/wiki/Hyperplane_separation_theorem#cite_note-4

Acronyms

ACR Average Closeness Rate. 12, 53

ADHD Attention Deficit Hyperactivity Disorder. 48, 49, 58, 60

ADODL Average Difference of Overall Depression Levels. 54

AHR Average Hit Rate. 12, 53

AI Artificial Intelligence. 2, 3, 5–7, 10, 11, 13, 14, 19, 64, 89, 117

ANEW Affective Norms for English Words. 42, 48, 49

API Application Programming Interface. 72, 73, 79, 129

BAI Beck Anxiety Inventory. 15

BART Bidirectional and Auto-Regressive Transformers. 3, 26, 35

BDI Beck's Depression Inventory. 15, 43, 53, 56, 57, 66

BERT Bidirectional Encoder Representations from Transformers. 21, 25–27, 32, 34–36, 39, 40, 43, 49, 81–83, 110, 111

BiGRU Bidirectional Gated Recurrent Unit. 24, 27, 34, 35, 81–83, 110

BiLSTM Bidirectional Long Short-Term Memory. 23, 24, 30, 35, 38, 40, 48, 69, 81–83, 88, 104, 110, 148

BoW Bag of Words. 20

CBOW Continuous Bag of Words. 21, 69

CES-D The Center for Epidemiologic Studies Depression. 15

CHARLS China Health and Retirement Longitudinal Study. 46

CNN Convolutional Neural Network. 22, 23, 35, 36, 38, 40, 81–83, 110, 111, 117

CS Conversational Systems. 91–93, 102, 104, 111, 115

DCHR Depression Category Hit Rate. 12, 53

DDA Depressive Disorder Annotation. 66

DL Deep Learning. 15, 22, 24, 28, 42–44, 63, 64, 117

DNN Deep Neural Network. 22, 30, 51, 63

DODL Difference between Overall Depression Level. 54

DS Disentanglement Score. 109, 113–115

DSM-5 Diagnostic and Statistical Manual of Mental Disorders 5. 15, 79

DySAT Dynamic Self-Attention Network. 28

EAT-26 Eating Attitudes Test 26. 15

ECG Electrocardiogram. 3

ED Eating Disorders. 15

EDE-Q Eating Disorder Examination Questionnaire. 15

ERDE Early Risk Detection Error. 12, 53

FN False Negative. 54

FP False Positive. 54

FQ Fear Questionnaire. 46

GAD-7 Generalised Anxiety Disorder 7. 15, 46

GAT Graph Attention Network. 28–30

GCN Graph Convolutional Network. 21, 28–32, 34, 35, 38, 41, 45, 50, 52

GINE Graph Isomorphism Network with Edge features. 76, 88, 126

GloVe Global Vectors for Word Representation. 21, 32, 41

GNN Graph Neural Network. vii, 12, 28–30, 32, 33, 37, 50–52, 62, 80, 85, 125, 130

GPT Generative Pre-trained Transformer. 49, 97, 110, 113

GRU Gated Recurrent Unit. 3, 22–24, 81–83, 110, 117

HAN Hierarchical Attention Network. 81–83, 132, 134–136

HCN Hierarchical Convolutional Network. 81–83

HDRS Hamilton Depression Rating Scale. 15

ICE Individual Conditional Expectation. vii, 12, 45, 49, 50

IG Integrated Gradients. 131–135

KW Kruskal–Wallis. 85

LASSO Least Absolute Shrinkage and Selection Operator. 46

LDA Latent Dirichlet Allocation. 20, 37, 42

LIME Local Interpretable Model-agnostic Explanations. vii, 11, 44–48, 50, 118, 131–135

LIWC Linguistic Inquiry and Word Count. 21, 39, 42, 48, 49, 99, 101, 104, 105, 113, 114, 148, 149

LLM Large Language Model. vii, 5–7, 52, 72, 73, 75, 92, 94, 97, 98, 149, 151

LR Logistic Regression. 110

LRP Layer-wise Relevance Propagation. vii, 45, 49, 50, 131, 132, 134, 135

LSTM Long Short-Term Memory. 3, 4, 6, 22–24, 26, 29, 32, 35–39, 80–83, 85, 110, 111, 117

MDD Mental Disorder Detection. vii, 11, 12, 14–16, 19–22, 24, 26, 27, 30–34, 42, 43, 45, 50, 53, 61–64, 151

MH Mental Health. ii, ix, 1, 2, 4–6, 9, 10, 19, 91, 118–123, 140, 146–148, 150

MHKG Mental Health Knowledge Graph. 8, 71, 72, 76, 78, 84, 87, 88, 125, 148, 150

MHS Mental Health Assessment. 120, 121

ML Machine Learning. 14, 20–22, 34, 46, 53, 62, 119

MLP Multi Layer Perceptron. 23, 37, 38, 40, 41, 76

MRI Magnetic resonance imaging. 3

NLP Natural Language Processing. 11, 14, 20, 24, 25, 48, 121, 146

OCD Obsessive Compulsive Disorder. 18, 58–60

OOV out-of-vocabulary. 69

OSM Online Social Media. ii, 10, 16–20, 22–27, 31–34, 42, 43, 61, 62, 64, 91–93, 97, 98, 102, 104, 108, 111, 115, 117, 151

PCA Principal Component Analysis. 21

PDP Partial Dependence Plots. vii, 12, 45, 49, 50

PDSS Panic Disorder Severity Scale. 46

PHQ-9 Patient Health Questionnaire 9. 5, 15, 46

PRAW Python Reddit API Wrapper. 93

PRISMA Reporting Items for Systematic Reviews and Meta-Analyses. 12, 13

PTSD Post-Traumatic Stress Disorder. 14, 58, 60, 128

RNN Recurrent Neural Network. 22–24, 82, 111

ROC-AUC Receiver Operating Characteristic Area Under the Curve. 53

SHAP SHapley Additive exPlanations. vii, 11, 44–48, 118, 131–135

SIAS Social Interaction Anxiety Scale. 46

SOTA State-Of-The-Art. 14, 43–45, 64, 81

SVM Support Vector Machines. 110, 111

TF-IDF Term Frequency-Inverse Document Frequency. 20, 29, 48

TPE Tree-structured Parzen Estimator. 80

WHO The World Health Organisation. 9

XAI Explainable Artificial Intelligence. ii, ix, 7, 10–15, 20, 34–41, 45, 61, 64, 118, 119, 121, 123, 130, 131, 140, 146, 147, 150

Bibliography

[1] Mohammad Abuhassan, Tarique Anwar, Matthew Fuller-Tyszkiewicz, Hannah K Jarman, Adrian Shatte, Chengfei Liu, and Suku Sukunesan. "Classification of Twitter users with eating disorder engagement: Learning from the biographies". In: *Computers in Human Behavior* 140 (2023), p. 107519.

[2] Mohammad Abuhassan, Tarique Anwar, Chengfei Liu, Hannah K Jarman, and Matthew Fuller-Tyszkiewicz. "EDNet: Attention-Based Multimodal Representation for Classification of Twitter Users Related to Eating Disorders". In: *Proceedings of the ACM Web Conference 2023*. 2023, pp. 4065–4074.

[3] Usman Ahmed, Jerry Chun-Wei Lin, and Gautam Srivastava. "Graph attention network for text classification and detection of mental disorder". In: *ACM Transactions on the Web* 17.3 (2023), pp. 1–31.

[4] Usman Ahmed, Gautam Srivastava, Unil Yun, and Jerry Chun-Wei Lin. "EANDC: An explainable attention network based deep adaptive clustering model for mental health treatment". In: *Future Generation Computer Systems* 130 (2022), pp. 106–113.

[5] Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A Gers. "Visbert: Hidden-state visualizations for transformers". In: *Companion Proceedings of the Web Conference 2020*. 2020, pp. 207–211.

[6] Daniyal Alghazzawi, Hayat Ullah, Naila Tabassum, Sahar K Badri, and Muhammad Zubair Asghar. "Explainable AI-based suicidal and non-suicidal ideations detection from social media text with enhanced ensemble technique". In: *Scientific Reports* 15.1 (2025), p. 1111.

[7] Jussi Alho, Mai Gutvilig, Ripsa Niemi, Kaisla Komulainen, Petri Böckerman, Roger T Webb, Marko Elovainio, and Christian Hakulinen. "Transmission of mental disorders in adolescent peer networks". In: *JAMA psychiatry* 81.9 (2024), pp. 882–888.

[8] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence". In: *Information fusion* 99 (2023), p. 101805.

[9] Hessam Amini and Leila Kosseim. "Towards explainability in using deep learning for the detection of anorexia in social media". In: *Natural Language Processing and Information Systems: 25th International Conference on Applications of Natural Language to Information Systems, NLDB 2020, Saarbrücken, Germany, June 24–26, 2020, Proceedings 25*. Springer. 2020, pp. 225–235.

[10] Noh Amit, Rozmi Ismail, Abdul Rahim Zumrah, Mohd Azmir Mohd Nizah, Tengku Elmi Azlina Tengku Muda, Edbert Chia Tat Meng, Norhayati Ibrahim, and Normah Che Din. "Relationship between debt and depression, anxiety, stress, or suicide ideation in Asia: a systematic review". In: *Frontiers in psychology* 11 (2020), p. 1336.

[11] Lou Ancillon, Mohamed Elgendi, and Carlo Menon. "Machine Learning for Anxiety Detection Using Biosignals: A Review". In: *Diagnostics (Basel)* 12.8 (July 2022).

[12] Carol S Aneshensel, Ralph R Frerichs, and George J Huba. "Depression and physical illness: A multiwave, nonrecursive causal model". In: *Journal of health and social behavior* (1984), pp. 350–371.

[13] Luna Ansari, Shaoxiong Ji, Qian Chen, and Erik Cambria. "Ensemble Hybrid Learning Methods for Automated Depression Detection". In: *IEEE Transactions on Computational Social Systems* 10.1 (2023), pp. 211–219.

[14] Ashutosh Anshul, Gumpili Sai Pranav, Mohammad Zia Ur Rehman, and Nagendra Kumar. "A Multimodal Framework for Depression Detection During COVID-19 via Harvesting Social Media". In: *IEEE Transactions on Computational Social Systems* 11.2 (2024), pp. 2872–2888.

[15] Tarique Anwar, Matthew Fuller-Tyszkiewicz, Hannah K Jarman, Mohammad Abuhassan, Adrian Shatte, WIRED Team, and Suku Sukunesan. "EDBase: Generating a Lexicon Base for Eating Disorders Via Social Me-

dia". In: *IEEE Journal of Biomedical and Health Informatics* 26.12 (2022), pp. 6116–6125.

[16] Tarique Anwar, Surya Nepal, Cecile Paris, Jian Yang, Jia Wu, and Quan Z Sheng. "Tracking the evolution of clusters in social media streams". In: *IEEE Transactions on Big Data* 9.2 (2022), pp. 701–715.

[17] Mario Ezra Aragón, Adrian Pastor López-Monroy, Luis Carlos González-Gurrola, and Manuel Montes-y-Gómez. "Detecting Mental Disorders in Social Media Through Emotional Patterns - The Case of Anorexia and Depression". In: *IEEE Transactions on Affective Computing* 14.1 (2023), pp. 211–222.

[18] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. "One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques". In: *arXiv preprint arXiv:1909.03012* (2019).

[19] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-5™, 5th ed.* American Psychiatric Publishing, Inc., 2013, pp. xliv, 947.

[20] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation". In: *PLOS ONE* 10.7 (2015), pp. 1–46.

[21] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *arXiv:1409.0473* (2014).

[22] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. "A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity". In: *arXiv preprint arXiv:2302.04023* (2023).

[23] Ulya Bayram and Lamia Benhiba. "Emotionally-Informed Models for Detecting Moments of Change and Suicide Risk Levels in Longitudinal Social Media Data". In: *CLPsych*. 2022, pp. 219–225.

[24] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh. "An inventory for measuring depression." In: *Archives of General Psychiatry* 4 (1961), pp. 561–571.

[25] Aaron T Beck, Robert A Steer, Roberta Ball, and William F Ranieri. "Comparison of Beck Depression Inventories-IA and-II in psychiatric outpatients". In: *Journal of Personality Assessment* 67.3 (1996), pp. 588–597.

[26] Rachael E Belcher, Danielle Sim, Marcella Meykler, Jeunice Owens-Walton, Naeemul Hassan, Rachel S Rubin, and Rena D Malik. "A qualitative analysis of female Reddit users' experiences with low libido: how do women perceive their changes in sexual desire?" In: *The Journal of Sexual Medicine* 20.3 (Jan. 2023), pp. 287–297. ISSN: 1743-6095. DOI: 10.1093/jsxmed/qdac045. eprint: https://academic.oup.com/jsm/article-pdf/20/3/287/49344705/qdac045.pdf. URL: https://doi.org/10.1093/jsxmed/qdac045.

[27] Shalini Bhatia, Munawar Hayat, and Roland Goecke. "A Multimodal System to Characterise Melancholia: Cascaded Bag of Words Approach". In: *ACM ICMI*. 2017, pp. 274–280.

[28] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. "Enriching word vectors with subword information". In: *Transactions of the association for computational linguistics* 5 (2017), pp. 135–146.

[29] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[30] Margaret M. Bradley and Peter J. Lang. "Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings". In: 1999.

[31] Sarah Bridges and Richard Disney. "Journal of health economics". In: *Journal of Health Econ.* 29.3 (2010), pp. 388–403.

[32] Egon L van den Broek, Frans van der Sluis, and Ton Dijkstra. "Cross-validation of bimodal health-related stress assessment". In: *Personal and Ubiquitous Computing* 17.2 (2013), pp. 215–227.

[33] Sarah Brown, Karl Taylor, and Stephen Wheatley Price. "Debt and distress: Evaluating the psychological cost of credit". In: *Journal of Economic Psychology* 26.5 (2005), pp. 642–663.

[34] Ana-Maria Bucur, Ioana R Podină, and Liviu P Dinu. "A Psychologically Informed Part-of-Speech Analysis of Depression in Social Media". In: *RANLP*. 2021, pp. 199–207.

[35] Pete Burnap, Walter Colombo, and Jonathan Scourfield. "Machine Classification and Analysis of Suicide-Related Communication on Twitter". In: *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. HT '15. ACM, Aug. 2015, pp. 75–84.

[36] Pere-Lluís Huguet Cabot and Roberto Navigli. "REBEL: Relation extraction by end-to-end language generation". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. 2021, pp. 2370–2381.

[37] Sonia Camacho, Khaled Hassanein, and Milena Head. "Cyberbullying impacts on victims' satisfaction with information and communication technologies: The role of Perceived Cyberbullying Severity". In: *Information and Management* 55.4 (2018), pp. 494–507.

[38] Elena Campillo-Ageitos, Juan Martinez-Romo, and Lourdes Araujo. *UNED-MED at eRisk 2022: depression detection with TF-IDF, linguistic features and Embeddings*. 2022.

[39] Lei Cao, Huijun Zhang, and Ling Feng. "Building and Using Personal Knowledge Graph to Improve Suicidal Ideation Detection on Social Media". In: *IEEE Transactions on Multimedia* 24 (2022), pp. 87–102.

[40] Lei Cao, Huijun Zhang, Ling Feng, Zihan Wei, Xin Wang, Ningyun Li, and Xiaohao He. "Latent Suicide Risk Detection on Microblog via Suicide-Oriented Word Embeddings and Layered Attention". In: *EMNLP-IJCNLP*. 2019, pp. 1718–1728.

[41] Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. "Artificial intelligence, bias and clinical safety". In: *BMJ quality & safety* 28.3 (2019), pp. 231–237.

[42] Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. "Hyperbolic Graph Convolutional Neural Networks". In: *NeurIPS*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. 2019.

[43] Stevie Chancellor and Munmun De Choudhury. "Methods in predictive techniques for mental health status on social media: a critical review". In: *NPJ digital medicine* 3.1 (2020), p. 43.

[44] Stevie Chancellor, Tanushree Mitra, and Munmun De Choudhury. "Recovery Amid Pro-Anorexia: Analysis of Recovery in Social Media". In: *ACM CHI*. 2016, pp. 2111–2123.

[45] Chun-Hao Chang, Elvis Saravia, and Yi-Shin Chen. "Subconscious Crowdsourcing: A feasible data collection mechanism for mental disorder detection on Social media". In: *IEEE/ACM ASONAM*. 2016, pp. 374–379.

[46] Keng-Hao Chang, Matthew K Chan, and John Canny. "AnalyzeThis: unobtrusive mental health monitoring by voice". In: *CHI '11 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '11. New York, NY, USA: ACM, May 2011, pp. 1951–1956.

[47] Guoxin Chen, Yongqing Wang, Fangda Guo, Qinglang Guo, Jiangli Shao, Huawei Shen, and Xueqi Cheng. "Causality and Independence Enhancement for Biased Node Classification". In: CIKM '23. 2023, pp. 203–212.

[48] Min Chen, Ke Shen, Rui Wang, Yiming Miao, Yingying Jiang, Kai Hwang, Yixue Hao, Guangming Tao, Long Hu, and Zhongchun Liu. "Negative Information Measurement at AI Edge: A New Perspective for Mental Health Monitoring". In: *ACM Transactions Internet Technol.* 22.3 (2022), pp. 1–16.

[49] Fahad Riaz Choudhry, Vasudevan Mani, Long Chiau Ming, and Tahir Mehmood Khan. "Beliefs and perception about mental health issues: a meta-synthesis". In: *Neuropsychiatric disease and treatment* (2016), pp. 2807–2818.

[50] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. "Empirical evaluation of gated recurrent neural networks on sequence modeling". In: *arXiv preprint arXiv:1412.3555* (2014).

[51] Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. "SMHD: a Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions". In: *COLING*. 2018, pp. 1485–1497.

[52] Kenneth R Conner, Martin Pinquart, and Stephanie A Gamble. "Meta-analysis of depression and substance use among individuals with alcohol use disorders". In: *Journal of substance abuse treat.* 37.2 (2009), pp. 127–137.

[53] Zafra Cooper and Christopher Fairburn. "The eating disorder examination: A semi-structured interview for the assessment of the specific psychopathology of eating disorders". In: *International Journal of Eating Disorders* 6.1 (1987), pp. 1–8.

[54] Glen Coppersmith, Mark Dredze, and Craig Harman. "Quantifying Mental Health Signals in Twitter". In: *CLPsych*. 2014, pp. 51–60.

[55] Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. "From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses". In: *CLPsych*. 2015, pp. 1–10.

[56] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. "CLPsych 2015 Shared Task: Depression and PTSD on Twitter". In: *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Denver, Colorado: Association for Computational Linguistics, June 2015,

pp. 31–39. DOI: 10.3115/v1/W15-1204. URL: https://aclanthology.org/W15-1204/.

[57] Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. "Exploratory Analysis of Social Media Prior to a Suicide Attempt". In: *CLPsych*. 2016, pp. 106–117.

[58] George Cybenko. "Approximation by superpositions of a sigmoidal function". In: *Mathematics of control, signals and systems* 2.4 (1989), pp. 303–314.

[59] Simon D'Alfonso. "AI in mental health". In: *Current Opinion in Psychology* 36 (2020), pp. 112–117.

[60] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. "A Survey of the State of Explainable AI for Natural Language Processing". In: *IJCNLP*. 2020, pp. 447–459.

[61] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. "Predicting depression via social media". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 7. 1. 2013, pp. 128–137.

[62] Orianna DeMasi and Benjamin Recht. "A step towards quantifying when an algorithm can and cannot predict an individual's wellbeing". In: *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. UbiComp '17. New York, NY, USA: ACM, Sept. 2017, pp. 763–771.

[63] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. "Toxicity in chatgpt: Analyzing persona-assigned language models". In: *arXiv preprint arXiv:2304.05335* (2023).

[64] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *NAACL*. June 2019, pp. 4171–4186.

[65] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. "ERASER: A Benchmark to Evaluate Rationalized NLP Models". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, July 2020, pp. 4443–4458. DOI: 10.18653/v1/2020.acl-main.408. URL: https://aclanthology.org/2020.acl-main.408/.

[66] Kodati Dheeraj and Tene Ramakrishnudu. "Negative emotions detection on online mental-health related patients texts using the deep learning with MHA-BCNN model". In: *Expert Systems with Applications* 182 (2021), p. 115265.

[67] Sahraoui Dhelim, Liming Chen, Sajal K. Das, Huansheng Ning, Chris Nugent, Gerard Leavey, Dirk Pesch, Eleanor Bantry-White, and Devin Burns. "Detecting Mental Distresses Using Social Behavior Analysis in the Context of COVID-19: A Survey". In: *ACM Comput. Surv.* 55.14s (2023).

[68] Nina H Di Cara, Valerio Maggio, Oliver SP Davis, and Claire MA Haworth. "Methodologies for monitoring mental health on twitter: systematic review". In: *Journal of Medical Internet Research* 25 (2023), e42734.

[69] Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. "A Survey of Natural Language Generation". In: 55.8 (Dec. 2022).

[70] Tobias Esch, George B Stefano, Gregory L Fricchione, and Herbert Benson. "The role of stress in neurodegenerative diseases and mental disorders". In: *Neuroendocrinology letters* 23.3 (2002), pp. 199–208.

[71] Catherine K. Ettman, Alice Y. Fan, Alexander P. Philips, Gaelen P. Adam, Grace Ringlein, Melissa A. Clark, Ira B. Wilson, Patrick M. Vivier, and Sandro Galea. "Financial strain and depression in the U.S.: a scoping review". In: *Translational Psychiatry* 13.1 (May 2023), p. 168.

[72] Yucai Fan, Yuhang Yao, and Carlee Joe-Wong. "GCN-SE: Attention as Explainability for Node Classification in Dynamic Graphs". In: *ICDM*. 2021, pp. 1060–1065.

[73] Ethan Fast, Binbin Chen, and Michael S Bernstein. "Empath: Understanding topic signals in large-scale text". In: *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2016, pp. 4647–4657.

[74] "Financial threat, hardship and distress predict depression, anxiety and stress among the unemployed youths: A Bangladeshi multi-city study". In: *Jour. of Affect. Disor.* 276 (2020), pp. 1149–1158.

[75] Chris Finlay and Adam M Oberman. "Scaleable input gradient regularization for adversarial robustness". In: *Machine Learning with Applications* 3 (2021), p. 100017.

[76] Jerome H. Friedman. "Greedy function approximation: A gradient boosting machine." In: *The Annals of Statistics* 29.5 (2001), pp. 1189–1232.

[77] Andrea Galassi, Marco Lippi, and Paolo Torroni. "Attention in natural language processing". In: *IEEE transactions on neural networks and learning systems* 32.10 (2020), pp. 4291–4308.

[78] Kyle T Ganson, Alexander C Tsai, Sheri D Weiser, Samuel E Benabou, and Jason M Nagata. "Job insecurity and symptoms of anxiety and depression among US young adults during COVID-19". In: *Journal of Adolescent Health* 68.1 (2021), pp. 53–56.

[79] Muskan Garg, Chandni Saxena, Sriparna Saha, Veena Krishnan, Ruchi Joshi, and Vijay Mago. "CAMS: An Annotated Corpus for Causal Analysis of Mental Health Issues in Social Media Posts". In: *Procc of the 13th LREC*. June 2022, pp. 6387–6396.

[80] David M. Garner and Paul E. Garfinkel. "The Eating Attitudes Test: an index of the symptoms of anorexia nervosa". In: *Psychological Med.* 9.2 (1979), pp. 273–279.

[81] Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. "Knowledge-aware assessment of severity of suicide risk for early intervention". In: *The world wide web conference*. 2019, pp. 514–525.

[82] Lixia Ge, Chun Wei Yap, Reuben Ong, and Bee Hoon Heng. "Social isolation, loneliness and their relationships with depressive symptoms: A population-based study". In: *PloS one* 12.8 (2017), e0182145.

[83] Shreya Ghosh and Tarique Anwar. "Depression Intensity Estimation via Social Media: A Deep Learning Approach". In: *IEEE Transactions on Computational Social Systems* 8.6 (2021), pp. 1465–1474.

[84] George Gkotsis, Anika Oellrich, Tim Hubbard, Richard Dobson, Maria Liakata, Sumithra Velupillai, and Rina Dutta. "The language of mental health problems in social media". In: *Proceedings of the third workshop on computational linguistics and clinical psychology*. 2016, pp. 63–73.

[85] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation". In: *journal of Computational and Graphical Statistics* 24.1 (2015), pp. 44–65.

[86] Narayan Gopalkrishnan. "Cultural diversity and mental health: Considerations for policy and practice". In: *Frontiers in public health* 6 (2018), p. 179.

[87] Brandon M Greenwell, Bradley C Boehmke, and Andrew J McCarthy. "A simple and effective model-based variable importance measure". In: *arXiv:1805.04755* (2018).

[88] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. "Measuring statistical dependence with Hilbert-Schmidt norms". In: *ALT*. 2005, pp. 63–77.

[89] Patricia Gual-Montolio, Irene Jaén, Verónica Martínez-Borba, Diana Castilla, and Carlos Suso-Ribera. "Using Artificial Intelligence to Enhance Ongoing Psychological Interventions for Emotional Problems in Real- or Close to Real-Time: A Systematic Review". In: *International Journal of Environmental Research and Public Health* 19.13 (2022).

[90] Sharath Chandra Guntuku, Anneke Buffone, Kokil Jaidka, Johannes C Eichstaedt, and Lyle H Ungar. "Understanding and measuring psychological stress using social media". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 13. 2019, pp. 214–225.

[91] Sharath Chandra Guntuku, J Russell Ramsay, Raina M Merchant, and Lyle H Ungar. "Language of ADHD in adults on social media". In: *Journal of attention disorders* 23.12 (2019), pp. 1475–1485.

[92] Ian Hacking. "The looping effects of human kinds". In: *Causal Cognition: A Multidisciplinary Debate*. Oxford University Press, July 1996. ISBN: 9780198524021. DOI: 10.1093/acprof:oso/9780198524021.003.0012. eprint: https://academic.oup.com/book/0/chapter/194529638/chapter-ag-pdf/44619379/book_26284_section_194529638.ag.pdf. URL: https://doi.org/10.1093/acprof:oso/9780198524021.003.0012.

[93] Lauryn J Hagg, Stephanie S Merkouris, Gypsy A O'Dea, Lauren M Francis, Christopher J Greenwood, Matthew Fuller-Tyszkiewicz, Elizabeth M Westrupp, Jacqui A Macdonald, and George J Youssef. "Examining Analytic Practices in Latent Dirichlet Allocation Within Psychological Science: Scoping Review". In: *J. Med. Internet Res.* 24.11 (2022), e33166.

[94] William L. Hamilton, Rex Ying, and Jure Leskovec. "Inductive Representation Learning on Large Graphs". In: *NeurIPS*. 2017, pp. 1025–1035.

[95] Sooji Han, Rui Mao, and Erik Cambria. "Hierarchical Attention Network for Explainable Depression Detection on Twitter Aided by Metaphor Concept Mappings". In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International

Committee on Computational Linguistics, Oct. 2022, pp. 94–104. URL: https://aclanthology.org/2022.coling-1.9/.

[96] Ayaan Haque, Viraaj Reddi, and Tyler Giallanza. "Deep Learning for Suicide and Depression Identification with Unsupervised Label Correction". In: *ICANN*. 2021, pp. 436–447.

[97] Khan Md Hasib, Md Rafiqul Islam, Shadman Sakib, Md. Ali Akbar, Imran Razzak, and Mohammad Shafiul Alam. "Depression Detection From Social Networks Data Based on Machine Learning and Deep Learning Techniques: An Interrogative Survey". In: *IEEE Transactions on Computational Social Systems* 10.4 (2023), pp. 1568–1586.

[98] Colleen M Heflin and John Iceland. "Poverty, material hardship, and depression". In: *Social science quarterly* 90.5 (2009), pp. 1051–1071.

[99] Bert Heinrichs. "What Is Discrimination and When Is It Morally Wrong?" In: *Jahrbuch für Wissenschaft und Ethik* 12.1 (2007), pp. 97–114.

[100] Claire Henderson, Laura Potts, and Emily J Robinson. "Mental illness stigma after a decade of Time to Change England: inequalities as targets for further improvement". In: *European journal of public health* 30.3 (2020), pp. 497–503.

[101] Paula L Hensley. "Treatment of bereavement-related depression and traumatic grief". In: *Journal of Affective disorders* 92.1 (2006), pp. 117–124.

[102] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[103] Matthew Holland. "Robustness and scalability under heavy tails, without strong convexity". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 865–873.

[104] Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. "exbert: A visual analysis tool to explore learned representations in transformers models". In: *arXiv preprint arXiv:1910.05276* (2019).

[105] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators". In: *Neural networks* 2.5 (1989), pp. 359–366.

[106] Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-Excitation Networks". In: *IEEE/CVF CVPR*. 2018, pp. 7132–7141.

[107] Lijie Hu, Yixin Liu, Ninghao Liu, Mengdi Huai, Lichao Sun, and Di Wang. "Seat: stable and explainable attention". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 11. 2023, pp. 12907–12915.

[108] Yusif Ibrahimov, Tarique Anwar, and Tommy Yuan. "DepressionX: Knowledge Infused Residual Attention for Explainable Depression Severity Assessment". In: *Proceedings of the AAAI workshop on Health Intelligence (W3PHIAI)*. 2025. URL: arXiv%20preprint%20arXiv:2501.14985.

[109] Yusif Ibrahimov, Tarique Anwar, and Tommy Yuan. "Explainable AI for Mental Disorder Detection via Social Media: A survey and outlook". In: *arXiv preprint arXiv:2406.05984* (2024).

[110] Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. "Understanding convolutional neural networks for text classification". In: *arXiv preprint arXiv:1809.08037* (2018).

[111] Sarthak Jain and Byron C. Wallace. "Attention is not Explanation". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 3543–3556. DOI: 10.18653/v1/N19-1357. URL: https://aclanthology.org/N19-1357/.

[112] Hannah K Jarman, Siân A McLean, Scott Griffiths, Samantha J Teague, Rachel F Rodgers, Susan J Paxton, Emma Austen, Emily Harris, Trevor Steward, Adrian Shatte, et al. "Critical measurement issues in the assessment of social media influence on body image". In: *Body Image* 40 (2022), pp. 225–236.

[113] Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. "MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, June 2022, pp. 7184–7190. URL: https://aclanthology.org/2022.lrec-1.778/.

[114] Yan Jia, John McDermid, Tom Lawton, and Ibrahim Habli. "The role of explainability in assuring safety of machine learning in healthcare". In: *IEEE Transactions on Emerging Topics in Computing* 10.4 (2022), pp. 1746–1760.

[115] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. "Mistral 7B". In: *arXiv preprint arXiv:2310.06825* (2023).

[116] Zhengping Jiang, Sarah Ita Levitan, Jonathan Zomick, and Julia Hirschberg. "Detection of Mental Health from Reddit via Deep Contextualized Representations". In: *LOUHI*. ACL, 2020, pp. 147–156.

[117] Anna Jobin, Marcello Ienca, and Effy Vayena. "The global landscape of AI ethics guidelines". In: *Nature machine intelligence* 1.9 (2019), pp. 389–399.

[118] Amy K Johnson, Runa Bhaumik, Debarghya Nandi, Abhishikta Roy, and Supriya D Mehta. "Sexually Transmitted Disease–Related Reddit Posts During the COVID-19 Pandemic: Latent Dirichlet Allocation Analysis". In: *Journal of medical Internet research* 24.10 (2022), e37258.

[119] Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. "How large language models encode context knowledge? a layer-wise probing study". In: *arXiv preprint arXiv:2402.16061* (2024).

[120] Elham Kadkhoda, Mahsa Khorasani, Fatemeh Pourgholamali, Mohsen Kahani, and Amir Rezaei Ardani. "Bipolar disorder detection over social media". In: *Informatics in Medicine Unlocked* 32 (2022), p. 101042.

[121] Simranjeet Kaur, Ritika Bhardwaj, Astha Jain, Muskan Garg, and Chandni Saxena. "Causal categorization of mental health posts using transformers". In: *Proc. of the 14th FIRE*. 2022, pp. 43–46.

[122] Pei Ke, Haozhe Ji, Siyang Liu, Xiaoyan Zhu, and Minlie Huang. "SentiLARE: Sentiment-Aware Language Representation Learning with Linguistic Knowledge". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 6975–6988. DOI: 10.18653/v1/2020.emnlp-main.567. URL: https://aclanthology.org/2020.emnlp-main.567/.

[123] Anthony Kelly, Esben Kjems Jensen, Eoin Martino Grua, Kim Mathiasen, and Pepijn Van de Ven. "An Interpretable Model With Probabilistic Integrated Scoring for Mental Health Treatment Prediction: Design Study". In: *JMIR Medical Informatics* 13 (2025), e64617.

[124] Elma Kerz, Sourabh Zanwar, Yu Qiao, and Daniel Wiechmann. "Toward explainable AI (XAI) for mental health detection based on language behavior". In: *Frontiers in psychiatry* 14 (2023), p. 1219479.

[125] Soheyl Khalilpourazari, Saman Khalilpourazary, Aybike Özyüksel Çiftçioğlu, and Gerhard-Wilhelm Weber. "Designing energy-efficient high-precision multi-pass turning processes via robust optimization and artificial in-

telligence". In: *Journal of Intelligent Manufacturing* 32.6 (2021), pp. 1621–1647.

[126] Mizanur Khondoker, Richard Dobson, Caroline Skirrow, Andrew Simmons, and Daniel Stahl. "A comparison of machine learning methods for classification using simulation with multiple real data examples from mental health studies". In: *Stat. Methods Med. Res.* 25.5 (Oct. 2016), pp. 1804–1823.

[127] Thomas N. Kipf and Max Welling. "Semi-Supervised Classification with Graph Convolutional Networks". In: *ICLR*. 2017.

[128] Noona Kiuru, William J Burk, Brett Laursen, Jari-Erik Nurmi, and Katariina Salmela-Aro. "Is depression contagious? A test of alternative peer socialization mechanisms of depressive symptoms in adolescent peer networks". In: *Journal of Adolescent Health* 50.3 (2012), pp. 250–255.

[129] Kurt Kroenke, Robert L. Spitzer, and Janet B. W. Williams. "The PHQ-9: Validity of a brief depression severity measure." In: *Journal of General Internal Med.* 16.9 (2001), pp. 606–613.

[130] Adam M Kuczynski, Max A Halvorson, Lily R Slater, and Jonathan W Kanter. "The effect of social interaction quantity and quality on depressed mood and loneliness: A daily diary study". In: *Journal of Social and Personal Relationships* 39.3 (2022), pp. 734–756.

[131] Ai-Te Kuo, Haiquan Chen, Yu-Hsuan Kuo, and Wei-Shinn Ku. "Dynamic Graph Representation Learning for Depression Screening with Transformer". In: *arXiv:2305.06447* (2023).

[132] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. "Gradient-based learning applied to document recognition". In: *Proc. of the IEEE* 86.11 (1998), pp. 2278–2324.

[133] Sang Ah Lee, Yeong Jun Ju, Kyu-Tae Han, Jae Woo Choi, Hyo Jung Yoon, and Eun-Cheol Park. "The association between loss of work ability and depression: a focus on employment status". In: *Inter. arch. of occup. and environ. health* 90 (2017), pp. 109–116.

[134] Sherman A. Lee. "Coronavirus Anxiety Scale: A brief mental health screener for COVID-19 related anxiety". In: *Death Studies* 44.7 (2020), pp. 393–401.

[135] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. URL: https://aclanthology.org/2020.acl-main.703/.

[136] Ying Liang and Struther Van Horn. "How do romantic breakups affect depression among American college students? The role of sexual conservativeness". In: *Journal of American college health* 70.4 (2022), pp. 1019–1029.

[137] Yan Qian Lim, Ming Jie Lee, and Yim Ling Loo. "Towards A Machine Learning Framework for Suicide Ideation Detection in Twitter". In: *AiDAS*. 2022, pp. 153–157.

[138] Deyan Liu, Yuge Tian, Min Liu, and Shangjian Yang. "Developing an interpretable machine learning model for screening depression in older adults with functional disability". In: *Journal of Affective Disorders* 379 (2025), pp. 529–539. ISSN: 0165-0327. DOI: https://doi.org/10.1016/j.jad.2025.02.110. URL: https://www.sciencedirect.com/science/article/pii/S0165032725003271.

[139] Shengzhong Liu, Franck Le, Supriyo Chakraborty, and Tarek Abdelzaher. "On Exploring Attention-based Explanation for Transformer Models in Text Classification". In: *2021 IEEE International Conference on Big Data (Big Data)*. 2021, pp. 1193–1203. DOI: 10.1109/BigData52589.2021.9671639.

[140] Shikang Liu, Fatemeh Vahedian, David Hachen, Omar Lizardo, Christian Poellabauer, Aaron Striegel, and Tijana Milenković. "Heterogeneous network approach to predict individuals' mental health". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15.2 (2021), pp. 1–26.

[141] Xien Liu, Xinxin You, Xiao Zhang, Ji Wu, and Ping Lv. "Tensor graph convolutional networks for text classification". In: *AAAI*. Vol. 34. 05. 2020, pp. 8409–8416.

[142] Yu Liu, Chen Xu, Xi Kuai, Hao Deng, Kaifeng Wang, and Qinyao Luo. "Analysis of the Causes of Inferiority Feelings Based on Social Media Data with Word2Vec". In: *Scientific Reports* 12.1 (2022), p. 5218.

[143] David E Losada and Fabio Crestani. "A test collection for research on depression and language use". In: *International conference of the cross-language evaluation forum for European languages*. Springer. 2016, pp. 28–39.

[144] David E Losada, Fabio Crestani, and Javier Parapar. "Overview of eRisk: early risk prediction on the internet". In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings 9*. Springer. 2018, pp. 343–361.

[145] David E. Losada, Fabio Crestani, and Javier Parapar. "eRisk 2020: Self-harm and Depression Challenges". In: *Advances in Information Retrieval*. 2020, pp. 557–563.

[146] David E. Losada, Fabio Crestani, and Javier Parapar. "Overview of ERisk 2019 Early Risk Prediction on the Internet". In: *CLEF*. 2019, pp. 340–357.

[147] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *NeurIPS*. 2017, pp. 4768–4777.

[148] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. "Parameterized Explainer for Graph Neural Network". In: *NeurIPS*. Vol. 33. 2020, pp. 19620–19631.

[149] Jian Ma, Lei Wang, Yuan-Rong Zhang, Wei Yuan, and Wei Guo. "An integrated latent Dirichlet allocation and Word2vec method for generating the topic evolution of mental models from global to local". In: *Expert Systems with Applications* 212 (2023), p. 118695.

[150] Sean MacAvaney, Anjali Mittu, Glen Coppersmith, Jeff Leintz, and Philip Resnik. "Community-level Research on Suicidality Prediction in a Secure Environment: Overview of the CLPsych 2021 Shared Task". In: *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*. Ed. by Nazli Goharian, Philip Resnik, Andrew Yates, Molly Ireland, Kate Niederhoffer, and Rebecca Resnik. Online: Association for Computational Linguistics, June 2021, pp. 70–80. DOI: 10.18653/v1/2021.clpsych-1.7. URL: https://aclanthology.org/2021.clpsych-1.7/.

[151] Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H. Andrew Schwartz. "Suicide Risk Assessment with Multi-level Dual-Context Language and BERT". In: *CLPsych*. 2019, pp. 39–44.

[152] Timothy Matthews, Andrea Danese, Jasmin Wertz, Candice L Odgers, Antony Ambler, Terrie E Moffitt, and Louise Arseneault. "Social isolation, loneliness and depression in young adulthood: a behavioural genetic analysis". In: *Social psychiatry and psychiatric epidemiology* 51 (2016), pp. 339–348.

[153] Matthew Louis Mauriello, Thierry Lincoln, Grace Hon, Dorien Simon, Dan Jurafsky, and Pablo Paredes. "SAD: A Stress Annotated Dataset for Recognizing Everyday Stressors in SMS-like Conversational Systems". In: CHI EA '21. 2021.

[154] Kathleen Ries Merikangas, Jian-ping He, Marcy Burstein, Sonja A Swanson, Shelli Avenevoli, Lihong Cui, Corina Benjet, Katholiki Georgiades, and Joel Swendsen. "Lifetime prevalence of mental disorders in US adolescents: results from the National Comorbidity Survey Replication–Adolescent Supplement (NCS-A)". In: *Journal of the American Academy of Child & Adolescent Psychiatry* 49.10 (2010), pp. 980–989.

[155] Jonathan M Metzl and Kenneth T MacLeish. "Mental illness, mass shootings, and the politics of American firearms". In: *American journal of public health* 105.2 (2015), pp. 240–249.

[156] Ivan Mihov, Haiquan Chen, Xiao Qin, Wei-Shinn Ku, Da Yan, and Yuhong Liu. "Mentalnet: Heterogeneous graph representation for early depression detection". In: *2022 IEEE International Conference on Data Mining (ICDM)*. IEEE. 2022, pp. 1113–1118.

[157] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space". In: *arXiv:1301.3781* (2013).

[158] Kirill Milintsevich, Kairit Sirts, and Gaël Dias. "Your model is not predicting depression well and that is why: A case study of primate dataset". In: *arXiv preprint arXiv:2403.00438* (2024).

[159] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial Intelligence* 267 (2019), pp. 1–38. ISSN: 0004-3702. DOI: https://doi.org/10.1016/j.artint.2018.07.007. URL: https://www.sciencedirect.com/science/article/pii/S0004370218305988.

[160] Rohan Mishra, Pradyumn Prakhar Sinha, Ramit Sawhney, Debanjan Mahata, Puneet Mathur, and Rajiv Ratn Shah. "SNAP-BATNET: Cascading Author Profiling and Social Network Graphs for Suicide Ideation Detection on Social Media". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Re-*

search Workshop. Ed. by Sudipta Kar, Farah Nadeem, Laura Burdick, Greg Durrett, and Na-Rae Han. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 147–156. DOI: 10.18653/v1/N19-3019. URL: https://aclanthology.org/N19-3019/.

[161] Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. "Quantifying the Language of Schizophrenia in Social Media". In: *CLPsych*. 2015, pp. 11–20.

[162] Vidhi Mody and Vrushti Mody. "Mental Health Monitoring System using Artificial Intelligence: A Review". In: *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*. 2019, pp. 1–6. DOI: 10.1109/I2CT45611.2019.9033652.

[163] Milad Moradi and Matthias Samwald. "Explaining black-box models for biomedical text classification". In: *IEEE journal of biomedical and health informatics* 25.8 (2021), pp. 3112–3120.

[164] Danielle Mowery, Craig Bryan, and Mike Conway. "Towards Developing an Annotation Scheme for Depressive Disorder Symptoms: A Preliminary Study using Twitter Data". In: *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Denver, Colorado: Association for Computational Linguistics, June 2015, pp. 89–98. DOI: 10.3115/v1/W15-1211. URL: https://aclanthology.org/W15-1211/.

[165] Ankit Murarka, Balaji Radhakrishnan, and Sushma Ravichandran. "Classification of mental illnesses on social media using RoBERTa". In: *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*. online: Association for Computational Linguistics, Apr. 2021, pp. 59–68.

[166] Usman Naseem, Adam G Dunn, Jinman Kim, and Matloob Khushi. "Early identification of depression severity levels on reddit using ordinal classification". In: *Proceedings of the ACM Web Conference 2022*. 2022, pp. 2563–2572.

[167] Usman Naseem, Matloob Khushi, Jinman Kim, and Adam G Dunn. "Hybrid text representation for explainable suicide risk identification on social media". In: *IEEE Transactions on Computational Social Systems* (2022).

[168] Usman Naseem, Jinman Kim, Matloob Khushi, and Adam Dunn. "Graph-Based Hierarchical Attention Network for Suicide Risk Detection on Social Media". In: *ACM Web Conference*. 2023, pp. 995–1003.

[169] Usman Naseem and Katarzyna Musial. "DICE: Deep Intelligent Contextual Embedding for Twitter Sentiment Analysis". In: *ICDAR*. 2019, pp. 953–958.

[170] Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh, and Michael Berk. "Affective and Content Analysis of Online Depression Communities". In: *IEEE Transactions on Affec. Computing* 5.3 (2014), pp. 217–226.

[171] Jin-Won Noh, Young Dae Kwon, Jumin Park, In-Hwan Oh, and Jinseok Kim. "Relationship between physical disability and depression by gender: a panel regression model". In: *PloS one* 11.11 (2016), e0166238.

[172] Bridianne O'Dea, Stephen Wan, Philip J. Batterham, Alison L. Calear, Cecile Paris, and Helen Christensen. "Detecting suicidality on Twitter". In: *Internet Interventions* 2.2 (2015), pp. 183–188.

[173] Bridianne O'Dea, Tjeerd W. Boonstra, Mark E. Larsen, Thin Nguyen, Svetha Venkatesh, and Helen Christensen. "The relationship between linguistic expression in blog content and symptoms of depression, anxiety, and suicidal thoughts: A longitudinal study". In: *PLOS ONE* 16.5 (May 2021), pp. 1–17. DOI: 10.1371/journal.pone.0251787. URL: https://doi.org/10.1371/journal.pone.0251787.

[174] Blessing Ojeme and Audrey Mbogho. "Selecting Learning Algorithms for Simultaneous Identification of Depression and Comorbid Disorders". In: *Procedia Comput. Sci.* 96 (Jan. 2016), pp. 1294–1303.

[175] Mark Olfson, Carlos Blanco, and Steven C Marcus. "Treatment of adult depression in the United States". In: *JAMA internal medicine* 176.10 (2016), pp. 1482–1491.

[176] OpenAI. *GPT-3.5 Turbo*. Accessed: 25 April 2025. 2023. URL: https://platform.openai.com/docs/models/gpt-3-5-turbo.

[177] OpenAI. *GPT-4o mini: Smaller, faster, cheaper GPT-4*. Accessed: 25 April 2025. 2024. URL: https://openai.com/blog/gpt-4o-mini.

[178] Yalcin Ozdemir. "Parent-adolescent conflict and depression symptoms of adolescents: Mediator role of self-esteem". In: *Dusunen Adam* 27.3 (2014), p. 211.

[179] Sean-Kelly Palicki, Shereen Fouad, Mariam Adedoyin-Olowe, and Zahraa S Abdallah. "Transfer learning approach for detecting psychological distress in brexit tweets". In: *Proceedings of the 36th Annual ACM Symposium on Applied Computing*. SAC '21. New York, NY, USA: Association for Computing Machinery, Apr. 2021, pp. 967–975.

[180] Andreas S. Panayides, Amir Amini, Nenad D. Filipovic, Ashish Sharma, Sotirios A. Tsaftaris, Alistair Young, David Foran, Nhan Do, Spyretta Golemati, Tahsin Kurc, Kun Huang, Konstantina S. Nikita, Ben P. Veasey, Michalis Zervakis, Joel H. Saltz, and Constantinos S. Pattichis. "AI in Medical Imaging Informatics: Current Challenges and Future Directions". In: *IEEE Journal of Biomedical and Health Informatics* 24.7 (2020), pp. 1837–1857. DOI: 10.1109/JBHI.2020.2991043.

[181] Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. "Overview of erisk 2023: Early risk prediction on the internet". In: *CLEF*. Springer. 2023, pp. 294–315.

[182] Javier Parapar, Patricia Martín-Rodilla, David E. Losada, and Fabio Crestani. "Overview of eRisk 2021: Early Risk Prediction on the Internet". In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Ed. by K. Selçuk Candan, Bogdan Ionescu, Lorraine Goeuriot, Birger Larsen, Henning Müller, Alexis Joly, Maria Maistro, Florina Piroi, Guglielmo Faggioli, and Nicola Ferro. Cham: Springer International Publishing, 2021, pp. 324–344. ISBN: 978-3-030-85251-1.

[183] Pablo Paredes, Ran Gilad-Bachrach, Mary Czerwinski, Asta Roseway, Kael Rowan, and Javier Hernandez. "PopTherapy: coping with stress through pop-culture". In: *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare*. PervasiveHealth '14. Brussels, BEL: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), May 2014, pp. 109–117.

[184] Minsu Park, Chiyoung Cha, and Meeyoung Cha. "Depressive moods of users portrayed in Twitter". In: *Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining, SIGKDD 2012*. 2012, pp. 1–8.

[185] Minsu Park, David McDonald, and Meeyoung Cha. "Perception differences between the depressed and non-depressed users in twitter". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 7. 1. 2013, pp. 476–485.

[186] David A Patterson and Richard N Cloud. "The Application of Artificial Neural Networks for Outcome Prediction in a Cohort of Severely Mentally Ill Outpatients". In: *J. Technol. Hum. Serv.* 16.2-3 (Jan. 2000), pp. 47–61.

[187] Rachelle Pavelko and Jessica Gall Myrick. "Tweeting and Trivializing: How the Trivialization of Obsessive–Compulsive Disorder via Social Media Impacts User Perceptions, Emotions, and Behaviors". In: *Imagination, Cognition and Personality* 36.1 (2016), pp. 41–63.

[188] Judea Pearl. "Causal diagrams for empirical research". In: *Biometrika* 82.4 (1995), pp. 669–688.

[189] Sachin R. Pendse, Kate Niederhoffer, and Amit Sharma. "Cross-Cultural Differences in the Use of Online Mental Health Support Forums". In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (2019).

[190] James W Pennebaker. *Linguistic inquiry and word count: LIWC 2001*. 2001.

[191] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation". In: *EMNLP*. 2014, pp. 1532–1543.

[192] Anxo Pérez, Javier Parapar, and Álvaro Barreiro. "Automatic depression score estimation with word embedding models". In: *AI in Med.* 132 (2022), p. 102380.

[193] Anxo Pérez, Javier Parapar, Álvaro Barreiro, and Silvia Lopez-Larrosa. "BDI-Sen: A Sentence Dataset for Clinical Symptoms of Depression". In: *ACM SIGIR*. 2023, pp. 2996–3006.

[194] Vitali Petsiuk, Abir Das, and Kate Saenko. "Rise: Randomized input sampling for explanation of black-box models". In: *arXiv preprint arXiv:1806.07421* (2018).

[195] Jahandad Pirayesh, Haiquan Chen, Xiao Qin, Wei-Shinn Ku, and Da Yan. "MentalSpot: Effective Early Screening for Depression Based on Social Contagion". In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. CIKM '21. Virtual Event, Queensland, Australia: Association for Computing Machinery, 2021, pp. 1437–1446. ISBN: 9781450384469. DOI: 10.1145/3459637.3482366. URL: https://doi.org/10.1145/3459637.3482366.

[196] Inna Pirina and Çağrı 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*. Ed. by Graciela Gonzalez-Hernandez, Davy Weissenbacher, Abeed Sarker, and Michael Paul. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 9–12. DOI: 10.18653/v1/W18-5903. URL: https://aclanthology.org/W18-5903/.

[197] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. "Manipulating and measuring model interpretability". In: *Proceedings of the 2021 CHI conference on human factors in computing systems*. 2021, pp. 1–52.

[198] Richard H Price, Jin Nam Choi, and Amiram D Vinokur. "Links in the chain of adversity following job loss: how financial strain and loss of personal control lead to depression, impaired functioning, and poor health." In: *Journal of occup. health psychol.* 7.4 (2002), p. 302.

[199] Bhavini Priyamvada, Shruti Singhal, Anand Nayyar, Rachna Jain, Priya Goel, Mehar Rani, and Muskan Srivastava. "Stacked CNN - LSTM approach for prediction of suicidal ideation on social media". In: *Multimed. Tools Appl.* 82.18 (July 2023), pp. 27883–27904.

[200] Milen L. Radell, Eid G. Abo Hamza, Wid H. Daghustani, Asma Perveen, and Ahmed A. Moustafa. "The Impact of Different Types of Abuse on Depression". In: *Depression Research and Treatment* 2021 (2021), p. 6654503.

[201] Lenore Sawyer Radloff. "The CES-D Scale: A Self-Report Depression Scale for Research in the General Population". In: *App. Psycho. Meas.* 1.3 (1977), pp. 385–401.

[202] Waleed Ragheb, Jérôme Azé, Sandra Bringay, and Maximilien Servajean. "Negatively Correlated Noisy Learners for At-Risk User Detection on Social Networks: A Study on Depression, Anorexia, Self-Harm, and Suicide". In: *IEEE TKDE* 35.1 (2023), pp. 770–783.

[203] Jürgen Rehm and Kevin D Shield. "Global burden of disease and the impact of mental and addictive disorders". In: *Current psychiatry reports* 21 (2019), pp. 1–7.

[204] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *ACM SIGKDD*. 2016, pp. 1135–1144.

[205] Thalia Richter, Barak Fishbain, Andrey Markus, Gal Richter-Levin, and Hadas Okon-Singer. "Using Machine Learning-Based Analysis for Behavioral Differentiation Between Anxiety and Depression". In: *Scientific Reports* 10.1 (2020), p. 16381.

[206] J. N. Rosenquist, J. H. Fowler, and N. A. Christakis. "Social Network Determinants of Depression". In: *Molecular Psychiatry* 16.3 (2011), pp. 273–281.

[207] Cynthia Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature machine intelligence* 1.5 (2019), pp. 206–215.

[208] Kayalvizhi S, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, and Jerin Mahibha C. "Findings of the Shared Task on Detecting Signs of Depression from Social Media". In: *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Ed. by Bharathi Raja Chakravarthi, B Bharathi, John P McCrae, Manel Zarrouk, Kalika Bali, and Paul Buitelaar. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 331–338. DOI: 10.18653/v1/2022.ltedi-1.51. URL: https://aclanthology.org/2022.ltedi-1.51/.

[209] Farig Sadeque, Dongfang Xu, and Steven Bethard. "Measuring the latency of depression detection in social media". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 2018, pp. 495–503.

[210] Kayalvizhi Sampath and Thenmozhi Durairaj. "Data set creation and empirical analysis for detecting signs of depression from social media postings". In: *International Conference on Computational Intelligence in Data Science*. Springer. 2022, pp. 136–151.

[211] Catherine Sanchez, Adrienne Grzenda, Andrea Varias, Alik S Widge, Linda L Carpenter, William M McDonald, Charles B Nemeroff, Ned H Kalin, Glenn Martin, Mauricio Tohen, Maria Filippou-Frye, Drew Ramsey, Eleni Linos, Christina Mangurian, and Carolyn I Rodriguez. "Social media recruitment for mental health research: A systematic review". In: *Compr. Psychiatry* 103 (2020), p. 152197.

[212] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *arXiv:1910.01108* (2019), arXiv–1910.

[213] Elvis Saravia, Chun-Hao Chang, Renaud Jollet De Lorenzo, and Yi-Shin Chen. "MIDAS: Mental illness detection and Analysis via Social media". In: *IEEE/ACM ASONAM*. 2016, pp. 1418–1421.

[214] Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. "A Time-Aware Transformer Based Model for Suicide Ideation Detection on Social Media". In: *EMNLP*. 2020, pp. 7685–7697.

[215] Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. "Towards ordinal suicide ideation detection on social media". In: *Proceedings of the 14th ACM international conference on web search and data mining*. 2021, pp. 22–30.

[216] Ramit Sawhney, Harshit Joshi, Rajiv Ratn Shah, and Lucie Flek. "Suicide Ideation Detection via Social and Temporal User Representations using Hyperbolic Learning". In: *NAACL*. 2021, pp. 2176–2190.

[217] Ramit Sawhney, Atula Neerkaje, and Manas Gaur. "A Risk-Averse Mechanism for Suicidality Assessment on Social Media". In: *ACL*. 2022, pp. 628–635.

[218] Glenn N Saxe, Leonard Bickman, Sisi Ma, and Constantin Aliferis. "Mental health progress requires causal diagnostic nosology and scalable causal discovery". In: *Front. in Psychiatry* 13 (2022), p. 898789.

[219] Chandni Saxena, Muskan Garg, and Gunjan Ansari. "Explainable Causal Analysis of Mental Health on Social Media Data". In: *ICONIP*. 2023, pp. 172–183.

[220] Annika M Schoene, George Lacey, Alexander P Turner, and Nina Dethlefs. "Dilated LSTM with attention for Classification of Suicide Notes". In: *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*. Ed. by Eben Holderness, Antonio Jimeno Yepes, Alberto Lavelli, Anne-Lyse Minard, James Pustejovsky, and Fabio Rinaldi. Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 136–145. DOI: 10.18653/v1/D19-6217. URL: https://aclanthology.org/D19-6217/.

[221] Annika Marie Schoene, Alexander P. Turner, Geeth De Mel, and Nina Dethlefs. "Hierarchical Multiscale Recurrent Neural Networks for Detecting Suicide Notes". In: *IEEE Transactions on Affective Computing* 14.1 (2023), pp. 153–164.

[222] Viktor Schønning, Gunnhild Johnsen Hjetland, Leif Edvard Aarø, and Jens Christoffer Skogen. "Social media use and mental health and well-being among adolescents–a scoping review". In: *Frontiers in psychology* 11 (2020), p. 542107.

[223] Ivan Sekulic, Matej Gjurković, and Jan Šnajder. "Not Just Depressed: Bipolar Disorder Prediction on Reddit". In: *WASSA*. ACL, 2018, pp. 72–78.

[224] Richard I Shader. "COVID-19 and depression". In: *Clinical therapeutics* 42.6 (2020), pp. 962–963.

[225] Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, Wenwu Zhu, et al. "Depression detection via harvesting social media: A multimodal dictionary learning solution." In: *IJCAI*. 2017, pp. 3838–3844.

[226] Jiaming Shen, Jialu Liu, Dan Finnie, Negar Rahmati, Mike Bendersky, and Marc Najork. ""Why is This Misleading?": Detecting News Headline Hallucinations with Explanations". In: *ACM Web Conference*. 2023, pp. 1662–1672.

[227] Benjamin Shickel and Parisa Rashidi. "Automatic Triage of Mental Health Forum Posts". In: *CLPsych*. 2016, pp. 188–192.

[228] Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. "Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings". In: *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. Ed. by Kate Loveys, Kate Niederhoffer, Emily Prud'hommeaux, Rebecca Resnik, and Philip Resnik. New Orleans, LA: Association for Computational Linguistics, June 2018, pp. 25–36. DOI: 10.18653/v1/W18-0603. URL: https://aclanthology.org/W18-0603/.

[229] Asmit Kumar Singh, Udit Arora, Somyadeep Shrivastava, Aryaveer Singh, Rajiv Ratn Shah, Ponnurangam Kumaraguru, et al. "Twitter-stmhd: An extensive user-level database of multiple mental health disorders". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 16. 2022, pp. 1182–1191.

[230] Pradyumna Prakhar Sinha, Rohan Mishra, Ramit Sawhney, Debanjan Mahata, Rajiv Ratn Shah, and Huan Liu. "# suicidal - A Multipronged Approach to Identify and Explore Suicidal Ideation in Twitter". In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. CIKM '19. Beijing, China: Association for Computing Machinery, 2019, pp. 941–950. ISBN: 9781450369763. DOI: 10.1145/3357384.3358060. URL: https://doi.org/10.1145/3357384.3358060.

[231] Ruba Skaik and Diana Inkpen. "Using Twitter Social Media for Depression Detection in the Canadian Population". In: *AICCC*. 2021, pp. 109–114.

[232] Hoyun Song, Jinseon You, Jin-Woo Chung, and Jong C Park. "Feature attention network: interpretable depression detection from social media". In: *Proceedings of the 32nd Pacific Asia conference on language, information and computation*. 2018.

[233] Robert L. Spitzer, Kurt Kroenke, Janet B. W. Williams, and Bernd Löwe. "A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7". In: *Archives of Internal Med.* 166.10 (2006), pp. 1092–1097.

[234] Robert A. Steer and Aaron T. Beck. "Beck Anxiety Inventory". In: *Evaluating stress: A book of resources*. Scarecrow Education, 1997, pp. 23–40.

[235] Isabel Straw and Chris Callison-Burch. "AI in mental health and the biases of language based models". In: *PLOS ONE* 15.12 (2020), pp. 1–19.

[236] Yongduo Sui, Xiang Wang, Jiancan Wu, Min Lin, Xiangnan He, and Tat-Seng Chua. "Causal Attention for Interpretable and Generalizable Graph Classification". In: KDD '22. 2022, pp. 1696–1705.

[237] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks". In: *International conference on machine learning*. PMLR. 2017, pp. 3319–3328.

[238] Michael Mesfin Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. "Detection of Suicide Ideation in Social Media Forums Using Deep Learning". In: *Algorithms* 13.1 (Dec. 2019), p. 7.

[239] Qwen Team. *Qwen2.5: A Party of Foundation Models*. Sept. 2024. URL: https://qwenlm.github.io/blog/qwen2.5/.

[240] Vankayala Tejaswini, Korra Sathya Babu, and Bibhudatta Sahoo. "Depression Detection from Social Media Text Analysis Using Natural Language Processing Techniques and Hybrid Deep Learning Model". In: *ACM TALLIP* (2022).

[241] Alan R Teo, HwaJung Choi, and Marcia Valenstein. "Social relationships and depression: ten-year follow-up from a nationally representative study". In: *PloS one* 8.4 (2013), e62396.

[242] Anja Thieme, Danielle Belgrave, and Gavin Doherty. "Machine Learning in Mental Health: A Systematic Review of the HCI Literature to Support the Development of Effective and Implementable ML Systems". In: *ACM Transactions Comput.-Hum. Interact.* 27.5 (Aug. 2020). ISSN: 1073-0516. DOI: 10.1145/3398069. URL: https://doi.org/10.1145/3398069.

[243] Anna Thompson, Caroline Hunt, and Cathy Issakidis. "Why wait? Reasons for delay and prompts to seek help for mental health problems in an Australian clinical sample". In: *Social Psychiatry and Psychiatric Epidem.* 39.10 (2004), pp. 810–817.

[244] Graham Thornicroft, Somnath Chatterji, Sara Evans-Lacko, Michael Gruber, Nancy Sampson, Sergio Aguilar-Gaxiola, Ali Al-Hamzawi, Jordi Alonso, Laura Andrade, Guilherme Borges, et al. "Undertreatment of people with major depressive disorder in 21 countries". In: *The British Journal of Psychiatry* 210.2 (2017), pp. 119–124.

[245] I-Hsien Ting, Chia Sung Yen, Chia-Chun Kang, and Shu-Chen Yang. "An Empirical Study of Automatic Social Media Content Labeling and Classification based on BERT Neural Network". In: *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 2022, pp. 411–414. DOI: 10.1109/ASONAM55673.2022.10068630.

[246] Erhan Tiryaki, Akshay Sonawane, and Lakshman Tamil. "Real-Time CNN Based ST Depression Episode Detection Using Single-Lead ECG". In: *2021 22nd International Symposium on Quality Electronic Design (ISQED)*. Apr. 2021, pp. 566–570.

[247] ML Tlachac and Elke Rundensteiner. "Screening For Depression With Retrospectively Harvested Private Versus Public Text". In: *IEEE Journal of Biomedical and Health Informatics* 24.11 (2020), pp. 3326–3332.

[248] Sebastian Trautmann, Jürgen Rehm, and Hans-Ulrich Wittchen. "The economic costs of mental disorders: Do our societies react appropriately to the burden of mental disorders?" In: *EMBO reports* 17.9 (2016), pp. 1245–1249.

[249] Elsbeth Turcan and Kathy McKeown. "Dreaddit: A Reddit Dataset for Stress Analysis in Social Media". In: *LOUHI*. Nov. 2019, pp. 97–107.

[250] Ana Sabina Uban, Berta Chulvi, and Paolo Rosso. "Understanding Patterns of Anorexia Manifestations in Social Media Data with Deep Learning". In: *CLPsych*. 2021, pp. 224–236.

[251] Ana-Sabina Uban, Berta Chulvi, and Paolo Rosso. "An emotion and cognitive based Analysis of mental health disorders from Social media data". In: *Future Generation Computer Systems* 124 (2021), pp. 480–494.

[252] A Vaswani. "Attention is all you need". In: *Advances in Neural Information Processing Systems* (2017).

[253] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. "Graph Attention Networks". In: *International Conference on Learning Representations*. 2018.

[254] Anne M Verhallen, Remco J Renken, Jan-Bernard C Marsman, and Gert J Ter Horst. "Romantic relationship breakup: An experimental model to study effects of stress on depression (-like) symptoms". In: *PloS one* 14.5 (2019), e0217320.

[255] Jesse Vig. "Visualizing Attention in Transformer-Based Language Representation Models". In: *arXiv e-prints* (2019), arXiv–1904.

[256] Erin A Vogel, Jason P Rose, Lindsay R Roberts, and Katheryn Eckles. "Social comparison, social media, and self-esteem." In: *Psychology of popular media culture* 3.4 (2014), p. 206.

[257] Michael Von Korff and Gregory Simon. "The relationship between pain and depression". In: *The British Journal of psychiatry* 168.S30 (1996), pp. 101–108.

[258] Ning Wang, Luo Fan, Yuvraj Shivtare, Varsha Badal, Koduvayur Subbalakshmi, Rajarathnam Chandramouli, and Ellen Lee. "Learning Models for Suicide Prediction from Social Media Posts". In: *CLPsych*. 2021, pp. 87–92.

[259] Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. "Resolving knowledge conflicts in large language models". In: *arXiv preprint arXiv:2310.00935* (2023).

[260] Madeline E. White and Lata Satyen. "Cross-cultural differences in intimate partner violence and depression: A systematic review". In: *Aggression and Violent Behavior* 24 (2015), pp. 120–130.

[261] WHO. *Depression*. Ed. by World Health Organization. [Online; accessed 20-Jan-2024]. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/depression.

[262] Priya J Wickramaratne, Tenzin Yangchen, Lauren Lepow, Braja G Patra, Benjamin Glicksburg, Ardesheer Talati, Prakash Adekkanattu, Euijung Ryu, Joanna M Biernacka, Alexander Charney, et al. "Social connectedness as a determinant of mental health: A scoping review". In: *PloS one* 17.10 (2022), e0275004.

[263] Sarah Wiegreffe and Yuval Pinter. "Attention is not not Explanation". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 11–20. DOI: 10.18653/v1/D19-1002. URL: https://aclanthology.org/D19-1002/.

[264] Jiageng Wu, Xian Wu, Yining Hua, Shixu Lin, Yefeng Zheng, and Jie Yang. "Exploring Social Media for Early Detection of Depression in COVID-19 Patients". In: *ACM Web Conference*. 2023, pp. 3968–3977.

[265] Min Wu, Haoze Wu, and Clark Barrett. "VeriX: towards verified explainability of deep neural networks". In: *Advances in neural information processing systems* 36 (2023), pp. 22247–22268.

[266] Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. "Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts". In: *The Twelfth International Conference on Learning Representations*. 2023.

[267] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. "How powerful are graph neural networks?" In: *arXiv preprint arXiv:1810.00826* (2018).

[268] Xinyuan Xu. "Detecting Suicide Ideation in the Online Environment: A Survey of Methods and Challenges". In: *IEEE Transactions on Computational Social Systems* 9.3 (2022), pp. 679–687.

[269] Yuanyuan Xue, Qi Li, Li Jin, Ling Feng, David A Clifton, and Gari D Clifford. "Detecting Adolescent Psychological Pressures from Micro-Blog". In: *Health Information Science*. Springer International Publishing, 2014, pp. 83–94.

[270] Hao Yan, Ellen E Fitzsimmons-Craft, Micah Goodman, Melissa Krauss, Sanmay Das, and Patricia Cavazos-Rehg. "Automatic Detection of Eating Disorder-Related Social Media Posts That Could Benefit From a Mental Health Intervention". In: *Internationa Journal of Eaing Disorders* 52.10 (2019), pp. 1150–1156.

[271] Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. "Towards interpretable mental health analysis with large language models". In: *EMNLP*. 2023, pp. 6056–6077.

[272] Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. "MentaLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models". In: *Proceedings of the ACM Web Conference 2024*. WWW '24. Singapore, Singapore: Association for Computing Machinery, 2024, pp. 4489–4500. ISBN: 9798400701719. DOI: 10.1145/3589334.3648137. URL: https://doi.org/10.1145/3589334.3648137.

[273] Zhilin Yang. "XLNet: Generalized Autoregressive Pretraining for Language Understanding". In: *arXiv preprint arXiv:1906.08237* (2019).

[274] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. "Hierarchical Attention Networks for Document Classification". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kevin Knight, Ani Nenkova, and Owen Rambow. San Diego, California: Association for Computational Linguistics, June 2016, pp. 1480–

1489. DOI: 10.18653/v1/N16-1174. URL: https://aclanthology.org/N16-1174/.

[275] Liang Yao, Chengsheng Mao, and Yuan Luo. "Graph convolutional networks for text classification". In: *AAAI*. Vol. 33. 01. 2019, pp. 7370–7377.

[276] Andrew Yates, Arman Cohan, and Nazli Goharian. "Depression and Self-Harm Risk Assessment in Online Forums". In: *EMNLP*. 2017, pp. 2968–2978.

[277] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. "Do large language models know what they don't know?" In: *arXiv preprint arXiv:2305.18153* (2023).

[278] Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. "GNNExplainer: generating explanations for graph neural networks". In: *NeurIPS*. 2019, pp. 9244–9255.

[279] H. Yuan, H. Yu, S. Gui, and S. Ji. "Explainability in Graph Neural Networks: A Taxonomic Survey". In: *IEEE Transactions PAMI* 45.05 (2023), pp. 5782–5799.

[280] Yunhao Yuan, Koustuv Saha, Barbara Keller, Erkki Tapio Isometsä, and Talayeh Aledavood. "Mental Health Coping Stories on Social Media: A Causal-Inference Study of Papageno Effect". In: *ACM Web Conference*. 2023, pp. 2677–2685.

[281] Bowen Zhang and Harold Soh. "Extract, define, canonicalize: An llm-based framework for knowledge graph construction". In: *arXiv preprint arXiv:2404.03868* (2024).

[282] Tianlin Zhang, Annika M. Schoene, Shaoxiong Ji, and Sophia Ananiadou. "Natural language Processing applied to mental illness detection: a narrative review". In: *npj Digital Medicine* 5.1 (2022), p. 46.

[283] Tianlin Zhang, Kailai Yang, Hassan Alhuzali, Boyang Liu, and Sophia Ananiadou. "PHQ-aware depressive symptoms identification with similarity contrastive learning on social media". In: *Information Processing & Management* 60.5 (2023), p. 103417.

[284] Tianlin Zhang, Kailai Yang, and Sophia Ananiadou. "Sentiment-guided transformer with severity-aware contrastive learning for depression detection on social media". In: *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*. 2023, pp. 114–126.

[285] Tianlin Zhang, Kailai Yang, Shaoxiong Ji, and Sophia Ananiadou. "Emotion fusion for mental illness detection from Social media: A survey". In: *Information Fusion* 92 (2023), pp. 231–246.

[286] Zhenwen Zhang, Zepeng Li, Jianghong Zhu, Zhihua Guo, Bin Shi, and Bin Hu. "Enhancing user sequence representation with cross-view collaborative learning for depression detection on Sina Weibo". In: *Knowledge-Based Systems* 293 (2024), p. 111650. DOI: https://doi.org/10.1016/j.knosys.2024.111650.

[287] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. "A Robustly Optimized BERT Pre-training Approach with Post-training". eng. In: *Proceedings of the 20th Chinese National Conference on Computational Linguistics*. Ed. by Sheng Li, Maosong Sun, Yang Liu, Hua Wu, Kang Liu, Wanxiang Che, Shizhu He, and Gaoqi Rao. Huhhot, China: Chinese Information Processing Society of China, Aug. 2021, pp. 1218–1227. URL: https://aclanthology.org/2021.ccl-1.108/.

[288] Frederick J Zimmerman and Wayne Katon. "Socioeconomic status, depression disparities, and financial strain: what lies behind the income-depression relationship?" In: *Health economics* 14.12 (2005), pp. 1197–1215.

[289] Sidney Zisook and Richard A DeVaul. "Grief, unresolved grief, and depression". In: *Psychosomatics* 24.3 (1983), pp. 247–256.

[290] Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. "Depressionnet: learning multi-modalities with user post summarization for depression detection on social media". In: *proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 2021, pp. 133–142.

[291] Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. "Hierarchical convolutional attention network for depression detection on social media and its impact during pandemic". In: *IEEE Journal of Biomedical and Health Informatics* 28.4 (2023), pp. 1815–1823.

[292] Hamad Zogan, Imran Razzak, Xianzhi Wang, Shoaib Jameel, and Guandong Xu. "Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media". In: *World Wide Web* 25.1 (2022), pp. 281–304.

[293] Maria Li Zou, Mandy Xiaoyang Li, and Vincent Cho. "Depression and disclosure behavior via social media: A study of university students in China". In: *Heliyon* 6.2 (2020).