Multimodal NLP in Mental Healthcare

RINA CARINES MANUMBALI CABRAL

Doctor of Philosophy

THE UNIVERSITY OF SYDNEY

Supervisor: Dr. Caren Han, Dr. Josiah Poon

A thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy

School of Computer Science
Faculty of Engineering
The University of Sydney
Australia

21 April 2026

Statement of Originality

This is to certify that the content of this thesis is my own work. This thesis has not been submitted for any other degree or purpose.

I certify that the intellectual content of this thesis is the product of my own work, and that all assistance received in preparing this thesis and all sources have been acknowledged.

Rina Carines Manumbali Cabral

Acknowledgements

I would like to express my deepest gratitude to Dr Caren Han, without whom my research journey might not even have started. Thank you for believing in me and for helping me realise that research is a path I can take. Thank you for your incredible dedication and guidance in every step of my PhD journey and for sharing your infectious passion for the field. I have learned so much from working alongside you and look forward to learning even more.

I would also like to thank Dr Josiah Poon for the constant support and guidance. Thank you for the opportunities to participate in collaborations with international universities and for the additional support that enabled me to attend various conferences in person.

To my ADNLP/IAMNLP colleagues, thank you for the camaraderie, collaborations, discussions, and check-ins. Thank you to Dr Siwen Luo, Dr Yihao Ding, Dr Feiqi Cao, Dr Jie Yang, Dr Sharon Long, Dr Henry Weld, Dr Eileen Wang, Yan Li, and Zhihao Zhang for their support when I started this journey. And to our newer members, Shuo Yang, Shan Ng, Mohammad Elfauri, Reza Madani, Haoran Zhao, Muku Akasaka, Nic Liang, Sehyuk Park, Jinwoo Kim, Dillon Blake, Dhita Pratama, Biao Xiang, and Zhenyuan He. Special thanks to David Chung for being a great food/cafe recommender and for being the best premium Uber service in Melbourne. Also to Dr Jean Lee for great conversations and advice over coffee runs.

Finally, to my small but steadfast support system. Thank you to my family. Without their support, this journey away from home would never have found a starting point. Thank you for always keeping me grounded as I navigate this chapter and figure out my own paths to take. To Kat and Leni, thank you for being the best constants in any journey I find myself in. Thank you for the words of encouragement and for keeping me sane while putting up with my random anxiety outbursts from miles and continents away. And lastly, to Fudgee, Tim Tam, and Tootsie, for never failing to scold me every year after I disappear for months.

This research was made possible through the support received from the University of Sydney International Tuition Fee Scholarship and International Stipend Scholarship. I also received scholarship top ups from the Google AIR Scholarship.

Author Attribution Statement

The publications in this thesis are summarised as follows:

(1) **Chapter 3** of this thesis is published as [33] in the Robotics journal.
I am a first author of this paper. I formulated the research aim and co-designed the methodology. I was primarily responsible for collecting and analysing datasets, conducting all of the the experiments, analysing the results, and writing most of the manuscript.

(2) **Chapter 4** of this thesis is published as [32] in CIKM 2024.
I am the first author of this paper. I formulated the research aim, analysed datasets, established the methodology, implemented models, ran different experiments, and analysed results. I lead the writing of the manuscript and co-wrote most parts of the paper.

(3) **Chapter 5** of this thesis is published as [34] in WWW 2025.
I am a first author of this paper. I co-developed the research aim and co-designed the methodology. I conducted all of the experiments, analysed the results, and wrote most of the paper.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Name: Rina Carines Manumbali Cabral
Signature:
Date: 29 December 2025

As a supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Name: Dr. Caren Han
Signature:
Date: 29 December 2025

Name: Dr. Josiah Poon
Signature:
Date: 30 December 2025

Generative AI Attribution Statement

During the preparation of this thesis, the following generative AI models have been used as part of the research design:

(1) Bark, a text-to-speech generative model, was used as part of the methodology in Chapter 4. Details of its use are outlined in Section 4.4.2. The full code used in the study is published for reproducibility.

(2) Gemini 1.5-flash and GPT-4o were evaluated for comparison to the proposed methodology in Chapter 5. Prompt templates and full samples are provided in Section 5.6.7 for reproducibility.

Where relevant, in-text citations and labels have been included for any section of text that was generated by the generative AI models. The generative AI models were not used to generate content or enhance text in this manuscript. I take full responsibility for this submitted thesis, and confirm that the work is my own. Any use of generative AI is done in accordance with University guidelines and policies.

Publication List

Cabral, R. C., Han, S. C., Poon, J., & Nenadic, G. (2024). “MM-EMOG: Multi-Label Emotion Graph Representation for Mental Health Classification on Social Media." Robotics, 13(3), 53. **Journal Robotics**

Cabral, R. C., Luo, S., Poon, J., & Han, S. C. (2024). “3M-Health: Multimodal Multi-Teacher Knowledge Distillation for Mental Health Detection." In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, Boise, ID, USA. **CIKM 2024 (Rank A Conference)**

Cabral, R. C., Han, S. C., Alhassan, A., Batista-Navarro, R., Nenadic, G., & Poon, J. (2025). “TriG-NER: Triplet-Grid Framework for Discontinuous Named Entity Recognition." In Proceedings of the ACM on Web Conference 2025, Sydney NSW, Australia. **WWW 2025 (Rank A* Conference)**

Alhassan, A., Schlegel, V., **Cabral, R. C.**, Batista-Navarro, R., Han, S. C., Poon, J., & Nenadic, G. (2025). “DocDiscNER: Enhanced document-level discontinuous NER via coordination ellipses resolution and self-consistency decoding." In Frontiers in Artificial Intelligence and Applications. IOS Press. **ECIR (Rank A Conference)**

Abstract

Mental health has been a growing concern for countries, communities, and individuals in the past decade. Despite advances in societal norms and mental health resources, mental healthcare systems still face significant challenges. Researchers have explored technological advances in deep learning and natural language processing in response, especially with the abundance of online data and the emerging, yet still limited, movement towards accessible healthcare data for research. However, recent research trends have shifted toward incorporating various media-based modalities, including videos, images, and physiological data. This shift, while promising, presents a different set of limitations, particularly in terms of data accessibility and research reproducibility.

This thesis addresses these challenges by leveraging the ubiquity of textual data in mental health-related settings, aiming to exhaust different text-derived complementary information at different abstraction levels to enrich textual representations beyond standard semantic contextualisation. The main contributions of this thesis are threefold, proposing three abstraction-level modalities and three different approaches to multimodal integration for improving mental health risk detection and information extraction.

First, inspired by the complexity of human emotions and language, affective information from the emotion modality is thoroughly integrated into contextual representations through multi-emotion graph pretraining for depression and suicide risk detection. Capturing complex heterogeneous emotions through global and local graph-based learning consistently outperformed unimodal baseline weighted F1-scores by 7-21% overall and 5-49% for the most concerning classes, showing up to 41% and 51% improvement, respectively.

Extending this, the second study introduces the acoustic modality, capturing prosodic information derived from textual data. It proposes a multi-teacher knowledge distillation framework to integrate emotion, audio, and textual abstractions for the same mental health

tasks. The modality-preserving distillation increased baseline scores by 3-8% overall, generating up to 14% and 19% improvement for the whole dataset and for the most concerning classes, respectively.

Finally, the word-pair modality is explored, proposing a novel perspective for relational-structural abstraction from raw textual input. It is integrated through a triplet-grid framework leveraging triplet loss, which effectively pulls together co-occurring word pairs and distances dissimilar pairs, improving word-boundary detection for the extraction of disjointed adverse drug reactions. The proposed framework demonstrated up to 5% improvement over baselines for the extraction of discontinuous entities, thereby facilitating a deeper understanding of substance use disorders.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Content warning: This thesis contains discussions surrounding mental health and provides examples of texts indicative of depression and suicide ideation, which may be triggering for some individuals.

1.1 Background

Mental health, being the state of mind of an individual, is a salient factor that contributes to the holistic well-being of both the individual and the society they belong to. Good mental wellness can translate to good physical and social well-being, allowing a person to thrive in various aspects of their life, both as an individual, a family member, and a member of different communities. On a larger scale, a society that invests in mental healthcare reaps economic benefits from its productive and functioning members. This all-encompassing characteristic makes mental health and mental healthcare salient topics of concern across different viewpoints.

Despite significant advances in normalising mental health topics and in investments in relevant resources in recent years, however, mental health disorders are still prevalent worldwide. The World Health Organisation has determined that one in seven individuals globally experiences mental health issues [251]. In Australia, it is estimated that 43% of the adult population has experienced a mental disorder at least once in their life [17], even with improvements in resources and services in the past decade [16].

One of the reasons why mental health remains a challenge is that, by nature, most mental health issues start unseen and are overlooked until symptoms become too severe. Internalised

mental health issues such as depression, anxiety, and post-traumatic stress disorder (PTSD) are typically brushed aside since sadness, worry and stress are normal human reactions to everyday situations. Externalised disorders, on the other hand, such as eating disorders and substance use disorders, tend to go unnoticed until the effects are physically evident. This subtlety makes detection and recognition of the diseases alone challenging.

Mental healthcare systems also face different challenges. As with any other healthcare system, health practitioners are only able to help individuals who are within the system. On a personal level, individuals may avoid or refuse to interact with mental health services because of the stigma associated with having mental health issues. Financial costs are also a significant concern for many since mental health services are still considered a luxury in many places. On a system level, even with proper facilities in place, resources are typically limited. Mental healthcare facilities are often scarce, sparse, and inaccessible to individuals in more rural places. Furthermore, human capital, including therapists, doctors, and support staff, is often understaffed and overworked.

Because of the limitations and challenges faced by mental healthcare systems, researchers from both healthcare and technology domains have turned to natural language processing (NLP) to help with mental health-related tasks, which has been beneficial in many ways. First, deep learning has enabled mental health systems to process large amounts of data instantaneously, allowing healthcare workers to focus on treating and interacting with patients rather than processing and analysing large quantities of data. NLP has also enabled mental health solutions to reach people outside of clinical settings through various channels, such as social media and online forums. Lastly, with recent advances in large language models (LLMs), mental healthcare NLP tasks have expanded from disease detection and information extraction to support focused applications, such as clinician support in terms of medical history summarisation and patient support through chatbot-assisted therapies.

However, despite the opportunities brought about by NLP, there are inherent limitations of textual data and conventional language modelling methods that warrant close consideration, especially when dealing with health and mental health-related tasks.

Text, by its very nature, is unstructured, subjective, and can contain implicit information. Despite advances in language modelling and contextualising representations, comprehending subjective and implicit information, which is highly relevant to many mental health conditions, such as emotion, sentiment, and sarcasm, remains a challenge for many language models, as well as many humans. Furthermore, its structurelessness opens it up for syntactic creativity, which, while great for expressing thoughts and emotions, presents a challenge for information extraction tasks such as named entity recognition (NER).

Prevailing language modelling methods, on the other hand, face related limitations primarily due to the token-based pretraining objective of most language models, namely masked language modelling and next word/sentence prediction (Section 2.2.2). These objectives are predominantly aimed at contextual semantic learning, which produces affectively neutral contextual representations. Affective information is only inferred from explicit emotion and sentiment words. Implicit expressions are therefore overlooked or neglected. Additionally, these pretraining methods have an inherent sequential bias due to context being inferred from adjacent, neighbouring tokens.

To overcome some of the limitations of textual data and current language modelling methods, other researchers have explored the incorporation of media-based multimodal data into mental health downstream tasks. Most of these studies focus on different media, including videos, images, and audio. With the prevalence of social media and online forums, network data and activity have also been included along with conversation trees, historical posts and user metadata. Lastly, physiological data, such as electroencephalograms (EEG) and electrocardiograms (ECG), are also utilised by researchers who have access to them. However, despite the availability of these multimodal data, accessibility remains a significant hindrance to research reproducibility and progress. Mental health is a very sensitive topic that is reasonably protected by numerous ethics and data privacy laws. Institutions that are able to collect this data must follow strict procedures for distributing it, and even more so for healthcare data. Furthermore, some countries restrict data access to within the country only. These privacy and access issues, even with social media data, prompt each researcher to use the data modalities available to them.

1.2 Research Aim

To address these limitations, this thesis takes an alternative perspective of exhausting multimodal abstractions from mental health-related textual data and developing novel ways of incorporating these modality representations to mental health downstream tasks. This thesis thus defines a modality at the representation and abstraction level, rather than strictly at the level of input medium or data streams. A modality, therefore, refers to a distinct representational view that encodes complementary information at its own abstraction level and contributes to downstream reasoning. From a single piece of raw text, these derived modalities introduces different information factorisation, inductive biases, and failure points.

It is motivated by the ubiquitousness of textual data in healthcare, both in clinical settings and outside of them. For instance, every visit to a clinical practitioner or therapist is likely to be recorded in the hospital's electronic health records (EHR) system. Highly specialised medical technologies or tests, such as diagnostic scans or blood tests, are similarly reported in text form with varying degrees of technicality. Likewise, on the patient's side, textual prescriptions for medications are typically handed out. Outside the clinical system, thanks to advances in digital communication, social media and other online forums have become valuable resources for individuals seeking information or advice, as well as for sharing experiences, thoughts, and emotions. While these avenues of information are technically not clinically validated, they still offer valuable insights that are not captured or hindered by interactions within clinical systems.

Moreover, text as a raw data is more accessible and more interpretable or comprehensible than other medical data such as physiological data or radiology images. Even without highly specialised knowledge, text can be read and interpreted by any literate person. Comprehension would vary based on technicality of text and individual knowledge, but an EEG report would still be easier to interpret than the EEG data itself.

Given these reasons which make textual data an extremely valuable source of information for mental healthcare and general healthcare alike, and to address limitations of current language

TABLE 1.1. Modalities at different abstraction and representation levels. These modalities are defined by their distinct representational view, encoding complementary information at their own abstraction-level.

Chapter	Modality	Encodes	What it corrects
3	**Emotion Modality**	Affective state and intensity	Affective neutrality of language models
4	**Acoustic Modality**	Paralinguistic and prosodic biomarkers	Loss of affective tonal cues in raw textual data
5	**Word-Pair Modality**	Relational and structural dependencies	Sequential bias of token-level language models

models in this application, this thesis expands on the conventional semantic contextual representation of textual data through the extraction and incorporation of multimodal information. It explores multimodality through different abstractions of textual data, underscoring complementary information derived from the same raw text data. It proposes different integration methods that highlights each modality's unique nature. Specifically, in each chapter, this thesis aims to answer three broader research questions:

(1) What abstraction-based modalities can be derived from mental health-related texts?
(2) What methods can be used to integrate different modalities together?
(3) What mental healthcare tasks would benefit from multimodal information?

1.3 Contribution

In summary, this thesis aims to extend textual information and representation through multimodal abstraction of raw text data and proposes novel methods to derive and incorporate them in mental healthcare-related tasks.

It explores three different abstraction modalities, each with different information factorisation and representational significance for mental health downstream tasks (Table 1.1). The emotion modality represents the affective abstractions from the textual data. It integrates affective states and intensities to capture implicit feelings and sentiments that are often missed due to the affective neutrality of word representations in language models. The acoustic modality,

even if derived, represents a paralinguistic and prosodic abstraction capturing affective tonal cues absent in the raw textual data. Finally, the word-pair modality represents a relational and structural abstraction of textual data, thereby reframing the sequential bias inherent in prevalent token-based language models.

With each modality representing different information abstractions, this thesis proposes three multifaceted integration methods to preserve each modality's unique nature and cater to their specific inductive biases (Table 1.2), as opposed to simple representation concatenations which homogenises heterogeneous information. For the emotion modality, recognising the complexity of human emotions where one word may convey multiple emotions, a multi-emotion graph-based pretraining method is proposed, which captures global and local affective abstractions to create emotion-rich contextualised representations. For the acoustic modality, as a distinct medium from text, vocal biomarkers and prosodic abstractions are incorporated through multi-teacher knowledge distillation. This distillation approach trains a highly specialised audio-based teacher model that extracts paralinguistic cues from the proxy audio before distilling learned information to a student model. Lastly, the word-pair modality is incorporated through a grid-based contrastive learning approach, thereby capturing token-based relational and structural abstractions that are overlooked by traditional sequential perspectives.

Finally, these abstractive modalities and integration methods are applied to mental health downstream tasks. These tasks present challenges that typical text processing pipelines are unable to address. In particular, for detecting internalised mental health problems such as depression, anxiety, and suicidality, emotions and affective states are significant information that must be considered since mental health is heavily intertwined with emotions. This need is further highlighted when dealing with emotion-rich, user-generated texts in social media or other online platforms. The first two studies focus on this disease risk detection task, wherein the first study addresses the complexity of associating emotions with words thus applying the emotion modality with the graph-based multi-emotion pretraining. The second study explores the task further through utilising three different abstractions, namely

TABLE 1.2. Summary of modalities, integration techniques, and motivations explored by this thesis.

Chapter	Modalities	Integration	Motivation
3	Emotion, Text	Multi-label Graph Pretraining	Incorporating complex human emotions in linguistic semantics to contextualise emotion-rich texts for identifying mental health risk
4	Audio, Emotion, Text	Multi-teacher Knowledge Distillation	Simulating multimodal human understanding of different senses to incorporate derived auditory cues for enhanced emotion comprehension in mental health-related texts
5	Word-Pair Associations, Token Relationships	Triplet-Grid Framework	Expanding word-boundary detection perspectives for improved extraction of medical entities in unstructured text for understanding drug effects and monitoring possible substance abuse

the semantic text modality, the emotion modality, and the acoustic modality, brought together in a multi-teacher distillation framework.

Another significant task in mental healthcare is to understand the effects of different medications on individuals outside clinical settings, particularly for monitoring adverse drug events/reactions (ADRs) that may signify possible substance abuse, especially since antidepressants and other mental health prescription drugs are often easily abused, leading to Substance Use Disorder. Online forums have been a significant resource for individuals seeking advice regarding medications, making them a valuable tool for understanding drug effects and monitoring abuse. However, syntactic differences introduced by unique writing styles pose a challenge in extracting information from unstructured texts, where entities of medical interest, such as symptoms, diseases, or drug reactions, are often disjointed within a sequence of words. Typical sequential processing of text data falls short in determining entity boundaries of these discontinuous entities. The third study focuses on this information extraction task with a particular focus on the extraction of discontinuous adverse drug events. It proposes the use of the word-pair modality to incorporate token-level structural and relational abstractions, improving word-boundary detection through a multi-view grid-based contrastive learning approach.

The detailed contribution of each study are as follows:

(1) Multi-label Emotion Graphs
   (a) introduces MM-EMOG, a multi-label emotion graph representation for mental health classification using only social media textual posts;
   (b) proposes the use of Graph Convolutional Neural Networks [110] in a purely textual capacity for multi-label emotion representation and social media mental health classification tasks. To our knowledge, this study is one of the first to apply multi-label and graph-based textual emotion representation;
   (c) demonstrates the proposed model's high performance on three publicly available social media mental health classification datasets compared to pretrained baselines.

(2) Multimodal Multi-teacher Knowledge Distillation with Acoustic Modality
   (a) introduces 3M-Health, a novel approach to mental health classification through a multimodal and multi-task knowledge distillation model;
   (b) introduces a new acoustic modality feature derived from original textual posts, motivated by the proven effectiveness of vocal biomarkers in indicating psychological distress and other medical conditions [93];
   (c) demonstrates that the proposed multimodal approach outperforms unimodal counterparts with the choice of modalities influencing performance across divers datasets.

(3) Word-Pair Triplet-Grid Framework
   (a) introduces Trig-NER, a Triplet-Grid Framework that leverages token-based triplet loss for learning fine-grained word-pair relationships for disconinuous named entity recognition (DNER);
   (b) introduces a novel token-based triplet loss that learns fine-grained token-level representations for discontinuous entity extraction;
   (c) proposes a grid-based triplet-loss that defines word-pair similarity based on co-occurrence within the same entity, enhancing the model's ability to capture non-adjacent entity segments;

(d) performs extensive evaluations on three widely used DNER benchmark datasets, including an adverse drug reaction dataset and two disease identification datasets, and provides qualitative analysis that demonstrate the effectiveness of the proposed grid-based triplet framework over existing baselines and prompted large language models.

1.4 Thesis Overview

Chapter 2 provides an overview of the literature on mental healthcare-related deep learning, with a particular focus on different tasks and data modalities. The first subchapter focuses on general mental health-related tasks. The second subchapter focuses on text-only data, highlighting data sources, datasets, and prevalent language models used for mental health tasks. The third subchapter introduces other media-based data modalities and multimodality in mental healthcare applications. It highlights the motivations of this thesis and offers insights into future directions for combining textual modality with other data modalities to improve downstream mental healthcare tasks.

Chapter 3 is based on the work **MM-EMOG: Multi-Label Emotion Graph Representation for Mental Health Classification on Social Media** [33] published in **Robotics** journal. This study explores the modality of emotions for depression and suicidality detection to integrate affective abstractions lost in the semantic space of prevalent language models. It is inspired by the complexity and heterogeneity of human emotions and words, where a single word can convey multiple emotions. It presents a novel approach to contextualising emotions through multi-label emotion classification pretraining in a graph-based learning framework. The efficacy of this multi-emotion representation is demonstrated in a mental health classification task evaluated on three social media datasets, highlighting the importance of contextualising complex emotions.

Chapter 4 is based on the work **3M-Health: Multimodal, Multi-teacher Knowledge Distillation for Mental Health** [32] published and presented in **The 33rd ACM International Conference on Information and Knowledge Management (CIKM 2024)**. Inspired by

natural human understanding, which processes information from different senses, this study explores the acoustic modality to incorporate paralinguistic and prosodic abstractions in vocal biomarkers through text-to-speech generation. The acoustic modality is primarily considered and explored due to its capability to convey emotional cues through different auditory signals. Building on the previous study, this approach is applied to a mental health classification framework that incorporates three abstractive modalities (text, emotion, and audio) through a multi-teacher knowledge distillation framework. Textual representations learned by a text-only student model are enhanced through knowledge acquired from the multimodal teachers. Evaluated on four social media datasets, the study highlights the significance of both contextualised emotions and auditory information on improving mental health classification and further demonstrates a need for modality selection or moderation.

Chapter 5 is based on the work **TriG-NER: Triplet-Grid Framework for Discontinuous Named Entity Recognition** [34] published in **The ACM Web Conference (WWW 2025)** as a full paper. This study extends the traditional sequential perspective on unstructured text through the word-pair modality, incorporating relational and structural dependencies, using triplet loss in a grid-based perspective. It explores a grid-based application for the identification and extraction of drug-related reactions, which helps understand pharmaceutical substance abuse and identify instances of Substance Use Disorder. The grid framework leverages word-pair relationships enhanced through the application of a triplet loss that simultaneously pulls together similar word pairs and pushes apart dissimilar ones. This novel framework is applied to a discontinuous named entity recognition task where identifying correct entity boundaries is crucial. In this application, similarity is defined by the co-occurrence of words and word pairs within an entity, effectively distinguishing entity words from non-entity words, even if they occur consecutively. The grid framework is evaluated on DNER datasets, identifying adverse drug reactions from an online health forum and diseases from clinical texts, showing significant improvements over other grid-based baselines.

Chapter 6 summarises key contributions, findings, and discussions for each chapter in relation to the research questions put forth by this thesis. It finally concludes with a discussion on potential research directions for mental health-related applications.

CHAPTER 2

Literature Review

Mental health has seen an exponential growth in research interest over the past five years, driven by the advent of deep learning, natural language processing (NLP), and, more recently, large language models (LLMs). Faster computational resources enabled an interdisciplinary field to leverage various types of data within and outside clinical settings. Large quantities of clinical data, such as electronic health records (EHR), physiological data, medical imaging data, or recorded counselling sessions, which are typically archived and untouched after patient diagnosis and treatment, are now utilised to gain more insights about mental disorders. Outside clinical settings, deep learning advances have enabled the extension of this capability to user-generated data and even post-mortem investigation data, allowing for insights into mental healthcare, particularly regarding candid and uninhibited thoughts, feelings, and emotions.

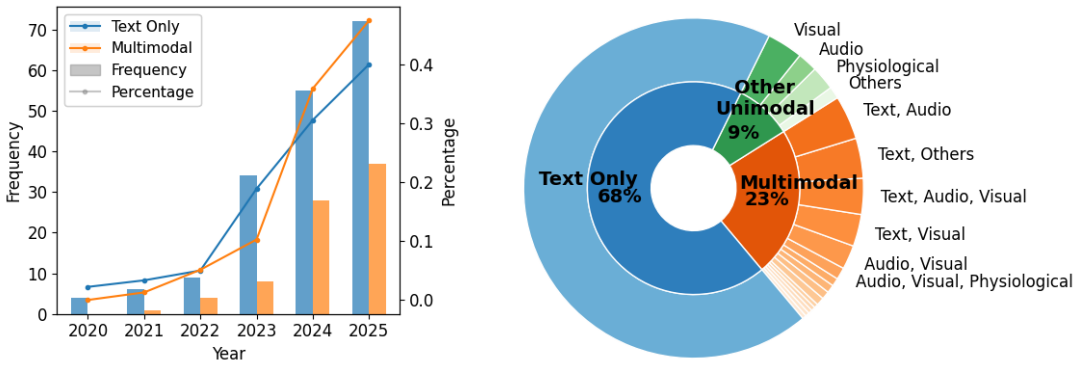


FIGURE 2.1. Medium-based modality trends in literature from 2020-2025. (left) A comparison of the number of studies using only the textual modality against studies using other modalities or a combination of different modalities. (right) Proportion of each modality and modality combinations.

At the core of this increased research attention are textual or language-based data, which largely dominate clinical and user-generated data. Even with the rapid rise of research involving multimedia or multimodal data, particularly between 2023 and 2024, textual data remains a fundamental resource for mental healthcare-related tasks in both unimodal and multimodal studies. Figure 2.1 illustrates this along with trends from other mediums of data utilised in mental health research retrieved from the last five years[1]. However, there is an undeniable increase in interest in other modalities as well, especially with the rise of multimodal large language models (MLLMs), which incorporate these together with textual modality.

This chapter summarises and explores relevant literature around mental health-related tasks employing different medium-based modalities. The first subchapter (Section 2.1) introduces various generalised tasks that encompass different single modality inputs or a combination of different modalities. The second subchapter (Section 2.2) discusses trends in unimodal mental health textual frameworks and enumerates popular text-based datasets. The third subchapter (Section 2.3) takes a broader perspective and expands from the textual modality to discuss current trends and literature in multimedia, multimodal frameworks for mental health. Detailed relevant literature for each study may be found in its corresponding chapter.

2.1 Mental Health-related Tasks

While tasks and datasets are typically interrelated to each other, this section focuses only on prevalent mental health-related tasks. Datasets will be discussed in more detail with respect to modalities in the succeeding subsections.

[1]This search includes research articles from 01 January 2020 to 31 October 2025. Detailed search methods are described in Appendix A.

2.1.1 Disease Risk Detection

An overwhelming majority (84% of the 263 studies retrieved from the search conducted for Figure 2.1) of deep learning mental health studies focus on the detection or classification of diseases, risk levels, or symptoms using any given modality of input data.

Classification Outcomes

Focusing on one particular mental health condition, some studies aim to distinguish between the presence and absence of the disease in a binary classification problem [86, 140, 182, 225, 265]. In contrast, other studies recognise different risk levels or severity, turning the problem into a multiclass classification task [84, 67, 103, 209]. While more granular, several studies have identified an important limitation in multiclass classification: risk levels are hardly exclusive from one another and must be treated as ordinal levels that increase or decrease in severity for each stage [166, 95], reformulating the task into a regression or regression-based classification.

There is also an inherent complexity when it comes to mental health in that most mental health diseases, symptoms, and factors co-occur with each other, prompting research into the interconnectedness of these aspects through the identification of multiple diseases, symptoms, or causes. For instance, acknowledging the nature of depression, Haque et al. [84] employ a binary classification to differentiate between depression and suicide related social media posts. Garg [65] further presents a multiclass classification on mental disturbance causes. Recognising inherent comorbidities of mental health diseases, Hengle et al. [86] propose multi-label classification between depression and anxiety disorders.

Classification Granularities

Another significant perspective on mental health classification is the granularity for which the classification is applied to. In clinical settings, classification tasks may be applied to individual clinical notes identifying related diseases contained within the note itself [139]. On a slightly larger view, using multiple clinical notes may be used to more accurately identify patient risk and changes of symptoms over time [61, 88, 152, 249]. In the same manner, outside clinical settings, social media posts have been a popular data source for many research

studies where classification is performed over individual posts [66, 70, 166, 182, 185] or over users as individuals [38, 47, 67, 75, 140, 199, 209].

While less common, researchers have also explored building forecasting models in a population or national level using large online data to inform health practitioners and policy makers of possible trigger events and enabling real-time large scale monitoring and timely management which traditional forecasting methods using historical statistics lack. In particular, Tuarob et al. [224], utilised social media text data and language-agnostic language models to extract different mental health signals. These signals were aggregated for input to a forecasting model evaluated against ground-truth self-harm injury and death statistics.

These different granularities often have different aims however, it is important to consider them when creating scalable mental healthcare models especially when dealing with post-level and user-level or record-level and patient-level data. Advancing from smaller granularity to larger ones is generally more straightforward and more scalable however, the opposite would be more complicated primarily because a user-level annotation will not translate to every post.

2.1.2 Information Extraction

Unlike detection/classification tasks, information extraction involves identifying and extracting specific aspects of mental health from textual data. Traditionally approached through two-step named-entity recognition (NER), the precise location of the entity within the text is first identified, then subsequently classified according to entity type. However, entities are defined differently depending on the requirements of the downstream task, creating a significant challenge for the task. Standardised entities, such as diseases or drug names, are relatively straightforward to identify and segment from the text, as the words within these entities are typically in a continuous sequence. On the other hand, symptoms and adverse drug events or reactions (ADE/ADR) are more complex entities that have less defined structures and can occur as overlapping, discontinuous, or nested (Figure 2.2). Due to these complex entities, recent NER models have proposed different tagging schemes, which significantly dictate the model's architecture, thereby limiting generalisability.

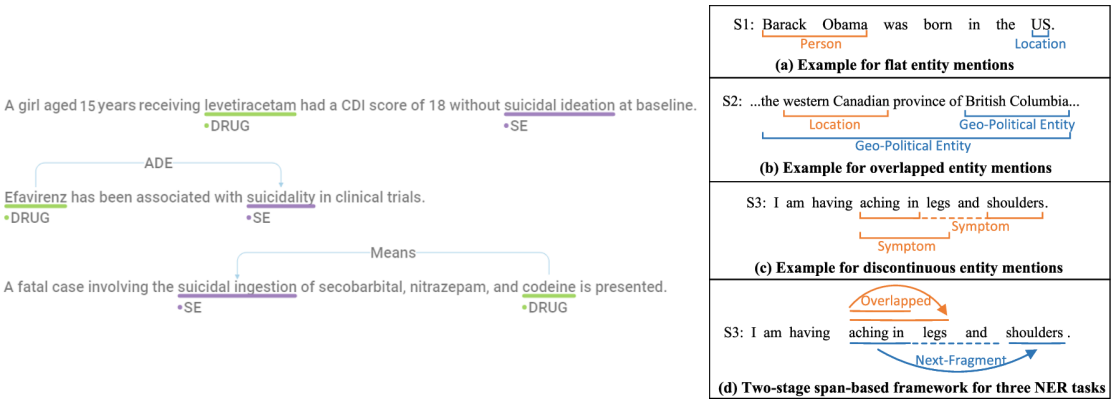


FIGURE 2.2. Examples of mental health-related entity recognition [106] (left) and different complex entities [90] (right).

Information extraction applied to mental health and healthcare aims to identify and extract various social, medical, and mental health aspects from psychiatric or psychological texts. From clinical notes and scientific texts, Researchers [137, 139, 170, 178] have extracted multiple health factors, including symptoms, disorders, medications, brands, therapies, medical departments, and more. Other researchers [241, 240] have explored social determinants of health (SDoH) or the conditions in which an individual lives with, using suicide death investigation notes, offering insights on social factors that influence suicide ideation and attempt. Karapetian et al. [106] takes information extraction further by not only identifying suicide-related phrases (e.g. "suicidal ideation", “suicide", “suicidal ingestion") and drugs but also classifying the relationship between them as either ADE, suicide means, and treatment.

2.1.3 Generative Tasks

Generative tasks surrounding mental health and mental healthcare has evolved in recent years with the rapid advancement of large language models. These tasks include explanation of symptoms or diseases, summarisation of patient profiles or therapy sessions, and LLM-assisted online therapies.

One significant benefit of recent LLMs is the incorporation of reasoning or thinking functions which can not only do risk classifications but also provide highly interpretable explanations in natural language. Li et al. [123] introduced Mental Analysis by Incorporating Mental

Scales (MAIMS) where disease risk detection and explanation is done in two steps. In the first step, using textual input from a social media post and a set of mental scale question, two LLMs imitates the poster and anaylses the post content for completion of the mental scale questions. In the second step, another set of LLMs are prompted to determine mental health issues with corresponding explanations. Singh et al. [210] on the other hand, incorporates text, audio, and video media modalities from recorded patient-doctor conversations for emotion and cognitive distortion detection and reasoning. Similar to Li et al., after prompting for emotion and cognitive distortion labels, an LLM is re-prompted for an explanation using the dialogue input and the predicted labels.

Generative LLMs have also benefited mental health therapists by generating summaries of counselling sessions, allowing them to focus more on the patient's needs and less on distracting tasks such as note-taking and recording. Adhikary et al. [6], recognising the limitations of standard summarisation approaches overlooking counselling-specific aspects, created MentalCLOUDS, a Mental Health Counselling-Component Guided Dialogue Summaries benchmarking dataset, and used it to evaluate 11 different LLMs. Srivastava et al. [212] took a different approach and proposed to improve counselling summarisation through an adaptable planning engine, PIECE, which injects domain and structural knowledge into any LLM foundation model.

The conversational capabilities of recent LLMs have also opened up mental health-specific tasks and applications. Several researchers have evaluated ChatGPT and other emerging LLMs, both proprietary and open source, in terms of their ability to function as a mental health support tool [63, 134, 165, 184, 203] or as aids in therapies and intervention [275, 260]. One particular study by Liu et al. [134] evaluates ChatGPT-generated responses and compares them with human-generated responses to depression-related questions, such as *"How about building confidence when things get tough?"* and *"Is it true that people who work and step into society don't believe in love anymore?"*. Different from standard Q&A tasks, these mental health support questions require meaningful, empathetic, and above all, safe answers. As such, researchers have evaluated LLM mental health support abilities with non-standard

metrics such as sentiment, response tone (e.g. politeness, professionalism), empathy, usability, and personalisation [134, 104, 165].

2.2 NLP in Mental Health

Mental health natural language processing studies have increased exponentially as various online platforms have enabled the sharing of personal thoughts, communication with peers and mental health systems, and even the sharing of datasets for research. This section discusses popular and emerging textual datasets, highlighting differences in data sources and annotation processes, as well as prominent text-only frameworks that have been utilised to build and evaluate deep learning models in various mental health-related tasks.

2.2.1 Text-based Datasets

One significant aspect that offer valuable insights regarding performance of NLP models using textual datasets in mental health studies is the source of the data. Knowing where the data was collected from sheds light on the nature of text which informs research and model building decisions. Table 2.1 enumerates publicly accessible popular and emerging mental health-related datasets highlighting data source, the task the dataset is typically used for, annotation method, annotation level, and availability as of the writing of this thesis. Data sources may be divided into five categories: social media platforms, dedicated health platforms, clinical interviews, electronic health records (EHR), and others.

Social Media

Social media (SM) platforms have dominated published datasets for text-based mental health studies in recent years primarily due to the large availability of user-generated content. Compared to the other categories, social media-based data offers the most candid expression of thoughts and feelings of an individual at the time a message is posted. Moreover, real-time posting enables time-based and longitudinal studies such as early risk detection [140], changes in moods or symptoms [91, 213], and disease relapse [7]. However, language in social media platforms are typically informal and are prone to syntactic and grammatical

TABLE 2.1. Prominent and Emerging Textual Mental Health Datasets

Dataset	Mental Health	Outcome	Task[a]	Source[b]	Platform	Annotation	Level	Avail.
CLPsych2015 [47]	Depression, PTSD	Disease	CLS	SM	Twitter	Human	User	DUA
eRisk2018 [140]	Depression, Anorexia	Early Det.	CLS	SM	Reddit	Auto	User	DUA
IdenDep [182]	Depression	Disease	CLS	SM	Reddit	Auto	Post	Public
UMD [209]	Suicide	Disease	CLS	SM	Reddit	Human	User	DUA
Dreaddit [225]	Stress	Disease	CLS	SM	Reddit	Hybrid	Post[d]	Public
R-CSSRS [67]	Suicide	Disease	CLS	SM	Reddit	Human	User	Public
SDCNL [84]	Depression, Suicide	Disease	CLS	SM	Reddit	Auto	Post	Public
DepSeverity [166]	Depression	Disease	CLS	SM	Reddit	Human	Post	Public
SWMH [99]	General MH	Disease	CLS	SM	Reddit	Auto	Post	Public
CAMS [66]	Depression, Suicide	Causes	CLS, EXP	SM	Reddit	Human	Post	Public
DEPTWEET [103]	Depression	Disease	CLS	SM	Twitter	Human	Post	Public[e]
ANGST [86]	Depression, Anxiety	Disease	CLS	SM	Reddit	Hybrid	Post	DUA
RMHD [185]	General MH	Causes	CLS	SM	Reddit	Human	Post	Public
ReDepress [7]	Depression	Relapse	CLS	SM	Reddit	Hybrid	User, Post	DUA
SWDD [36]	Depression	Disease	CLS	SM	Weibo	Human	User	Authors
KoMOS [105]	Depression, Anxiety, Sleep, Eat	Disease, Symptoms	CLS	SM	Naver	Hybrid	Conv.	Public
DepreSym [180]	Depression	Symptoms	CLS	SM	Reddit	Human	User	Authors
CounselChat [25]	General MH	Topic	CLS	HP	counselchat	Auto	Post	Public
EMS-Counsel [70]	General MH	Symptoms	CLS	HP	counselchat	Human	Post	Public
DAIC-WOZ[f] [75] E-DAIC[f] [54]	Depression, Anxiety, PTSD	Disease	CLS	IN	-	-	Patient	DUA
Vance et al. [229]	Suicide	Symptoms	CLS	EHR	NeuroBlu	Human	Sentence	Authors
DTD [139]	Depression	Clinical Features	IE	EHR	-	Auto	Note	Public
DSR [106]	Suicide	Relation	IE	SCI	PubMed	Hybrid	Sentence	Public
IMHI [262]	Multiple[g]	Explanation	EXP	SM, SMS	Reddit Twitter	Auto	Post	Public

[a] CLS: Classification, IE: Information Extraction, EXP: Generative Expalantion
[b] SM: Social Media, HP: Health Platforms, IN: Clinical Interviews, SMS: Text Messaging, SCI: Scientific Literature, EHR: Electronic Health Records, SCI: Scientific Papers
[c] Availability: Public - Data is unrestricted in an online repository; DUA - Requires signing a Data Use Agreement; Authors - Data may be requested from corresponding authors.
[d] Dreaddit samples are post segments instead of full posts.
[e] Only Tweet IDs are publicly available. Full dataset may be requested from authors.
[f] Multimodal datasets which audio transcripts are widely used.
[g] IMHI compiles 10 publicly available datasets involving different mental health conditions.

errors which contributes to the complexity to any downstream task. Social media language further incorporates emojis and text-based emoticons (eg. o_O) which, while contributing to expression of emotions, adds additional challenges in processing and contextualisation.

Different social media platforms exhibit different characteristics as well. **Reddit**, as a platform that prides itself in preserving the anonymity of users, is the most utilised platform in a significant number of studies since privacy concerns restrict it less. With anonymity, the target audience of a post on Reddit is often strangers, which makes many users more forthcoming with their thoughts. As an online forum, Reddit has subreddits or smaller dedicated communities, which most researchers [7, 86, 140, 182, 209, 225] have taken advantage of. Subreddits such as *r/Depression* and *r/SuicideWatch* not only serve as a data collection point but also as a weak labelling method where samples are labelled primarily based on which subreddit they belong to [99, 166, 182]. **Twitter** (now X) is another widely used platform for mental health-related datasets. More of a microblogging platform, texts posted on Twitter are significantly shorter, with only a maximum of 280 characters. The short character limit prompts users to use shortened words or connect multiple posts to convey their message. Without subspaces like subreddits in Reddit, data collection and automatic annotation in Twitter is slightly more complex. Hashtags (e.g., `#depressed`) are a common component on Twitter and serve as indicators of topics. Researchers have used these hashtags, along with keyword or keyphrase matching, to collect and, in some cases, automatically label mental health-related posts [138, 103, 94, 171]. Non-English-centric platforms have also been explored by other researchers focusing on language-specific or multilingual frameworks. For Chinese-language focused studies, Weibo [36, 270, 278] and Zhihu [77], a microblogging and Q&A platform, respectively, are more prominently used. Korean-based studies, in turn, have used the Naver Knowledge iN [105, 176], a Q&A platform, to build mental health deep learning systems.

Public availability of social media datasets has sparked numerous discussion on data privacy and ethics. Despite being a popular data source, a lot of research exploring mental health in social media are unable to share their datasets fully or at all. Bucur et al. [29] has even identified initially popular datasets which are no longer available. Some studies only provide pre-extracted features. Others require API access and reasonable effort to repopulate which doesn't always result in the same dataset as originally published due to post deletion or account restrictions making fair evaluation among different models difficult. Most of recent

studies have stricter access rules requiring data use agreements and ethics board approvals which, depending on the approving body, can take considerable time and effort.

Dedicated Health Platforms

Dedicated health platforms, including peer-to-expert (or trained peer) support platforms, and crisis helplines, both online and offline, are another source of data for NLP mental health studies. Unlike social media data, the type of content on these platforms generally poses a greater sense of urgency, as help seekers may be in heightened emotional and extreme mental states. Furthermore, these platforms are firmly committed to preserving the privacy of help seekers, making publicly available datasets very rare. Research involving these dedicated platforms is subject to extensive ethics reviews, and access to the datasets is typically gained through partnerships with the platforms themselves. Nonetheless, to aid in advancing digital mental health technologies, some platforms have made their data exclusively accessible to some researchers through close collaborations.

One publicly available dataset from an online crisis helpline platform is CounselChat [25] from the counselchat.com platform. The founders of the platform graciously provided the Q&A-type data, which spans 31 topics and involves 307 expert contributors ranging from psychologists, social workers, and mental health counsellors. The top 5 topics with the most questions posted to are depression, relationships, intimacy, anxiety, and family conflict. Gollapalli et al. [70] extended this dataset by re-annotating a subset for early maladaptive schemas (negative perceptions of an individual) determined by recruited experts.

Other studies have partnered with online support platforms such as Talklife.com [206] and Reachout.com [89, 155]. Some researchers utilised local or national text or call helplines such as On The Line (Australia) [93], Shout (United Kingdom) [136], Cerebral (United States of America) [214], Taiwan Lifeline International [245], and OpenUp (Hong Kong) [258]. Salmi et al. [193] and Thomas et al. [222] studied helplines in the Netherlands and Germany, respectively. While still offering valuable insights and effective models, most of these studies are only evaluated using their own dataset, limiting the generalisability of the developed frameworks. Furthermore, the reproduction of research results and the comparison of performances among different frameworks is very limited.

Clinical Interviews

Another prominent data source for mental health-related tasks is clinical interviews, including counselling sessions. The previous two data sources involve users or help-seekers initiating a post or conversation. Clinical interviews, on the other hand, are prompted by a mental health professional, eliciting a more focused response. The depth of relevant information on patient responses is greater since an expert is guiding the conversation in real-time with the aim of helping a patient open up and understand their mental state. Experts, such as psychiatrists, are extensively trained to identify symptoms and diagnose mental health conditions through observation, and the use of quantifiable diagnostic tools, including the DSM-5 [188] and PHQ-8 [111] questionnaires. Because of these, clinical interviews are considered a gold standard in-terms of clinical validity for mental health risk detection tasks [43]. However, some crucial downsides of clinical interviews include the high costs and the extensive time required to set up and conduct the interviews due to the involvement of mental health experts and the need for recruiting interviewees. Therapies and counselling sessions also tend to be services that not all individuals can afford or even seek.

DAIC-WOZ [75] is a widely used clinical interview-based dataset for depression risk detection collected through semi-structured Wizard-of-OZ (WOZ) type interviews, where patients converse with a digital avatar of the interviewer. Despite being a multimodal dataset comprising audio and video modalities, researchers have utilised interview transcripts to develop different text-based classification models [19, 115, 129, 141, 221, 272]. E-DAIC [191], an extension of the DAIC-WOZ dataset adding interviews conducted by a fully automated interviewer, is another interview-based dataset explored by researchers [3, 83, 192]. Other studies use transcripts from interviews conducted with recruited cohorts [53, 172, 246, 197]. These transcripts, however, are not publicly shared due to strict privacy and ethical laws suffering similar limitations of studies using non-public data.

Electronic Health Records

When it comes to clinical validity for any medical or healthcare-related tasks, electronic health records (EHRs) are the most credible gold standard, involving expert-written and validated clinical notes and reports. However, standard EHR data are highly structured with

specialised language and contain very objective information from the third-person point of view of the expert. While these experts are trained to identify, understand, and record an individual's mental health state, a third-person point of view, even with the help of diagnostic questionnaires, may still not comprehensively capture another person's internalised thoughts and feelings, including specific causes of emotions. This outsider perspective limits EHR-based risk detection to history [5], symptom-based [41, 145, 150] features, and analysis of comorbidities [88, 61, 233], rather than considering thoughts and emotions.

Risk detection aside, EHR data are very valuable for other tasks such as information extraction and treatment retention. Lorge et al. [139] utilises EHR records to train a model for extracting 40 mental health-related factors involving patient, illness, and treatment domains. Haredasht et al. [167] predict treatment retention versus attrition in opioid use disorder EHRs.

Similar to the previous data source types, EHR-based disease risk detection studies typically use their own datasets, obtained by partnering with hospitals and other institutions, which limits their capability to share the datasets with other researchers [5, 49, 145, 150, 250]. Other studies [41, 167, 139, 229, 268] utilise subsets of more accessible and larger datasets and data repositories that make de-identified EHR data publicly available, such as the MIMIC datasets [101, 102] and the Neuroblu platform [177]. The MIMIC datasets are available to researchers on the Physionet[2] platform after registration, completion of online training, and signing a Data Use Agreement (DUA). The Neuroblu platform is accessible through a paid subscription. Despite using accessible platforms, studies that utilise these data providers still rarely publish the specific subsets and annotations used in their research, due to restrictions on publishing derived datasets.

Recognising the lack of publicly available, annotated EHR data specifically for information extraction for difficult-to-detect depression, Lorge et al. [139] used ChatGPT [28] with extensive few-shot prompt engineering to create a synthetic dataset[3] of psychiatric clinical notes. However, the authors note that, despite careful prompting, ChatGPT still produced

[2] https://physionet.org/
[3] https://github.com/isabellelorge/dtd

errors. They further encourage limiting the use of the synthetic dataset as a silver standard rather than a gold standard.

Other Sources

Other researchers have also explored other data sources. Taking a posthumous perspective, Wang et al. [241] used death investigation narratives of suicide incidents to identify SDoH-related circumstances that affect an individual's mental health. Zhou et al. [280] explored narratives from coroners and medical examiners with LLMs to investigate female firearm suicides. On the other hand, Karapetian et al [106] annotated research articles published in the PubMed archive for extracting relationships between drugs and suicide.

2.2.2 Text-based Frameworks

With deep learning advancements, textual representations and methods have advanced from traditional frequency-based methods such as One-Hot-Encoding (OHE), Bag-Of-Words (BOW), and Term Frequency-Inverse Document Frequency (TF-IDF) to deep contextualisations including Word2Vec [154], GloVe [179], BERT [56], and large language models. This section provides an overview of pretrained language models (PLMs) and large language models (LLMs) used in different mental health studies. Table 2.2 enumerates widely used PLMs and emerging LLMs.

General-purpose pretrained language models are widely used in numerous mental health research studies, demonstrating the significance of pretraining and the effectiveness of deep contextualisation of texts. Researchers have incorporated various transformer-based language models [230], including BERT [56], RoBERTa [135], and Longformer [22], into mental health-related pipelines in various ways. Some researchers [69, 146, 268] have used them in a standalone manner, whether with finetuning or not, modifying the final layer head depending on the type of task. Others [27, 108, 145, 112] have primarily utilised them as encoders or feature extractors within a more complex framework.

TABLE 2.2. Widely utilised pretrained language models (PLMs) and emerging large language models (LLMs) for mental health downstream tasks.

Model Name	Domain	Base Model	Objective[a]	Training Data
BERT [56]	General	-	MLM , NSP	BooksCorpus, English Wikipedia
RoBERTa [135]	General	-	MLM	BookCorpus, CCNews, OpenWebText, Stories
Longformer [22]	General	-	MLM	BookCorpus, English Wikipedia, Realnews, Stories
BioBERT [116]	Medical Lit.	BERT	MLM	PubMed Abstracts, PMC Full-text articles
BioClinicalBERT [9]	Clinical Notes	BERT	MLM, NSP	MIMIC-III
PharmBERT [228]	Drugs	BERT	MLM	DailyMed
PubMedBERT [78]	Medical Lit.	BERT	MLM, NSP	PubMed Abstracts
BERTweet [168]	Social Media	BERT	MLM	Twitter Stream, Covid-19 Tweets
MentalBERT [98]	Mental Health	BERT	MLM	Reddit mental health posts
MentalRoBERTa [98]	Mental Health	RoBERTa	MLM	Reddit mental health posts
MentalXLNet [96]	Mental Health	XLNet	PLM	Reddit mental health posts
MentalLongformer [96]	Mental Health	Longformer	MLM	Reddit mental health posts
DisorBERT [15]	Mental Health	Bert	MLM	Reddit TIFU [109], Reddit mental health posts
MentalLLaMA [262]	Mental Health	LLaMA	CLM†	IMHI
MentalBART [262]	Mental Health	BART	CLM†	IMHI
MentalT5 [262]	Mental Health	T5	CLM†	IMHI

[a] MLM - Masked Language Modelling; NSP - Next Sentence Prediction; PLM - Permutation Language Modelling; CLM - Causal Language Modelling
† Objective of the finetuned base models.

Specialised PLMs have also been explored in a similar manner, where general-purpose language models are finetuned using large volumes of data for a particular domain. In a broader perspective, medical-related PLMs have been utilised for mental health-related tasks, such as BioBERT [116], BioClinicalBERT [9], PharmBERT [228], and PubMedBERT [78]. Social media-specific language models, such as BERTweet [168], have also been explored to capture the nuances of social media language that are not captured when pretraining with clinical documents.

Focusing on mental health in particular, researchers have explored the continued pretraining of general-purpose models using social media mental health data, as social media sources are more likely to contain the large amounts of data required for effective pretraining. MentalBERT and MentalRoBERTa [98] continue the pretraining of the BERT and RoBERTa models using approximately 13 million sentences from mental health-related social media datasets, employing the same masked language modelling and dynamic masking as the base models.

MentalXLNet and MentalLongformer [96] were then proposed and made available to handle long sequences of texts, with a maximum length of 512 and 4096 tokens, respectively, as BERT and RoBERTa are both limited to only 128 tokens. DisorBERT [15], on the other hand, recognises the computational complexity of MentalBERT and proposes a double-domain adaptation strategy which first adapts a base BERT model with Reddit language, then applies a guided masking strategy to adapt it to the mental health domain using only around 225k sentences for both steps, combined.

Finally, different large language models have also been adapted for mental health-related tasks. Yang et al. [262] have adapted multiple LLM base models for interpretable mental health analysis by reformulating different downstream tasks into a text generation one. They compiled various mental health datasets, primarily sourced from social media, and expanded them by generating task-based explanations using ChatGPT, resulting in the Interpretable Mental Health Instruction (IMHI) dataset. LLaMA-based models [223] were trained using instruction tuning methods [247] to create MentalLLaMA-chat models[4] of varying sizes. MentalBART and MentalT5 were also trained for completion-based generation, making them lighter than the instruction-based counterparts.

2.3 Multimedia Multimodality Mental Healthcare

Other media-based modalities have been extensively explored for mental healthcare-related tasks, particularly for disease detection. However, research on multimedia modality alignment and fusion in mental health has seen a steep increase in the past couple of years (Figure 2.1). This increase in attention is further highlighted by the integration of different media in large language models, where textual prompts are used to query non-textual inputs, even if the prompt itself does not contain the same contextual information as the input. For instance, with visual input, a question prompt may simply ask, "Does the person look depressed?" or an instruction prompt may be, "Describe the emotions shown by the person." Since a significant element in these use cases is the interaction between visual input and textual prompts, they are

[4] https://github.com/SteveKGYang/MentalLLaMA

TABLE 2.3. Mental health datasets in other media and multimedia modalities.

Dataset	Mental Health	Source/ Participants	Text	Audio	Visual	Physio./ Sensor	Others[a]	Avail.[b]
DAIC-WOZ [75]	Depression, PTSD	Recruited	✓	✓	✓	×	✓	DUA
E-DAIC [54]	Anxiety, Depression, PTSD	Recruited	✓	✓	✓	×	✓	DUA
CMDC [285]	Depression	Recruited	✓	✓	✓	×	×	DUA
RADAR-MDD [147]	Depression, Anxiety	Recruited	✓	✓	×	✓	×	Authors
MODMA [35]	Depression	Recruited	×	✓	×	✓	×	DUA
KUAH [117]	Schizophrenia	Recruited	×	×	✓	✓	×	Authors
MCIC [71]	Schizophrenia	Recruited	×	×	✓	✓	✓	Authors
HUSM [164]	Depression	Recruited	×	×	×	✓	×	Public
TDBRAIN [57]	Depression, ADHD, OCD	Recruited	×	×	×	✓	×	DUA
WU3D [243]	Depression	SM - Weibo	✓	×	✓	×	✓	Public
Sina-Weibo [40]	Suicide	SM - Weibo	✓	×	✓	×	×	DUA
D-Vlog [265]	Depression	SM - YouTube	✓	✓	✓	×	×	Authors
multiRedditDep [227]	Depression	SM - Reddit	✓	×	✓	×	×	Authors
RESTORE [259]	Depression	SM - Multiple	✓	×	✓	×	×	Authors
Axiom [149]	Anxiety	SM - Multiple	✓	×	✓	×	×	Authors

[a] Other modalities including other structured data such as demographic data and social media metadata.
[b] Availability: Public - Data is unrestricted in an online repository; DUA - Requires signing a data use agreement; Authors - Data may be requested from corresponding authors.

treated as multimodal studies. This section introduces other media modalities of mental health-related data and corresponding prevalent datasets. Table 2.3 enumerates several accessible datasets in non-textual modalities, including both unimodal and multimodal data.

Visual Data

Visual data used in mental health-related tasks may be explored in two ways. First, there are different types of image-based data, including social media-posted images, medical images, and satellite-based images. Second, different visual modalities may be extracted to decouple the features from the raw image source.

As with social media textual data, social media images have a wide range of variability, which presents a challenge in contextualising them. A publicly available social media dataset which includes image modalities is Weibo User Depression Detection Dataset (WU3D)[5] [243], providing data of 35k users from Weibo with a total of 2.1M posts and 1.2M images annotated

[5] https://github.com/aidenwang9867/Weibo-User-Depression-Detection-Dataset

by data labelling specialists and reviewed by experts. The dataset is mainly in the Chinese language and also provides text and other metadata information for each user and each post. Other studies have also collected data from image-based platforms, such as Flickr [62, 257] and Instagram [23, 44, 143], but these datasets are not publicly available.

In a medically grounded perspective, neuroimaging or brain imaging, specifically Structural Magnetic Resonance Imaging (MRI) data, has also been widely explored for mental health-related tasks, including schizophrenia [26, 196], depression [117, 276], bipolar disorder [117], and anxiety disorders [276] detection. From a macro perspective, Ouyang et al. [174] used satellite and street imagery to predict depression rates in New York City, highlighting features such as green spaces and infrastructure.

Derived visual modalities or features have also been widely used in numerous studies, as opposed to using raw image inputs. These features allow deep learning models to focus on specific attributes and atypical behaviour, including eye movements for schizophrenia [51] and depression [124, 248, 266] and human skeleton for ADHD [122, 125, 121]. Some studies have focused on faces to focus on facial expressions using face-specific images [126], face detection [74], and facial landmarks or action point units [1, 265]. Pre-extracting visual modalities has enabled some researchers to publicly release datasets that circumvent privacy issues, as the raw image or video is not shared; however, this limits the visual information that models can use.

Audio Data

Audio data for mental health-related tasks are mainly speech-based recordings since extensive studies have shown that vocal biomarkers have distinguishing capabilities among healthy participants and participants showing suicide ideation, depression, bipolar disorder, and schizophrenia [93, 114, 175]. These studies show that abnormal speech patterns typically manifest as low intonation, different speech rates, frequent and more prolonged pauses, and frequent throat clearing.

Widely used speech datasets are interview-based audio-visual datasets with transcribed textual conversations, which will be discussed in the following subsection. MODMA [35] is a

multimodal dataset for depression detection, providing Chinese speech data recorded during three tasks: an interview, a reading task, and a picture description task. Wang et al. [237], on the other hand, collected the Emotional Word Reading Experiment (EWRE) dataset, which focuses on word readings instead of dialogues. Despite this, several researchers have focused solely on using the audio modality. Common audio processing techniques for use in mental health deep learning models include direct encoding of raw audio signals [20, 58, 195], extraction of Mel-Frequency cepstral coefficients (MFCC) features [237, 265], filterbank energy (FBE) features [79], and spectrograms [273].

Audio-Visual Data

As discussed in previous sections, audio-visual data with textual transcriptions has been widely utilised in various ways, including text-only, audio-only, video/image-only, or any combination of the three, because of its inherent multimodality. Section 2.2.1 briefly introduced DAIC-WOZ and E-DAIC, which both provide raw audio and their corresponding transcripts. To maintain patient privacy, the raw video data is not released; instead, pre-extracted features are provided, including facial action points, gaze, and pose.

D-Vlog[6] [265] is another audio-visual dataset for depression detection containing 961 vlogs from the YouTube platform. Unlike clinical interview-based datasets, this dataset highlights depression in real-world, daily life settings. However, as a social media-based dataset, data collection is done through key phrase matching with human annotators labelling each video based on what the user in the video says. Raw audio and video are not publicly available for similar reasons, but the authors provide pre-extracted acoustic and visual features.

Physiological and Sensor-based Data

Physiological data are measures of real-time bodily functions through different biological signals in a person's body. These types of data are the most objective for mental health detection, as they present unbiased information regarding the actual physical effects of emotions, rather than relying on subjective thoughts and feelings. Because of this, mental health research has explored the use of different physiological modalities among which, common signals utilised are EEG (electroencephalogram) [21, 52, 76, 73, 187, 275, 281] and

[6] https://sites.google.com/view/jeewoo-yoon/dataset

MEG (magnetoencephalography) [187] measuring brain activity, ECG (elecrocardiogram) [187] measuring heart rhythm, and fNIRS (functional near-infrared spectroscopy) [242] or fMRI (functional MRI) [173] measuring brain blood oxygen.

While most research using physiological data mainly uses their own private data, a few datasets are available for research after signing data use agreements (DUA) or verification. The Two Decades Brainclinics Research Archive for Insights in Neuroscience (TDBRAIN)[7] dataset [57] is an EEG dataset of 1,274 psychiatric patients with different diagnoses, including major depressive disorder (MDD), attention deficit hyperactivity disorder (ADD), and obsessive-compulsive disorder (OCD) over 20 years. The MODMA dataset, introduced in the audio section, also provides full brain and 3-electrode EEG data from each healthy and depressed participant.

Other sensor-based data have also been explored since physiological data are not easily accessible and require specialised equipment and a certain level of medical and technical expertise to collect and interpret. RADAR-MDD [147] utilises wearable sensors and mobile remote measurement technologies to collect data from participants with depression over multiple sites. It collects passive data, such as ambient noise, ambient light, and GPS location, using smartphones. Additionally, it gathers activity data from a FitBit and mental health data through questionnaires.

2.4 Summary

This chapter provides a broad overview of mental health-related tasks, datasets, and frameworks, focusing on media-based modalities. It highlights the advantages and limitations of each medium and various data sources, presenting opportunities for applying NLP and deep learning in advancing mental health systems.

While many researchers have explored visual, audio, and physiological data, there is still an evident dominance of textual data both in unimodal and multimodal settings (Figure 2.1). In

[7] https://www.brainclinics.com/resources

unimodal text-only research, social media-based datasets are most prominent due to the ease of data collection in online settings. This results in larger datasets, enabling better contextual learning for language models. Researchers are also drawn to the candidness and real-time recording of thoughts and feelings in social media user-generated texts, which is a particular interest in most mental health studies.

In multimodal settings, textual data remains prevalent among other media, particularly when used with audio and visual data. This prevalence is in part due to the flexibility of textual data, which can represent and be derived from other modalities, such as transcripts from speech audio or textual descriptions of visual data, including image captions for posted images and accompanying reports of medical imaging. While still less common, studies incorporating physiological and sensor-based data with large language models [59, 275] have emerged in the past year and are likely to show a rapid increase in the succeeding years.

One prevailing challenge with mental health deep learning research, especially among non-text-based datasets, is the availability and accessibility of data used for training and evaluating models and frameworks. Due to heightened privacy laws and ethical concerns, a large portion of studies are unable to share the data they used publicly, as strict privacy and ethics laws prevent them from doing so. Some researchers share their data with other researchers only after the requesting party has provided Institutional Review Board (IRB) or ethics committee approvals, which is typically a lengthy process and is exclusive to each dataset. Even previously available social media-based datasets have been withdrawn from public access due to ethical considerations [29].

2.4.1 Implications for this Thesis

Limitations in recent literature

This chapter highlights two limitations that this thesis aims to address. First, as an overarching motivation, are the limitations brought about by the accessibility and availability of non-textual datasets. Despite the availability of a few public non-text-based datasets, their rarity limits the generalisability of the models trained and evaluated on them. Most multimodal

studies also build architectures based solely on the modalities available to them, hindering the development of a general framework. Text-based mental health data remains more available, accessible, and easier to collect in large quantities, motivating this research to focus on exhausting textual information through textual abstractions.

Moreover, this thesis recognises the limitations of conventional language modelling methods. In particular, language model semantic spaces lose affective information, especially for words that lack emotional connotations. Without tonal cues from verbal delivery, implicit emotions continue to be a challenge for language models. Furthermore, there is an inherent sequential bias in language models resulting from pretraining objectives that rely on neighbouring words to contextualise a token. This sequential bias overlooks the underlying structural dependencies of words, thereby limiting tasks that involve complex structures.

Approcahes explored by this thesis

The ubiquitousness of mental health-related textual data, along with its growing integration with other media-based modalities, especially with LLMs, and the limitations of conventional semantic spaces from language models, demonstrate the persistent need to improve textual representations in text-based architectures. This thesis takes an exhaustive perspective on mental health-related texts through abstractive multimodality, as opposed to medium-based multimodality or multimedia data.

In particular, this thesis explores three distinct complementary information represented at their own abstraction level. Addressing the affective neutrality in the semantic space of language models, the emotion modality captures affective abstractions from the raw text. Concurrently, the acoustic modality, even if derived, abstracts paralinguistic and prosodic biomarkers addressing the loss of tonal cues in textual data. Finally, the word-pair modality, representing relational and structural abstractions of textual components, disentangles the sequential bias of current language models. The succeeding chapters detail novel methods of integrating these modalities and their applications text-based mental health downstream tasks.

CHAPTER 3

Multi-Emotion Representations for Mental Health Classification

---

This chapter is the published work **MM-EMOG: Multi-Label Emotion Graph Representation for Mental Health Classification on Social Media** [33] published in the **Robotics** journal. I am a first author of this paper. I formulated the research aim and co-designed the methodology. I was primarily responsible for collecting and analysing datasets, conducting all of the the experiments, analysing the results, and writing most of the manuscript.

This work explores the integration of complex emotions for mental health detection in social media texts. In particular, this study addresses the affective neutrality in the learned semantic space of language models, especially for implicit emotions conveyed through words lacking emotional connotations (Table 1.1). By incorporating the emotion modality at the representational level, both implicit and explicit, heterogeneous affective abstractions from the textual data are incorporated in the representation space. The emotion modality is integrated with semantic representations through a multi-emotion graph-based pretraining framework, enabling the simultaneous learning of global and local affect patterns (Table 1.2). This framework is applied to emotion-charged mental health-related texts for the detection of depression and suicidal ideation expressed in social media.

More than 80% of people who commit suicide disclose their intention to do so on social media. The main information we can use in social media is user-generated posts, since personal information is not always available. Identifying all possible emotions in a single textual post is crucial to detecting the user's mental state; however, human emotions are very complex, and a single text instance likely expresses multiple emotions. This paper proposes a new multi-label emotion graph representation for social media post-based mental health classification. We first construct a word-document graph tensor to describe emotion-based

contextual representation using emotion lexicons. Then, it is trained by multi-label emotions and conducts a graph propagation for harmonising heterogeneous emotional information, and is applied to a textual graph mental health classification. We perform extensive experiments on three publicly available social media mental health classification datasets, and the results show clear improvements[1].

3.1 Introduction

3.1.1 Background

According to the World Health Organisation (WHO) [252], it is revealed that a staggering majority of individuals who tragically succumb to suicide, surpassing 80%, choose to divulge their suicidal ideation and intentions on social media platforms. This trend underscores the profound impact and far-reaching implications that social media can have on mental health and well-being. The disclosure of such deeply personal and troubling thoughts on these digital platforms presents a unique opportunity for early intervention and detection of mental disorders, as well as potential suicidal tendencies. The significance of recognising and addressing these issues promptly extends beyond individual well-being to the broader spectrum of societal welfare, making early detection a vital component for fostering good governance. By understanding the intricate connection between social media expressions and mental health indicators, policymakers and healthcare professionals can proactively implement measures to provide timely support and intervention, thereby contributing to the overall enhancement of mental health outcomes on a societal level. In essence, the revelation of suicidal ideation on social media acts as a crucial signal for the imperative need to prioritise and enhance mental health surveillance and support systems for the greater well-being of communities.

One challenge of social media research, however, is the privacy and protection of one's identity. Users tend to prefer to be completely anonymous or withhold personal details and confidential information, making it impossible to observe. Accessing contextual components,

[1]Code is available at https://github.com/adlnlp/mm_emog

such as historical posts and user or post metadata information, has also become increasingly restrictive due to heightened data protocols from social media platforms, further complicating research reproducibility. Due to this trend, the primary information easily accessible for mental health detection from social media is user-generated posts. Our research focuses on detecting mental illnesses through the analysis of social media textual posts only, as opposed to other forms of social media data, with the question, *'What would be the most important component from which we can identify the mental health condition using pure text from social media?'* The answer can be found in the WHO's definition of mental disorder, stating that *'A mental disorder is characterized by a clinically significant disturbance in an individual's cognition, emotional regulation, or behaviour.'* [251]. The ideal setup for mental state detection via textual posts would identify all possible emotions and integrate those feelings and emotional statuses.

Recent studies incorporate emotions into mental health classification by fine-tuning pre-trained embeddings through a single emotion classification task [113, 201]. Due to the complexity of human emotions, multiple emotions are likely to be expressed in a single textual post, and those emotions can be correlated. To represent emotions and their correlation with the text, we can consider two types of textual representation techniques: sequential text representation and graph-based text representation. While sequential text representation promotes capturing text features from local consecutive word sequences, graph-based text representation can attract widespread attention and successfully understand word and document relationships [264, 133, 239].

3.1.2 Research Aims

This paper proposes MM-EMOG, a novel multi-label, graph-based emotion representation for mental health classification, utilising user-generated social media posts. Note that we focus on post-only-based mental health classification due to privacy issues and restrictions in social media. To achieve this aim, we first construct a word-document graph tensor to generate emotion-based contextual representations using emotion lexicons. These representations are then pre-trained through a multi-label emotion classification task by conducting graph

propagation and transformer backpropagation to harmonise heterogeneous emotional information. The trained multi-emotion representation is then applied to a textual graph mental health classification model.

3.1.3 Main Contributions

The main contributions of this paper are as follows:

(1) We propose a novel multi-label emotion representation for mental health classification, utilising only social media textual posts.
(2) To our knowledge, no other studies have utilised Graph Convolutional Neural Networks [110] (GCN) in a purely textual capacity for multi-label emotion representation and social media mental health classification tasks before this. We create a novel multi-label and graph-based textual emotion representation.
(3) Our proposed model, MM-EMOG, achieved the highest performance on three publicly available social media datasets for mental health classification.

3.2 Related Works

3.2.1 Social Media Mental Health Classification

Social media has opened up new opportunities for suicide ideation and mental health studies by creating a new way to access information-rich data, not just of clinical patients but of a broader subset of the public outside clinical settings. Recent studies have focused on incorporating more social media components to capture as much available contextual information as possible. Among these are historical posts [31, 39, 40, 148, 199, 200, 198, 201, 209, 211, 284], conversation trees [202], social and interaction graphs [39, 148, 158, 201, 211], user and post metadata information [39, 40], and images [39]. While more contextual sources may be ideal for assessing an individual's mental health state, access to these data has become increasingly restrictive due to heightened data privacy concerns. It also complicates research reproducibility, as each study selects features based on the social media components

available to them. Because of this, our proposed system focuses on improving contextual information by incorporating emotions from the most basic social media component– a single textual post.

Emotions have been a long-standing area of interest for natural language processing (NLP) researchers, as language has been one of the primary avenues for conveying emotions. We are particularly interested in emotions since mental health is deeply rooted in an individual's thoughts and feelings. Traditional methods of incorporating emotions into mental health classification tasks involved frequency- or score-based emotion features [13, 67, 158, 162, 209, 226, 283]. For instance, Aragón et al. [13] proposed Bag-of-Sub-Emotions (BoSE), a histogram-based document emotion representation for the detection of depression and anorexia in online forums. Zogan et al. [283] implemented a feature selection process that incorporates positive, negative, and neutral emoji frequencies, as well as valence, arousal, and dominance scores, for depression detection. More recent studies have used machine and deep learning to fine-tune contextual embeddings using mental health classification as a downstream task [113, 190, 199, 198, 201]. Ren et al. [190] proposed an Emotion Understanding Network (EUN) where positive and negative word embeddings are learned separately by two attention networks later combined with a Semantic Understanding Network (SUN) for depression detection. Sawhney et al. [198] fine-tuned a pre-trained transformer on the EmoNet dataset [2] to extract the emotional spectrum of each post for suicide ideation detection in social media. However, these studies focus on learning a single emotion for a single word or an entire text, which undermines the complexity of human emotions, wherein a single word may express multiple emotions. Our proposed system integrates emotional context by harmonising heterogeneous emotions through a multi-label, corpus-based representation pre-training framework.

3.2.2 Graph Convolutional Networks

Graph Convolutional Networks (GCN) [110] have seen increased applications in recent years due to their versatility in learning entity representations. TextGCN [264] further adopts the GCN framework to learn representations for both words and documents without the use of

external embeddings. Recent studies have further improved upon this by learning multiple types of information through multi-edge [239] and multi-aspect [133] graphs. These have since been applied to various tasks, such as long and short document classification [130, 132], aspect-based sentiment analysis (ABSA) [127, 219, 271], and general text classification tasks [264, 277]. However, despite the incorporation of graph-based data, mental health classification in social media has been limited to user-user networks and user-post graphs [39, 148, 158, 211]. To our knowledge, no other studies have utilised GCN in a purely textual capacity for this task. Our study will leverage the corpus-wide learning of TextGCN by thoroughly exhausting all co-occurrence relationships between words and documents. Our system utilises this corpus-based learning in a multi-label emotion pre-training framework to harmonise complex contextual and emotional information contained in mental-health-related textual posts.

3.3 MM-EMOG

3.3.1 MM-EMOG Construction

We adapt TextGCN [264] to learn the local and global emotional trend of mental health in social media though a graph-based structure $G = (V, E, A)$, where V is a set of word and document nodes, E is a set of word-word edges $E_{w_i w_j}$, word-doc edges $E_{w_i d_j}$, and doc-doc edges $E_{d_i d_j}$, and $A \in R^{N \times N}$ defines the weights of these associations. Figure 3.1 Step 1 shows the MM-EMOG architecture.

Node Construction

We first preprocess the post text in two steps. First, we further de-identified emails, usernames, and URLs by replacing them with reserved tokens. Second, we conduct emoticon preservation by retaining emoticons and emojis to be contextualised as individual tokens.

After preprocessing, we then create nodes by using each post as a document node and each token in the corpus as the word or token node. Token nodes are created either through (1) word split tokenisation or (2) wordpiece tokenisation using the pre-trained tokeniser from

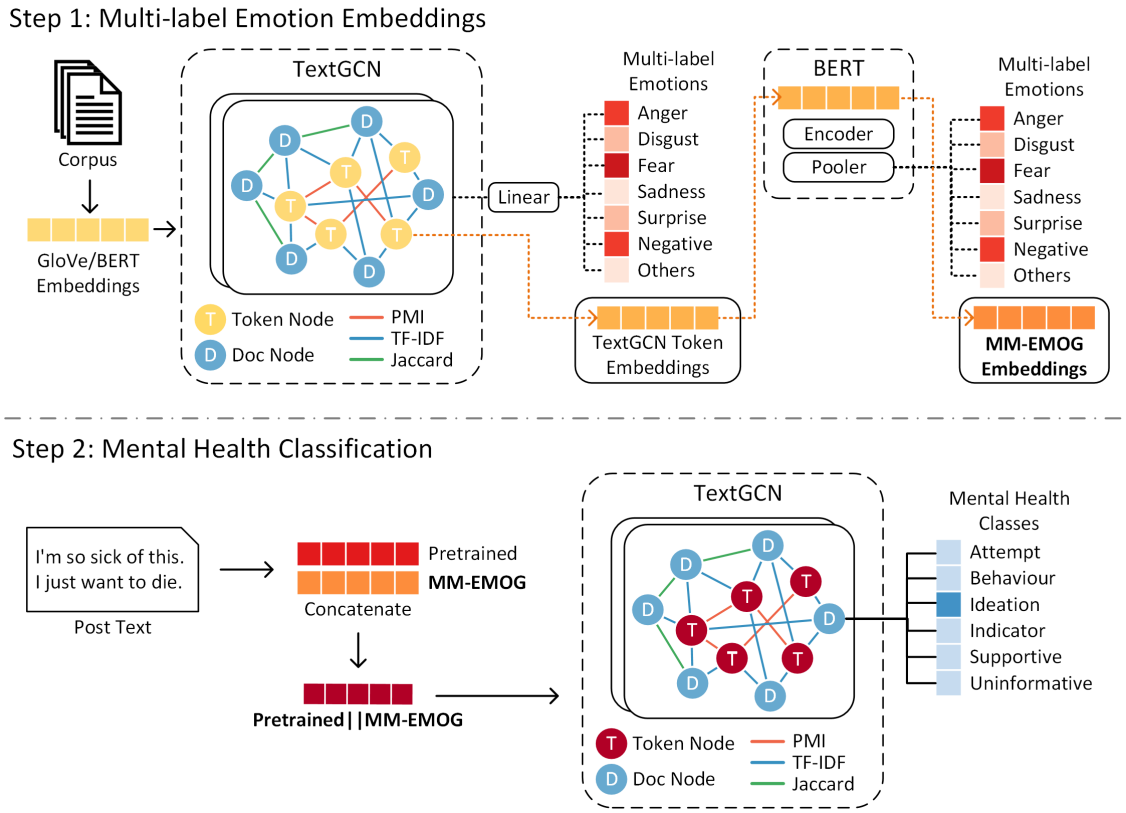


FIGURE 3.1. An overview of the proposed system illustrating the architectures for MM-EMOG pre-training (Step 1) and their application to mental health classification (Step 2). A textual graph is created to learn multi-emotion embeddings, where nodes represent tokens and documents in the corpus, while edges represent the relationships between them. The graph is passed to a two-layer GCN and a linear layer for a multi-label emotion classification task. After training, token node representations are extracted from the second GCN layer and used as initial weights for fine-tuning a pre-trained BERT model for the same task. After fine-tuning, the embeddings are extracted as the MM-EMOG embeddings for a graph-based mental health classification task.

the Bidirectional Encoder Representations from Transformers (BERT) [55] implementation. For wordpiece tokens, we incorporate emoticons into the tokeniser vocabulary for emoticon preservation and only apply lowercasing without additional text preprocessing. For word split tokens, we employ a simple text cleaning process that removes some punctuation and separates contractions. Stopwords are kept to retain negation words.

Finally, we initialise word split token nodes using Global Vectors (GloVe) [179] as embeddings and use the average weight of all token nodes to represent the document node. For wordpiece

tokens, we follow [82] and use contextualised BERT embeddings, where the learned vector for the `[CLS]` token is used to initialise each document node. For each unique token, all contextualised vector representations of the token are collected from the entire corpus. Minimum pooling is then performed to initialise each token node.

Edge Construction

Inspired by [82], we leverage all types of co-occurrence relationships between and among tokens and documents using three different types of edges. First, we construct token-to-token edges $E_{w_i w_j}$ representing the relationship between token w_i and w_j using the Pointwise Mutual Information (PMI) [45], a measure of association between two words. Token-to-document edges $E_{w_i d_j}$ are represented by the Term Frequency-Inverse Document Frequency (TF-IDF) between token w_i and document d_j. Finally, the association between documents d_i and d_j are represented by the Jaccard similarity to represent the document-to-document edges $E_{d_i d_j}$. The final set of edges is $E = \{E_{w_i w_j}, E_{w_i d_j}, E_{d_i d_j}\}$. Formally, the graph adjacency matrix is defined as:

$$A_{ij} = \begin{cases} PMI_{ij} & : i, j \text{ are words; } PMI > 0 \\ TF-IDF_{ij} & : i \text{ is word, } j \text{ is document} \\ Jaccard_{ij} & : i, j \text{ are documents} \\ 0 & : \text{otherwise} \end{cases}$$

By representing different granularities of social media textual posts and incorporating all possible co-occurrence relationships among the different textual components, the MM-EMOG graph ensures a comprehensive capture of both local and global information that is beneficial for understanding and contexualising emotional nuances in words and textual posts.

3.3.2 MM-EMOG Pre-training

To produce emotion-rich contextual representations from mental-health related texts, we leverage the corpus-wide neighbourhood information from the textual graph constructed in Section 3.3.1 and focus on learning heterogeneous emotions in a multi-label framework,

which has yet been explored by other studies. This multi-emotion classification task serves as the MM-EMOG pre-training objective which doesn't expose it to any mental health risk annotation or labels.

Multi-label Document Emotions

We first extract multi-label emotion classes at the document-level using emotion lexicons (Section 3.4.2) containing word-emotion and word-sentiment associations[2]. Assume a document with words W=$\{w_1, \ldots w_p\}$, where p is the number of unique words in the document, and a lexicon containing terms K=$\{k_1, \ldots k_q\}$ and emotion set EM=$\{em_1, \ldots em_r\}$, where q is the number of lexicon terms, r is the number of emotion classes in the lexicon, and where each lexicon term k_j is associated with one or more emotions EM_{k_j}. When a word from the document matches a lexicon term w_i=k_j, we extract the emotions EM_{k_j} and associate it with w_i. We note that positive emotions are grouped into the "other" class to motivate learning more nuance from negative emotions more related to higher-risk mental health classes. The final multi-label emotion class for the document is the union of all emotions associated with all the words in the document EM_d=$\{EM_{w_1} \cup EM_{w_2} \cup \ldots EM_{w_p}\}$.

Multi-label Emotion Training

To incorporate complex emotions into contextual embeddings, we conduct a multi-label emotion classification task by passing the node representations V and the adjacency matrix A to a two-layer GCN followed by a linear layer with an output dimension of r emotions. The second layer of the GCN has a dimension of $s = 768$ to follow popular pre-trained embeddings such as BERT [55] and RoBERTa [135].

Graph propagation takes the input and maps each instance to multiple emotions. In particular, the first GCN layer takes the input graph G to learn an input-to-hidden weight matrix $W^{(0)} \in R^{d \times H}$ [110], where d is the dimension of node representations and H is the hidden dimension size. This weight matrix is subsequently passed to the second GCN layer to produce a hidden-to-hidden weight matrix $W^{(1)} \in R^{H \times s}$. The final linear layer learns the hidden-to-output weights $W^{out} \in R^{s \times r}$. ReLu is used with binary cross-entropy loss for multi-label learning.

[2]Note that we refer to both emotion and sentiment as "emotion" for the purposes of this paper.

Backpropagation updates the initial representations to incorporate emotional information during model training.

The learned token node representations $W^{(1)}$ from the second GCN layer are extracted and used as initial weights to further finetune the representations using the BERT framework for the same multi-label emotion classification task. Similarly, binary cross entropy is utilised as loss function. Finally, the learned weights are extracted as multi-emotion contextual representations, MM-EMOG EmoWord (EW) or EmoWordPiece (EWP) embeddings.

The two-step multi-label emotion training pipeline ensures thorough incorporation of emotion information to the contextual embeddings by combining the benefits from the global and local component association of graph networks and from the temporal dependencies of sequential networks.

3.3.3 Mental Health Post Classification

We evaluate the effectiveness of our multi-label emotion training through the use of the final MM-EMOG representations in a post-based mental health classification task (Figure 3.1 Step 2). While we recognise the benefits and use cases of a user-based classification, the high granularity in post-based classification allows our model to be more scalable, more straightforward, and less invasive than a user-based classification that requires access to other social media components, such as historical posts and user metadata. Furthermore, post-based risk levels can be valuable interpretability points in a user-based analysis or can be aggregated to be incorporated in a user-based model; however, the opposite is not possible.

Similar to the pre-training in Step 1, we leverage the corpus-wide co-occurrence information from TextGCN using the same graph construction method. For token node representations, we concatenate BERT and MM-EMOG embeddings, while for document node representations, the average of all token representations within each document is used. Finally, the graph is passed to a two-layer GCN for post-based mental health classification following a similar graph propagation as the MM-EMOG training. However, for this task, the second GCN layer learns a hidden-to-output weight matrix $W^{mh} \in R^{H \times c}$, where c is the number of mental health

TABLE 3.1. Dataset statistics. CV: cross-validation

	TwitSuicide	**CSSRS**	**Depression**
Platform	Twitter	Reddit	Twitter
Total Posts	660	2,680	3,200
Total Users	645	375	-
Number of Classes	3	6	2
Evaluation Method	10-fold CV	5-fold CV	80/20
Length	13-147	2-6,221	6-374
Average Length	90.32	451.67	90.08
Word Count	3-31	1-1,051	1-77
Average Word Count	16.85	85.51	17.43

classes defined in each dataset evaluated. Categorical cross-entropy is used for single-label classification.

3.4 Experimental Setup

3.4.1 Datasets

We use three publicly available post-based datasets to evaluate MM-EMOG. Table 3.1 summarises the statistics, while Figure 3.2 shows the distribution of classes for each dataset.

The **TwitSuicide Dataset**[3] (TwitSuicide) [138] replicates the data collection and processing methods of Odea et. al. [171]. Posts are gathered from X (formerly called Twitter) and filtered using suicide-related terms and phrases such as *"kill myself"*, *"want to die"*, and *"better off without me"*. A sample of 660 tweets from 645 users was annotated by one psychologist and two computer scientists following previously established guidelines and three suicide risk levels [171]. The *Strongly Concerning* (SC) class is assigned to posts with a convincing display of serious suicidal ideation. *Possibly Concerning* (PC) is a designated default category where a post is only removed from this class if it falls into other categories. Lastly, the *Safe to Ignore* (SI) class presents no reasonable evidence to suggest that suicide risk is present. The distribution of the three classes are 15.61%, 40.00%, and 44.39% respectively.

[3]Data available upon request.

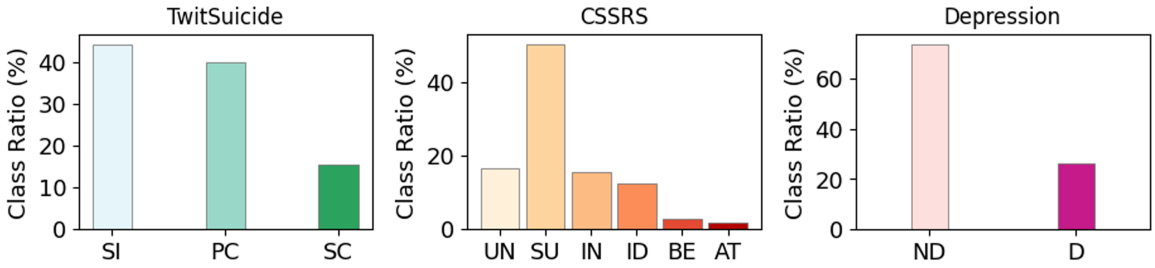


FIGURE 3.2. Class distribution for each dataset. For TwitSuicide (left), SI: Safe to Ignore, PC: Possibly Concerning, SC: Strongly Concerning. For CSSRS (center), UN: Uninformative, SU: Supportive, IN: Indicator, ID: Ideation, BE: Behaviour, and AT: Attempt. For Depression (right), ND: Non-depression and D: Depression.

The **Reddit C-SSRS Dataset**[4] (CSSRS) [67] is acquired from 15 mental health-related subreddits or forums in the Reddit platform. Users who actively participated in the subreddit *r/SuicideWatch* are treated as potentially suicidal. Four clinical psychiatrists refined this user classification using the Columbia-Suicide Severity Rating Scale (C-SSRS), a short questionnaire used by mental health professionals to quickly asses suicide risk in clinical practice. The questionnaire categorises a patient into three risk levels; however, to adapt for social media use and post-level annotations, three lower-risk levels were added for distinguishing unlikely suicidal users merely using suicide-related terms for discussion and support and for identifying irrelevant posts. The final six classes from high to low risk are *Actual Attempt* (AT; 1.83%), *Suicidal Behavior* (BE; 2.87%), *Suicidal Ideation*(ID; 12.57%), *Suicidal Indicator* (IN; 15.67%), *Supportive* (SU; 50.45%), and *Uninformative* (UN; 16.60%). We use the annonymised post-level dataset as provided with 375 users and medical entity normalised posts.

This **Twitter Depression Dataset**[5] (Depression) is the basis for the practice dataset of the CLPsych 2021 shared task [142]. The data is collected from Twitter using the hashtags "#depressed", "#depression", "#loneliness", and "#hopelessness1". After further filtering, all hashtags are removed to minimise bias introduced by the search parameters. Binary annotation was done by the authors where all posts are classified as depressive unless deemed otherwise

[4] https://github.com/AmanuelF/Suicide-Risk-Assessment-using-Reddit
[5] https://github.com/swcwang/depression-detection

TABLE 3.2. Emotion types for each lexicon.

Lexicon	Description	Emotion Types
EmoLex [159]	Crowdsourced word-emotion and word-polarity pairings	Anger, Anticipation*, Disgust, Fear, Joy*, Sadness, Surprise, Trust*, Positive*, Negative
TEC [160]	Automatic annotation from emotion-related hashtags on Twitter	Anger, Anticipation*, Disgust, Fear, Joy*, Sadness, Surprise, Trust*
SenticNet [37]	Concepts from common sense knowledge graphs associated with emotions through similarity prediction	Anger, Calmness*, Disgust, Eagerness*, Fear, Joy*, Pleasantness*, Sadness, Positive*, Negative

* Combined into “other".

after a manual review. Additional non-depressive texts were added creating a dataset of 3,200 posts. We refer to these classes as *Depression* (D; 26.34%) and *Non-Depression* (ND; 73.66%) classes.

3.4.2 Emotion Lexicons

To create emotion-rich contextual embeddings that capture the complex emotions within mental health-related posts, we use three widely used emotion lexicons that associate one or more emotions or sentiments to words or concepts. Table 3.2 enumerates the emotion types for each lexicon.

The **NRC Emotion Lexicon**[6] (EmoLex) [159] is a crowdsourced word-emotion and word-polarity pairings. The lexicon contains 6,453 terms matched to at least one of two sentiments or eight emotions.

The **NRC Twitter Emotion Corpus**[7](TEC) [160] is lexicon generated using emotion hashtags from Twitter. Word co-occurrence scores determine the word-emotion association. We apply a threshold of at least 0.5 to remove weakly associated pairs. A total of 16,862 terms, including hashtags, emoticons, common stop words, proper names, and numerical figures, are associated with at least one of eight emotions.

[6] https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
[7] http://saifmohammad.com/WebPages/lexicons.html

SenticNet[8] [37] is a concept-level knowledge base created through commonsense knowledge graphs. We use SenticNet7, which generates symbolic representations through subsymbolic techniques. A total of 149,673 concepts, including emoticons and emojis, are associated with one sentiment and two of 24 fine-grained emotions. We simplify these to eight primary emotions by grouping them based on their positive and negative intensity levels. Furthermore, for simplicity, we only utilise one-word concepts.

3.4.3 Baselines and Metrics

We evaluate the performance of our proposed system against four transformer-based pre-trained language models (PLM). **BERT** [55] was trained on next sentence prediction and masked language modelling that generated state-of-the-art performance on multiple NLP-related tasks. **RoBERTa** [135] replicated the BERT training setup but used a dynamic masking pattern and removed the next sentence prediction objective. **MentalBERT** and **MentalRoBERTa** [97] followed the training procedures of BERT and RoBERTa, but used data from different mental health subreddits, including *r/SuicideWatch*, *r/Anxiety*, *r/bipolar*, *r/mentalillness*, and *r/mentalhealth*. More details on these PLMs may be found on Table 2.2.

All baseline models are trained for 15 epochs with a 1e-4 learning rate, a maximum length of 256, and a batch size of 8. Other hyperparameters are left to the default values set by the HuggingFace[9] library. Due to class imbalance, we evaluate overall performance using accuracy and weighted F1 (F1w) scores. Class F1 scores are provided to show detailed breakdown of class performance, particularly the more concerning or higher risk classes.

3.4.4 Implementation Details

MM-EMOG is modelled using the entire corpus of each dataset with a train/validation split of 90:10 and is trained for 200 epochs with a 10-epoch early stop using the Adam optimiser for both TextGCN and BERT training phases. The TextGCN phase uses the following

[8] https://sentic.net/downloads/
[9] https://huggingface.co/

TABLE 3.3. Best-found hyperparameters for each dataset using all lexicons and all preprocessing setups. Emo: EmoLex; Sen: SenticNet.

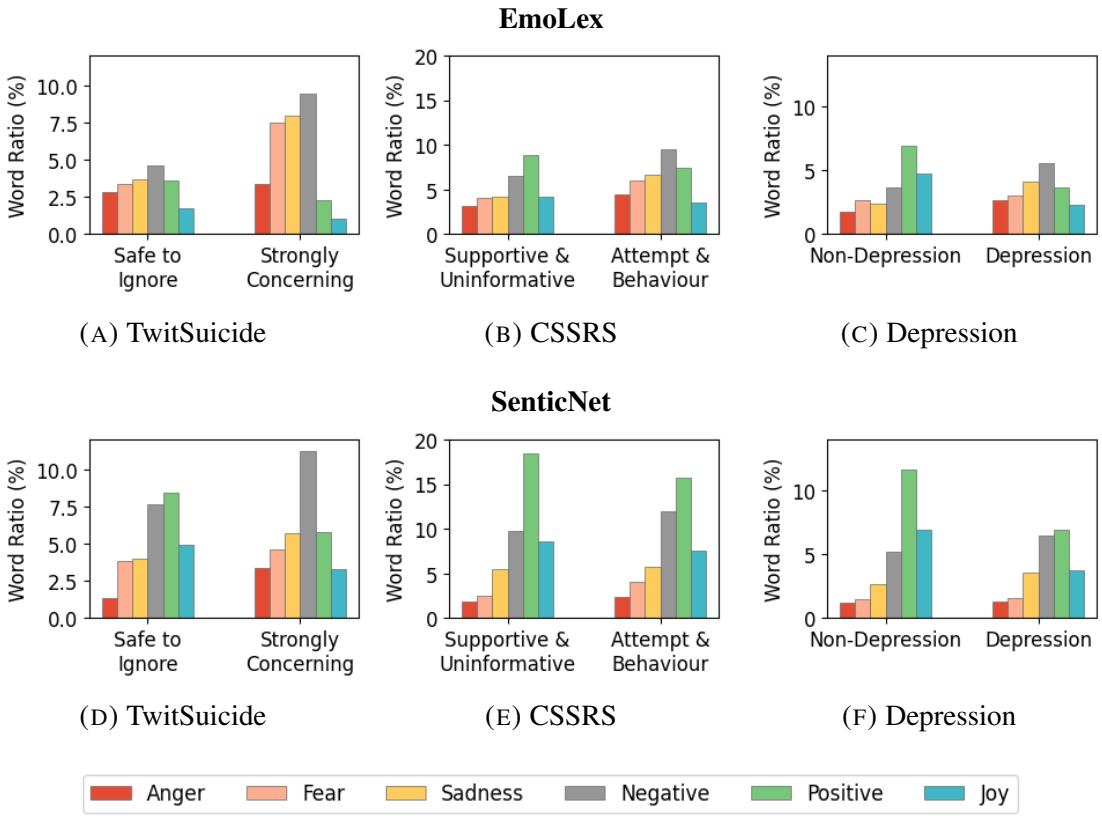
	TwitSuicide			CSSRS			Depression		
	Emo	**TEC**	**Sen**	**Emo**	**TEC**	**Sen**	**Emo**	**TEC**	**Sen**
EW1									
dropout	0.5	0.01	0.5	0.01	0.1	0.05	0.01	0.1	0.05
number of layers	4	2	2	2	2	2	2	2	2
hidden dimension	200	400	400	300	200	500	200	200	200
learning rate	0.01	0.03	0.04	0.05	0.03	0.03	0.05	0.02	0.05
EW2									
dropout	0.5	0.01	0.01	0.01	0.01	0.05	0.01	0.5	0.05
number of layers	2	2	2	2	2	2	2	2	2
hidden dimension	200	200	400	300	400	200	200	200	200
learning rate	0.02	0.05	0.01	0.03	0.05	0.04	0.05	0.03	0.01
EWP1									
dropout	0.5	0.01	0.1	0.01	0.1	0.5	0.1	0.5	0.1
number of layers	2	2	2	2	2	2	2	2	2
hidden dimension	100	100	200	200	200	400	200	200	200
learning rate	0.05	0.01	0.04	0.04	0.04	0.05	0.01	0.02	0.05
EWP2									
dropout	0.5	0.5	0.1	0.01	0.5	0.05	0.05	0.1	0.01
number of layers	2	2	5	2	2	2	2	2	2
hidden dimension	200	500	300	200	500	200	200	200	200
learning rate	0.02	0.04	0.04	0.04	0.04	0.05	0.04	0.05	0.02

EW: word split; EWP: word piece; 1: simple cleaning; 2: added de-identification and emoticon preservation.

hyperparameters: 200 hidden dimension size, 0.5 dropout, 0.02 learning rate, and 2 GCN layers. The BERT phase uses 0.5 dropout, 1e-5 learning rate, and 256 maximum length. Batch size is set to 64, 32, and 16 for the TwitSuicide, CSSRS, and Depression datasets, respectively.

For the post-based mental health classification task, this reasearch uses established evaluation setups following previous studies for each dataset: cross-validation (CV) with 10- and 5-fold setups for TwitSuicide and CSSRS, respectively, and a 80:20 train/test split for Depression. All classification models are tuned using Optuna[10], a hyperparameter optimisation framework, at a 90:10 validation split. The hyperparameters tuned and the search space for each are the number

[10] https://github.com/optuna/optuna



(A) TwitSuicide (B) CSSRS (C) Depression

(D) TwitSuicide (E) CSSRS (F) Depression

FIGURE 3.3. Emotion label distribution using the EmoLex (top) and SenticNet (bottom) lexicons on the three benchmark datasets.

of hidden layers $L = \{2, 3, 4, 5\}$, hidden layer dimension $H = \{100, 200, 300, 400, 500\}$, dropout $dr = \{0.01, 0.05, 0.1, 0.5\}$, learning rate $lr = \{0.01, 0.02, 0.03, 0.04, 0.05\}$, and weight decay $wd = \{0, 0.005, 0.05\}$. Each model setup was searched separately for 50 trials, maximising accuracy using a 90:10 split of the whole corpus for cross-validated datasets or of the training set for datasets with defined splits. Best-found hyperparameters for each model are presented on Table 3.3. Results are reported in Section 3.6.1 based on an average of 10 independent runs on Google Colab GPU-hosted runtimes.

3.5 Emotion Analysis

An aim of this study is the creation of contextualised representations incorporating heterogeneous emotions associated with mental-health-related text. This goal heavily relies on learning

complex emotions associated with words. Because of this, we evaluate the use of emotion lexicons to create multi-label emotion classes. This is done through matching each post word or token to corresponding emotion classes from each lexicon. Figure 3.3 summarises the resulting emotion label distribution using the EmoLex and SenticNet lexicons where an increase of negative emotions from the least to the most concerning classes is observed. In contrast, a negative trend emerges for the positive emotions. These trends demonstrate how social media posts contain emotional markers consistent with different levels of suicide ideation and depression. The heterogeneity of these emotions motivates the use of a multi-label approach in learning emotional contextual representations for mental health classification.

3.6 Results

3.6.1 Overall Performance

The MM-EMOG representations are evaluated through a mental health classification task using a graph-based model. Table 3.4 shows results from our proposed system against pre-trained language model baselines finetuned for the mental health classification task. Due to the small percentages of the most concerning classes for the CSSRS dataset, we report the averaged weighted F1-score for *Attempt* (AT), *Behaviour* (BE), and *Ideation* (ID) classes (Section 3.4.1).

Overall, our system outperforms all the baselines with 8%, 21%, and 14% improvement for TwitSuicide, CSSRS, and Depression datasets, respectively. Moreover, the notable increase in performance over the most concerning classes shows that, through multi-label contextual emotion representation learning, MM-EMOG can capture emotional intricacies where heightened negative emotions are present. We note that due to the severe binary class imbalance of 74:26, all the baselines for the Depression dataset are only predicting the majority class. Without using class weights or balancing methods, using MM-EMOG produces better performance.

TABLE 3.4. The overall results of our mental health classification model using MM-EMOG with BERT over different emotion-based lexicons. The best scores are **bold faced**; the second best are underlined. Class-based scores are shown for the most and least concerning classes for each dataset. For TwitSuicide: Strongly Concerning (SC) and Safe to Ignore (SI). For CSSRS: weighted average of Attempt, Behaviour, and Ideation (A,B,I) and Uninformative (UN). For Depression: Depression (D) and Non-Depression (ND).

	Overall F1-Scores		**Class F1-Scores**	
TwitSuicide	**Accuracy**	**F1 weighted**	**(SC)**	**(SI)**
BERT	55.15	54.25	33.96	61.49
RoBERTa	45.00	38.86	00.00	60.43
MentalBERT	63.33	63.29	48.00	71.23
MentalRoBERTa	45.75	44.02	24.46	53.22
Ours (EW2-TEC)	**71.86**	**71.03**	**52.64**	**78.03**
CSSRS	**Accuracy**	**F1 weighted**	**(A,B,I)**	**(UN)**
BERT	53.02	44.38	16.75	22.59
RoBERTa	28.66	25.86	00.00	23.38
MentalBERT	51.75	50.02	28.84	35.16
MentalRoBERTa	36.04	30.92	00.00	21.75
Ours (EWP1-EmoLex)	**73.07**	**70.79**	**43.82**	**72.71**
Depression	**Accuracy**	**F1 weighted**	**(D)**	**(ND)**
BERT	73.59	62.40	00.00	84.79
RoBERTa	73.59	62.40	00.00	84.79
MentalBERT	73.59	62.40	00.00	84.79
MentalRoBERTa	73.59	62.40	00.00	84.79
Ours (EWP2-SenticNet)	**78.16**	**76.20**	**48.51**	**86.13**

EW: word split; EWP: word piece; 1: simple cleaning; 2: added de-identification and emoticon preservation.

3.6.2 Ablation Results

To analyse which lexical components are beneficial for pre-training contextual emotional representations, we compare different embeddings based on the lexicon used to train them in Table 3.5. Twitter-based datasets achieve better performance when trained with TEC and SenticNet, which both include hashtags, emoticons, or emojis that are more frequently used on Twitter than on Reddit. These results underscores the importance of including these components in learning emotion representations especially for social media.

TABLE 3.5. Ablation study comparing the different emotion lexicons used for the multi-emotion classification pre-training task. The best scores are **bold faced**; the second best are underlined. Class-based scores are shown for the most and least concerning classes for each dataset. For TwitSuicide: Strongly Concerning (SC) and Safe to Ignore (SI). For CSSRS: weighted average of Attempt, Behaviour, and Ideation (A,B,I) and Uninformative (UN). For Depression: Depression (D) and Non-Depression (ND).

	Overall F1-Scores		**Class F1-Scores**	
TwitSuicide (EW2)	**Accuracy**	**F1 weighted**	**(SC)**	**(SI)**
EmoLex	67.97	65.26	28.06	75.96
TEC	**71.86**	**71.03**	**52.64**	**78.03**
SenticNet	70.12	68.80	44.09	76.84
CSSRS (EWP1)	**Accuracy**	**F1 weighted**	**(A,B,I)**	**(UN)**
EmoLex	**73.07**	**70.79**	**43.82**	**72.71**
TEC	72.34	69.79	41.54	72.09
SenticNet	70.07	67.41	37.86	71.14
Depression (EWP2)	**Accuracy**	**F1 weighted**	**(D)**	**(ND)**
EmoLex	77.56	**76.61**	**52.31**	85.33
TEC	77.64	**76.61**	49.40	85.61
SenticNet	**78.16**	76.20	48.51	**86.13**

EW: word split; EWP: word piece; 1: simple cleaning; 2: added de-identification and emoticon preservation.

We also compare the effect of different preprocessing methods (Section 3.3.1) to the mental health classification task. First, different tokenisation methods comparing simple word split tokenisation (EW) and word piece tokenisation (EWP). Second, a simple text cleaning method (1) versus the application of further de-identification and emoticon preservation (2). Table 3.6 shows that Twitter-based datasets perform better for de-identified and emoticon-preserved setups, possibly due to the frequent use of usernames, URLs, emoticons, and emojis on the platform. De-identification reduces noise during model training, while preserving emoticons as separate tokens contextualises them in the same way that words are contextualised with different meanings. Comparing tokenisation setups, both the CSSRS and the Depression datasets achieve better performance when wordpiece tokenised, while a simple word split is better for TwitSuicide. We note that during graph construction using the word split setup, TwitSuicide's vocabulary size is only 330, while Depression and CSSRS datasets have 1178 and 2673, respectively. The smaller vocabulary graph of TwitSuicide might have allowed it to

TABLE 3.6. Ablation study comparing the accuracy achieved using different text preprocessing setups. The best scores are **bold faced**; the second best are underlined.

Setup		TwitSuicide (TEC)		CSSRS (EmoLex)		Depression (SenticNet)	
		Accuracy	**F1weighted**	**Accuracy**	**F1weighted**	**Accuracy**	**F1weighted**
EW	1	69.24	67.01	70.30	66.99	76.06	66.46
EW	2	**71.86**	**71.03**	72.33	69.81	77.45	66.22
EWP	1	67.52	64.81	**73.07**	**70.79**	77.27	68.83
EWP	2	68.09	65.73	72.59	70.28	**78.16**	**76.20**

EW: word split; EWP: word piece;1: simple cleaning; 2: added de-identification and emoticon preservation

TABLE 3.7. Comparison of BERT and MentalBERT as the pre-trained embedding concatenated with MM-EMOG for mental health classification. The best scores are **bold faced**.

Dataset	**Setup**	**Lexicon**	**Embedding**	**Acccuracy**	**F1 weighted**
TwitSuicide	EW2	TEC	BERT	**71.86**	**71.03**
			MentalBERT	70.55	69.44
CSSRS	EWP1	EmoLex	BERT	**73.07**	**70.79**
			MentalBERT	72.02	69.47
Depression	EWP2	SenticNet	BERT	**78.16**	**76.20**
			MentalBERT	77.48	**76.20**

EW: word split; EWP: word piece;1: simple cleaning; 2: added de-identification and emoticon preservation

perform better on a simple word split setup. Longer and larger datasets benefit more from wordpiece tokenisation because of the deconstruction of out-of-vocabulary words.

Finally, we compared concatenating MM-EMOG embeddings with BERT or MentalBERT embeddings for the mental health classification task in Table 3.7. Using the best setup and lexicon for each dataset from the previous tests, BERT performs marginally better than MentalBERT despite MentalBERT being trained on mental-health-related data. Because of this, we retain the use of BERT embeddings for the rest of the experiments.

3.6.3 Qualitative Analysis

We further evaluate MM-EMOG with a qualitative assessment of the produced predictions. In Table 3.8, each sample is compared to the prediction of the two best-performing baseline

TABLE 3.8. Qualitative comparison of MM-EMOG predictions over the two best performing baseline models: BERT and MentalBERT. Parts of the examples are masked with *** to prevent a reverse search of each post. The **bold text** shows correct predictions by the model.

Example	Actual	Ours	BERT	MBERT
TwitSuicide				
i'm SO fucking tired i want to die. *** adrenal exhaustion *** since surgery, I've not been well ***	**SC**	**SC**	PC	PC
*** tired, *** foot hurts *** don't want to be here	**PC**	**PC**	SC	SC
*** victim of a failed suicide attempt *** I dont wet-shave my neck. Ouch	**SI**	**SI**	PC	SC
CSSRS				
Aannnnnnnd I failed... again. *** pills *** stomach Muscle cramp and Common cold chills...	**AT**	**AT**	SU	IN
*** VA hospital for three months *** awesome.	**BE**	**BE**	SU	BE
I know what you mean. I think about blowing my brains *** the immensely sweet relief *** constant Anxiety and Fear no longer exist. All of my issues will disappear, and thats all that matters. Why is suicide bad, again? *** why should I continue? ***	**ID**	**ID**	SU	SU
*** Im still sad that I had to go trough my life, sometimes bit angry to fate, *** nothing to show of my life. *** no longer bitter and *** that I was/am bad and deserved this.***	**IN**	**IN**	SU	ID
*** you didnt study the right way :) Things change *** so dont give up! I thought I wouldnt make it *** but then I changed majors ***	**SU**	**SU**	IN	UN
*** dressed in some of my finer casual *** made myself some coffee. *** today is better ***	**UN**	**UN**	SU	AT
Depression				
*** scares get re opened *** pooring salt in them. I hate this feeling. *** pain im in again	**D**	**D**	ND	ND
ine *** so revolting, yet so irresistible *** I must have it	**ND**	**ND**	**ND**	**ND**

models, BERT and MentalBERT. We note that for the *Ideation* (ID) class of CSSRS, our system distinguishes between simultaneous expression of support and ideation. Expressions of empathy such as "I know what you mean" and "I feel the same way" are frequently expressed in the *Supportive* (SU) class; however, these are directed toward situations that trigger negative emotions, such as having no one to talk to or being in an unpleasant environment. For the ID class, empathy is expressed towards hopelessness and self-harm. MM-EMOG captures emotional context that differentiates these better.

3.7 Ethical Considerations

While our work is mainly at a foundational research stage and not yet for production and deployment, we recognise that mental health classification using social media may be used to profile and disadvantage people with mental health issues in certain situations, such as employment and housing applications. However, we aim for the safeguarded use of any future healthcare application borne from this research, primarily for early detection and prevention of extreme outcomes of mental illnesses, such as self-harm and suicide. Two possible future applications are (1) for individual patient monitoring at the hands of mental health experts with proper patient consent or (2) for population-level monitoring for better mental health resource planning.

The use of publicly available data from social media comes with inherent risks, which we attempt to mitigate when conducting the study. First, we further de-identified each post in each dataset by replacing usernames, hyperlinks, and email addresses. Furthermore, we made it a point to mask published examples to prevent reverse searches leading back to the poster's account.

3.8 Limitations

We acknowledge three limitations of this study. First, we use mainly English-based datasets, lexicons, and baseline models. Low-resource languages were not explored in this study, but it is an open direction for the future. We also note that despite being marked as English, some posts may contain a mix of different languages. Second, the computational resource needed for building and training graph networks grows exponentially with the length and size of the datasets. We are limited by the resources available to us, which only allow a maximum of 256 words from each post. Lastly, there are not enough publicly available state-of-the-art models for single post-only, text-based mental health classification. Thus, we provide baselines based on widely used pre-trained language models.

3.9 Conclusion

Mental Illness Detection through individual social media posts is a challenging task due to limited information. Since mental health is deeply rooted in emotions, identifying all possible emotions within the text is crucial to enrich contextual representations further. We introduced MM-EMOG (**M**ulti-label **M**ental Health **Emo**tion **G**raph) representations, which contextualise and harmonise complex heterogeneous emotions through a corpus-based, multi-label pre-training framework. MM-EMOG representations are learned through a multi-label emotion classification task using GCN to leverage the global and local relationship of words and documents, followed by BERT for enhanced semantic contextualisation of each token embedding. We evaluate MM-EMOG through a graph-based mental health classification task. Our results show that MM-EMOG successfully outperforms baselines in three social media mental health datasets with notable improvements over the most concerning classes regardless of the lexicon used for pre-training. Tokenisation methods and further de-identification and emoticon preservation affect the datasets differently due to the different characteristics of the datasets.

This chapter, through the introduction of MM-EMOG, addresses the three research questions put forth by this thesis. First, it identifies the emotion modality as a complementary abstraction to textual semantics at the representation level, addressing the inherent affective neutrality in standard word representation spaces, especially for words lacking emotional connotations (RQ1). Next, it proposes a multi-emotion graph-based pretraining effectively incorporating complex, heterogeneous emotions to produce emotion-rich contextual representations (RQ2). Finally, the effectiveness of both at capturing nuanced implicit emotions is highlighted through depression and suicide ideation detection in social media textual posts, enabling early risk identification and intervention for enhanced mental health surveillance and support systems (RQ3).

CHAPTER 4

Multimodal Knowledge Distillation for Mental Health Classification

---

This chapter is the published work **3M-Health: Multimodal, Multi-Teacher Knowledge Distillation for Mental Health Detection** [32] published in **CIKM 2024**. I am the first author of this paper. I formulated the research aim, analysed datasets, established the methodology, implemented models, ran different experiments, and analysed results. I lead the writing of the manuscript and co-wrote most parts of the paper.

This work explores the addition of the acoustic modality as a proxy audio modality derived from the same textual data. This acoustic modality integrates derived paralinguistic and prosodic abstractions addressing the loss of affective tonality in textual data (Table 1.1) Similar to the emotion modality, it reorganises textual information to gain affective information through different inductive biases. Multi-teacher knowledge distillation integrates emotion, acoustic, and semantic modalities to enrich the textual representations learned by a smaller student model (Table 1.2), thereby improving mental health detection in social media.

The significance of mental health classification is paramount in contemporary society, where digital platforms serve as crucial sources for monitoring individuals' well-being. However, existing social media mental health datasets primarily consist of text-only samples, potentially limiting the efficacy of models trained on such data. Recognising that humans utilise cross-modal information to comprehend complex situations or issues, we present a novel approach to address the limitations of current methodologies. In this work, we introduce a **M**ultimodal and **M**ulti-Teacher Knowledge Distillation model for **M**ental **Health** Classification, leveraging insights from cross-modal human understanding. Unlike conventional approaches that often rely on simple concatenation to integrate diverse features, our model addresses the challenge of appropriately representing inputs of varying natures (e.g., texts and sounds). To mitigate

the computational complexity associated with integrating all features into a single model, we employ a multimodal and multi-teacher architecture. By distributing the learning process across multiple teachers, each specialising in a particular feature extraction aspect, we enhance the overall mental health classification performance. Through experimental validation, we demonstrate the efficacy of our model in achieving improved performance.[1]

4.1 Introduction

Mental health is a critical aspect of individual well-being, influencing both personal lives and societal structures [64]. Despite advancements in mental healthcare, not everyone with mental health concerns actively seeks professional help. The widespread use of social media platforms, such as Twitter and Reddit, has opened avenues for detecting mental health issues by analysing text-oriented posts. This shift towards online expression has prompted research into text-based mental health classification, focusing on identifying the presence and categories of mental health concerns within social media posts. Recent studies in mental health classification from social media content have embraced diverse components, ranging from historical posts and conversation trees to social graphs and user metadata [40, 39, 128, 199, 202]. However, the availability of these additional sources varies due to data privacy restrictions or user preferences, introducing challenges in research and system reproducibility.

In light of these challenges, our research addresses the limitations of existing methodologies by focusing on the analysis of text-only social media posts, a fundamental and universally available component. While semantic pre-trained textual embedding from text-only input may capture explicit emotional words related to mental health, they may fall short in capturing less explicit emotions, limiting their robustness. For instance, some textual posts may lack explicit emotional language yet imply an unhealthy mental state.

Recognising the potential shortcomings of text-only datasets, we introduce a novel approach to mental health classification through a Multimodal and Multi-Teacher Knowledge Distillation Model. Inspired by human comprehension strategies that involve multimodal information

[1] Code available at https://github.com/adlnlp/3mhealth

integration, our model leverages insights from multimodal human understanding to enhance the efficacy of mental health risk detection.

Our approach introduces a new acoustic modality feature generated from original textual posts, motivated by the proven effectiveness of vocal biomarkers in indicating psychological distress and other medical conditions [93]. This would derive a new modality from text-only input for unimodal text-based mental health risk detection. Simultaneously, we also incorporate emotion-enriched features as additional information. Instead of integrating all modalities into one model, we employ a multimodal and multi-teacher architecture to address the computational complexity of integrating diverse features. This approach distributes the learning process across multiple teachers, each specialising in a particular feature extraction aspect. To the best of our humble knowledge, there have been no attempts to create a new media-based modality from text-only input for the unimodal text-based mental health risk detection tasks. Additionally, we propose a new multimodal knowledge distillation model for the mental health risk detection domain.

4.2 Related Works

4.2.1 Mental Health Classification

Recent studies in mental health classification from social media content have incorporated diverse social media components. These components encompass various elements, including historical posts, conversation trees, social and interaction graphs, user or post metadata information, and profile pictures or posted images [40, 39, 128, 199, 202]. However, these additional components are not always be available or accessible due to data privacy restrictions or user preferences. This complicates research reproducibility since each study selects features based on what social media components are available to them.

Our research focuses on exploring mental health detection by analysing only social media textual posts, which is a compulsory component of text-based social media posts related to mental health. Based on the textual aspect, existing studies have worked on frequency- or

score-based emotion features [14, 283]. More recent works fine-tuned contextual embeddings on emotion-based tasks to use as emotion features [113, 199]. These studies mainly focus on identifying and matching one type of emotion to each word or entire textual content. On the other hand, this research builds on the study from the previous chapter highlighting the complexity of human emotions wherein a single word could be associated with multiple types of emotions. We integrate the emotion-enriched features generated through multi-label, corpus-based representation pretraining.

In addition, motivated by the effectiveness of vocal biomarkers in psychological distress and other medical condition indications [93], we propose a novel way to include acoustic features generated by original textual posts in this task. These simultaneously processes the emotion-enriched text and audio features as additional information. To the best of our knowledge, there have been no attempts to create a new modality from the text-only input to achieve unimodal text-based mental health risk detection tasks.

4.2.2 Multi-teacher Knowledge Distillation

Other multimodal social media-based mental health detection studies mainly integrate modalities through the concatenation of features [81, 30] or a joint encoder [11]. However, a simple concatenation of features does not represent the unique nature of each modality properly. To integrate the multimodal knowledge efficiently, we design our model using knowledge distillation to compress a complex and large multimodal integration model into a smaller and simpler one while still retaining the accuracy and performance of the resultant model.

Knowledge distillation [87] involves transferring knowledge from a teacher model to a student model, commonly applied to compress large models by mapping intermediate layer outputs [42, 100] or minimising KL divergence in class distribution [157]. Traditionally, knowledge distillation is used within the same modality. However, recent approaches extend it to different modalities [263, 169, 120]. Some studies explore collaborative learning with multiple teachers for improved compression, such as in language [253] and vision [181] models. Other applications include multilingual language translation [216] and multi-teacher

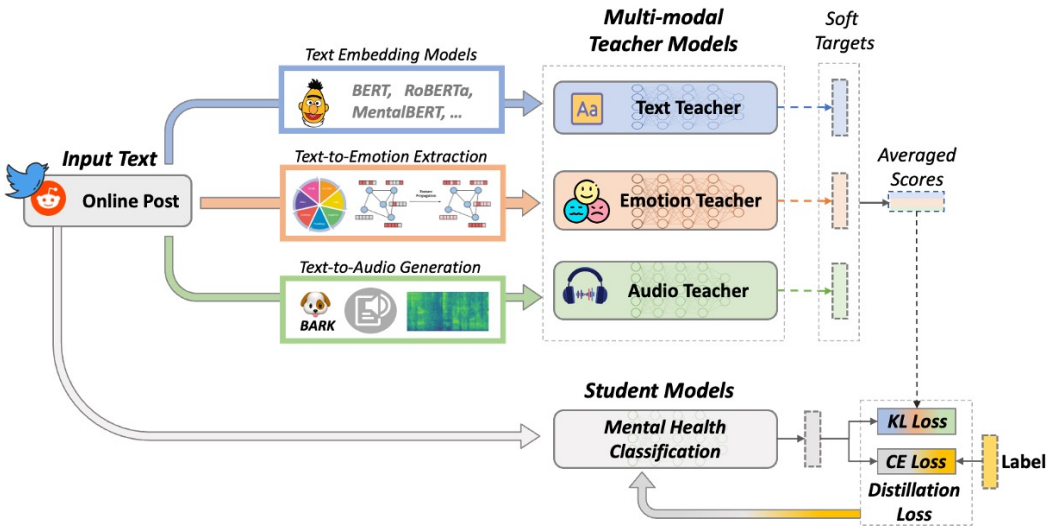


FIGURE 4.1. Architecture of 3M-Health consisting of three modality-based teachers distilling knowledge to a text-only student model.

knowledge distillation to unify classifiers trained on distinct data sources [231]. Inspired by this, we propose a new Multimodal Multi-Teacher Knowledge Distillation framework for mental health risk detection.

4.3 3M-Health

In this section, we introduce our **M**ultimodal **M**ulti-teacher knowledge distillation model for **M**ental **Health** detection, **3M-Health**. Figure 4.1 illustrates the overall architecture. This model consists of three distinct teacher models, each focusing on different modalities to independently learn diverse aspects of features crucial for interpreting mental health-related posts. The acquired features from multimodal teachers serve as a valuable source of knowledge for instructing the student model by utilising the average output distribution of the teacher models as soft targets. We introduce three essential multimodal teacher models for mental health risk detection, including 1) a text-based teacher for understanding semantic aspects from input texts, 2) an emotion-based teacher for interpreting emotion aspects from input texts, and 3) an audio-based teacher for discerning emotions conveyed through audio sounds.

4.3.1 Multimodal Multi-Teacher Construction

This section articulates each teacher model's objective and construction process. Teacher model fine-tuning is detailed in Section 4.3.2.

4.3.1.1 Text-based Teacher

The text-based teacher aims to teach contextual semantic comprehension of mental health-related textual posts. We leverage pre-trained large language models (PLMs) since contextualised embeddings from PLMs represent different meanings based on the context (e.g. *blue* means *a kind of colour*, but *gloomy* in other emotional contexts). More specifically, some words may have opposite meanings in the medical domain (e.g., *positive* usually means something good but often refers to the presence of a specific condition, which is typically not a desirable outcome). Inspired by this, we explore several general and domain-specific PLMs. For general PLMs, we compare BERT [56] and RoBERTa [135]. For health-specific PLMs, MentalBERT [98] and ClinicalBERT [236]. More information regarding these PLMs may be found in Table 2.2.

4.3.1.2 Emotion-based Teacher

The emotion-based teacher aims to represent affective abstractions from the input text of mental health-related posts. It is initialised with emotion-rich contextualised representations extracted from the multi-emotion, graph-based model proposed in Chapter 3.

As a brief recap, we first obtain a multi-label emotion class indicating *anger*, *disgust*, *fear*, *sadness*, *surprise*, *negative*, and *other*[2] for each post using the SenticNet7 lexicon [37][3], mapping identified words to their corresponding emotion types. This emotion lexicon consists of terms $K = \{k_1, ..., k_q\}$ associated with one or more emotion types from $EM = \{em_1, ..., em_r\}$. For each word $W = \{w_1, ..., w_p\}$ in a post, we assign EM_{k_j} to w_i whenever $w_i = k_j$

[2]Positive sentiment/emotions are grouped into *other* to focus on the different negative emotion on mental health-related text.

[3]https://sentic.net/downloads/

in K in this document. Consequently, each post is associated with a multi-label class $EM_d = \{EM_{w_1} \cup EM_{w_2} \cup ...EM_{w_p}\}$.

We construct a graph $G = (V, E, A)$ representing all posts and their word tokens, where V is the set of all post nodes and token nodes tokenised through wordpiece tokenisation with emoticon preservation[4]. Here, E encompasses token-token edges E_{w_i,w_j}, token-post edges E_{w_i,d_j}, and post-post edges E_{d_i,d_j}, while A specifies weights between related nodes [82]. Post node and token node representations are initialised with the [CLS] embedding and the minimum of contextualised token embeddings from pre-trained BERT word embeddings, respectively. Edge values are determined by Pointwise Mutual Information (PMI) for E_{w_i,w_j}, Term Frequency-Inverse Document Frequency (TF-IDF) for E_{w_i,d_j}, and Jaccard similarity for E_{d_i,d_j}. Utilising these initialised representations and edge values, a two-layer Graph Convolutional Neural Network [110] (GCN) is trained with ReLu for the multi-label emotion classification task based on the SenticNet7 lexicon. The updated second-layer hidden states are extracted and used as initial weights for fine-tuning BERT on the same multi-label emotion classification task to further comprehend the associated emotions in the posts. The updated word embeddings are extracted as the multi-emotion contextual embeddings which is used to initialise the emotion teacher model.

4.3.1.3 Audio-based Teacher

Most publicly available social media mental health datasets primarily consist of text-only samples, which motivates the initial two teacher models to be based purely on textual information. To address the need for a more comprehensive understanding of complex mental health and emotional contexts, we propose integrating multimodal information derived from textual posts. According to research, individuals can more accurately interpret the emotions of others through listening rather than observing facial expressions/body language or reading written text [46]. Drawing inspiration from this insight, we introduce an audio-based teacher to enhance knowledge distillation, enabling the interpretation of emotions in mental health posts through affective tonal cues.

[4]To further preserve and integrate emotions in the posts, emoticons and emojis are added to the tokeniser vocabulary.

To achieve this, we first employ Bark[5], a pre-trained transformer-based text-to-audio model, to generate corresponding audio for each post as Bark can capture emotional sounds detected from the text (e.g. *[laughs]*, *[gasps]* and *"..."* for hesitations)[6]. Note that Bark can generate only 13 seconds of audio. Hence, we tokenise each post at the sentence level, generate audio for each sentence, and then aggregate these audio segments into a complete audio representation for the entire post. Particularly long sentences or texts that lack punctuation are further split into segments with a maximum of 45 tokens.

4.3.2 Multimodal Multi-Teacher Fine-tuning

Researchers has emphasised the significance of fine-tuning teacher models for effective student instruction [100, 253]. In this section, we describe the independent fine-tuning process of each teacher model.

Text-based teacher

We fine-tuned pre-trained language models for the mental health classification task with labels $C = \{c_1, c_2, ..., c_{|C|}\}$ to build the text-based teachers. This process enables the teacher to learn the nuances of mental health-related contexts within each dataset.

Emotion-based teacher

For the emotion-based teacher, following the generation of emotion-rich representations for each post and its words, these embeddings serve as inputs for fine-tuning a Multi-Layer Perceptron (MLP) for mental health classification, operating over the same mental health labels C.

Audio-based teacher

We employ the Audio Spectrogram Transformer [72] (AST) for the audio-based teacher to classify each generated audio into mental health risk classes. AST is a transformer-based model that takes a sequence of audio spectrogram patches as inputs. An audio waveform is first converted into a 128x100t spectrogram based on a sequence of 128-dimensional log

[5] https://github.com/suno-ai/bark

[6] A theoretical and practical comparison of text-to-audio generation APIs and a list of Bark's sound cues in Section 4.4.2.

Mel filterbank features computed using a 25ms Hamming window every 10ms. Such a spectrogram is then split into a sequence of 16x16 patches of images with an overlap of six in both time and frequency dimensions. A special token [CLS] is added to the beginning of the sequence of spectrogram patches. After passing through transformer encoder layers, the [CLS] embedding is fed into a linear layer with sigmoid activation to classify mental health risk class labels C.

Each teacher model is individually constructed and fine-tuned to facilitate optimal learning. We are aware of the concerns raised by some researchers highlighting the potential inconsistency in the feature space when different teachers are separately pre-trained with distinct settings and then fine-tuned independently [253]. Based on our initial testing, co-finetuning multimodal teachers yields little improvement especially when distilling knowledge through soft labels from the prediction layer; in fact, it tends to result in lower performance. We speculate that integrating heterogenous representation spaces from multimodal information with different inductive biases results in representation collapse and may not perform optimally during co-finetuning.

4.3.3 Multi-Teacher Knowledge Distillation

For the student model, we use a single modality involving textual posts as input for a pre-trained BERT, which processes the sequence of tokenised words. The student model is trained through the same mental health risk classification task over the same class labels C using knowledge distilled from the text-based, emotion-based, and audio-based teacher models. To incorporate the acquired knowledge from these various multimodal sources, the student model is trained to minimise the distillation loss given by $L = L_{task} + L_{kd}$. Here, L_{task} represents the cross-entropy loss between the student model's predictions and the ground truth of mental health risk categories, while L_{kd} stands for the Kullback-Leibler (KL) divergence between the student model predictions and the soft targets from the teacher models' predictions. Given the presence of multiple teacher models, as a preliminary evaluation, we calculate L_{kd} by averaging the predicted probability distributions from all three teacher models.

TABLE 4.1. Data statistics. Durations are in a minute:second (mm:ss) format.

	TwitSuicide	**DEPTWEET**	**IdenDep**	**SDCNL**
Task	Suicide	Depression	Depression	Suicide/Depression$^{\pm}$
Platform	Twitter	Twitter	Reddit	Reddit
Num. Classes	3	4	2	2
Total Samples	660	5128	1841	1895
Evaluation	10-fold	60/20/20	10-fold	80/20
Train/Val	-	4,102	-	1,516
Test	-	1,026	-	379
Length	13-147	1-926	11-17,641	13-24,590
Avg. Length	90.32	163.28	1,127.57	936.76
Words	3-31	101	3,477	4,411
Avg. Words	16.85	28.15	215.1	178.53
Min Duration	00:01.903	00:01.250	00:02.900	00:02.463
Max Duration	00:32.853	00:56.596	22:07.740	28:09.546
Avg. Duration	00:11.545	00:17.860	01:45.193	01.27.568

$^{\pm}$SDCNL distinguishes between suicide and depression-related posts.

4.4 Experimental Setup

4.4.1 Datasets

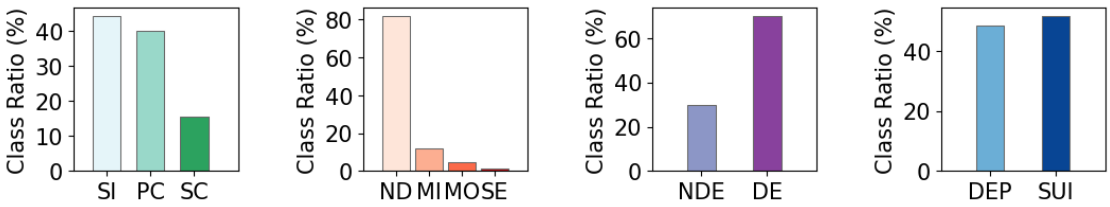
We evaluate our proposed model using four publicly available datasets related to mental health on social media. Table 4.1 and Figure 4.2 provide a summary of statistics and class distribution.

The **TwitSuicide Dataset**[7] [138] replicates the data collection, processing, and annotation methods of Odea et al [171]. A sample of 660 tweets is annotated into three risk levels. The *Strongly Concerning* (SC) class is assigned to posts with a convincing display of severe suicidal ideation, while *Safe to Ignore* (SI) shows no evidence of suicide risk. If it doesn't fall into other categories, a post remains in the *Possibly Concerning* (PC) class.

DEPTWEET[8] [103] is collected from Twitter using seed terms based on the Patient Health Questionnaire (PHQ-9). The dataset comprises 40,191 tweets; however, only 5,128 tweets

[7]Data available upon request.

[8]https://github.com/mohsinulkabir14/DEPTWEET



(A) **TwitSuicide** (B) **DEPTWEET** (C) **IdenDep** (D) **SDCNL**

FIGURE 4.2. Class distribution. For (a) TwitSuicide, SI: Safe to Ignore; PC: Possibly Concerning; SC: Strongly Concerning. For (b) DEPTWEET, ND: Non-depression; MI: Mild; MO: Moderate; SE: Severe. For (c) IdenDep, NDE: Non-depression; DE: Depression. For (d) SDCNL, DEP: Depression; SUI: Suicide.

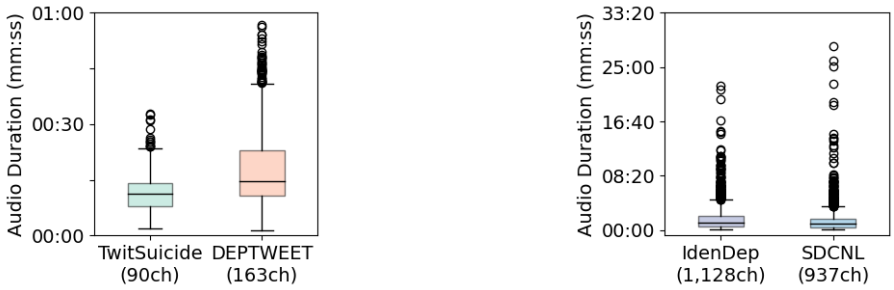
were retrieved during this study. The labels include *Non-Depressed* (ND), *Mild* (MI), *Moderate* (MO), and *Severe* (SE), maintaining an imbalanced class distribution, with around 80% labelled as ND and less than 2% SE.

The **Identifying Depression Dataset**[9] (IdenDep) [182] consists of 1,841 Reddit posts, with "depression indicative" (DE) posts sourced from the Depression subreddit and non-depressive (NDE) posts from the "family" and "friendship advice" subreddits. No further manual check was done on the samples, increasing the probability of false negatives.

The **SDCNL Dataset**[10] [84] involves distinguishing between Reddit suicide-related and depression-related posts. The dataset contains 1,895 nearly balanced posts labelled as *Suicide* (SUI) or *Depression/Not Suicide* (DEP) based on their subreddit. In accordance with ethical principles outlined by Benton et al. [24], all posts are de-identified before any analysis, audio generation, and model training.

We provide a detailed breakdown of text and audio statistics in Tables 4.2 and 4.3 to provide more information regarding the nature of each class, which may influence model learning and performance. Notably, DEPTWEET and IdenDep datasets have highly skewed data, with 82.16% and 70.23% on a single class, respectively. Figure 4.3 illustrates the differences in the generated audio in terms of duration. The Reddit-based datasets, IdenDep and SDCNL,

[9] https://github.com/Inusette/Identifying-depression
[10] https://github.com/ayaanzhaque/SDCNL



(A) Twitter-based (B) Reddit-based

FIGURE 4.3. Audio length comparison. ch: character average

TABLE 4.2. Text statistics for each class per dataset.

Class	Total	%	Length		Tokens	
			Range	Ave.	Range	Ave.
TwitSuicide						
Safe to Ignore (SI)	103	15.61	13-139	77.89	4-31	15.25
Possibly Concerning (PC)	264	40.00	24-147	88.16	4-31	16.35
Strongly Concerning (SC)	293	44.39	13-147	96.65	3-30	17.86
DEPTWEET						
Non-Depressed (ND)	4213	82.16	1-816	164.47	1-101	28.08
Mild (MI)	606	11.82	4-885	144.74	1-87	26.38
Moderate (MO)	232	4.52	32-926	184.95	5-99	33.25
Severe (SE)	77	1.50	23-398	178.81	1-62	30.57
IdenDep						
Non-Depression (NDE)	548	29.77	11-17641	1546.34	1-3477	295.75
Depression (DE)	1293	70.23	11-13803	950.09	2-2487	180.92
SDCNL						
Depression (DEP)	915	48.28	43-16015	1000.68	8-3200	192.84
Suicide (SUI)	980	51.72	13-24590	977.07	2-4411	165.16

are significantly longer than the Twitter-based datasets, possibly providing more auditory information inferred from the textual posts.

TABLE 4.3. Audio statistics for each class per dataset in a minute:second (mm:ss) format.

Dataset	Class	Min Duration	Max Duration	Ave. Duration
TwitSuicide	Safe to Ignore (SI)	00:01.903	00:31.000	00:12.215
	Possibly Concerning (PC)	00:02.143	00:32.853	00:11.393
	Strongly Concerning (SC)	00:01.943	00:24.760	00:10.280
DEPTWEEET	Non-Depressed (ND)	00:01.250	00:56.200	00:17.210
	Mild (MI)	00:02.230	00:56.596	00:15.413
	Moderate (MO)	00:04.460	00:47.770	00:18.958
	Severe (SE)	00:02.250	00:40.160	00:17.865
IdenDep	Non-Depression (NDE)	00:02.900	22:07.740	02:20.941
	Depression (DE)	00:03.100	21:28.643	01:30.420
SDCNL	Depression (DEP)	00:05.756	25:58.966	01:33.802
	Suicide (SUI)	00:02.463	28.09.546	01:21.747

4.4.2 Text-to-Audio Generators

In order to generate the best possible audio to represent each textual post in our benchmark datasets, we performed a theoretical and practical comparison between five publicly accessible text-to-speech and text-to-audio generative APIs.

(1) **Tacotron2**[11] [207] uses a recurrent neural network architecture to predict mel spectrogram sequences from text followed by a modified WaveNet vocoder.

(2) **SpeechT5**[12] [12] unifies modalities with a shared encoder-decoder architecture that uses cross-modal vector quantisation for speech and text alignment.

(3) **SpeechBrain**[13] [186] is a speech toolkit offering various speech-related tasks. Their text-to-speech model is based on Tacotron2 but is trained further on the LJSpeech [92] and LibriTTS [269] datasets.

[11] https://github.com/NVIDIA/tacotron2
[12] https://github.com/microsoft/SpeechT5
[13] https://github.com/speechbrain/speechbrain/

(4) **Balacoon**[14] packages offer lightweight and fast text analysis and speech generation, going against larger but slower TTS models. It sacrifices multi-speaker and multilingual features for lightning-fast speed on the CPU. The detailed model architecture was not publicly available at the time of this paper's writing.

(5) **Bark**[15] has a GPT-based architecture using a quantised audio representation that does not require the use of phonemes, allowing it to generalise beyond speech, thus making it a text-to-audio model more than a TTS model.

Upon comparison of the five generators, we use Bark due to the expressiveness of the audio generated by the model. While the other models suffer from a robotic delivery of the generated speech, not verbalising numerical figures, and reading of emoticons as individual punctuations (e.g. ">:|" as *greater than, semicolon, pipeline*), Bark produces the most naturally sounding audio recognising textual markers like "," for pauses, "–" and "..." for hesitations, capitalisation for emphasis (e.g. *goodbye* vs. *GOODBYE*), and sentence punctuations for producing tonal shifts (e.g. *huh?* vs *huh!*). Bark also verbalises non-speech sounds such as [laughter], [laughs], [sighs], [music], [gasps], [clears throat], *haha*, *uhm*, *waaah*, and *ooh*. Bark's ability to infer and convey emotions from an input text would be valuable to our mental health risk detection model, as it can provide additional emotional cues from the generated sound.

4.4.3 Baselines and Metrics

We assess the performance of our model by comparing it to previously published results using post-only[16] and post-level classification on the same datasets, employing identical class labels and similar evaluation setups. We use results reported in the following studies as our baselines: **Bi-LSTM Char+Word** [138] for TwitSuicide; **MLP** [215] and **EAN** [190] for IdenDep; and **GUSE-DENSE** [84] and **AugBERT+LR** [10] for SDCNL. For the DEPTWEET dataset, we use the published **DistilBERT** code from [103] to replicate baseline results for the retrieved dataset. In addition, we provide strong baselines from fine-tuning state-of-the-art PLMs: **BERT** [56], **RoBERTa** [135], **MentalBERT**, and **MentalRoBERTa** [98]. More details of

[14] https://huggingface.co/balacoon/tts
[15] https://github.com/suno-ai/bark
[16] In contrast to studies incorporating other components such as posted images or user network and activity.

TABLE 4.4. Teacher model hyperparameter search space and best found parameters.

Parameters	Search Space	TwitSuicide	DEPTWEET	IdenDep	SDCNL
Text-based Teachers					
Language Model	BERT, RoBERTa, Mental BERT, ClinicalBERT	BERT	MentalBERT	BERT	BERT
Dropout	{0.01, 0.05, 0.1, 0.5}	0.01	0.01	0.05	0.01
Weight Decay	{0, 0.01, 0.1}	0	0	0	0
Learning Rate	{1e-04, 1e-05, 2e-05, 3e-05, 4e-05, 5e-05}	4e-05	4e-05	4e-05	5e-05
Epochs	[2-5]	4	3	5	4
Hidden Layers	{2, 4, 6, 8, 10, 12}	10	8	2	4
Attention Heads	{2, 4, 6, 8, 12}	8	8	6	12
Batch Size	128	64	64	64	
Emotion-based Teachers					
Dropout	{0.01, 0.05, 0.1, 0.5}	0.1	0.1	0.05	0.01
Weight Decay	{0, 0.01, 0.1}	0	0.01	0.1	0
Learning Rate	{1e-03, 1e-04, 1e-05}	1e-05	1e-04	1e-04	1e-05
Hidden Layers	[2-5]	2	4	3	5
Hidden Dim	{100, 200, 300, 400, 500}	400	400	100	100
Batch Size	{64, 124}	128	64	64	64
Audio-based Teachers					
Dropout	{0.01, 0.05, 0.1, 0.5}	0.1	0.1	0.01	0.01
Learning Rate	{1e-03, 1e-04, 1e-05, 5e-05}	5e-05	1e-05	1e-05	5e-05
Hidden Layers	{2, 4, 6, 8, 10, 12}	8	6	6	8
Attention Heads	{2, 4, 6, 8, 12}	4	6	8	8
Scheduler Patience	[2-5]	4	4	3	4
Scheduler Factor	0.1, 0.5	0.5	0.5	0.5	0.1
Batch Size	32, 64	32	32	32	32

which may be found on Table 2.2. All PLM baselines follow the training setup used by [138] with a batch size of 8 and a learning rate of 1e-04 trained for three epochs. Given the class imbalance, we evaluate our system based on macro F1 (F1m) and weighted F1 (F1w) scores, followed by accuracy and class F1 scores.

4.4.4 Implementation Details

We evaluate our model following established evaluation setups from previous literature using the same datasets on the same classification task setup for fair benchmark comparisons. We use 10-fold cross-validation for TwitSuicide and IdenDep, while a train/test split is used for

TABLE 4.5. Student hyperparameter search space and best found parameters for different combinations of teacher modalities.

Parameters	Search Space	TwitSuicide	DEPTWEET	IdenDep	SDCNL
Text ✓Emo ✓Aud ✓					
Dropout	{0.01, 0.05, 0.1, 0.5}	0.05	0.01	0.1	0.05
Learning Rate	{1e-04, 1e-05, 2e-05, 3e-05, 4e-05, 5e-05}	1e-04	3e-05	5e-05	5e-05
Weight Decay	{0, 0.01, 0.1}	0	0	0	0.01
Hidden Layers	{2, 4, 6, 8, 10, 12}	10	6	10	12
Attention Heads	{2, 4, 6, 8, 12}	3	12	8	12
Activation	{“relu”, “gelu”}	gelu	gelu	gelu	gelu
Epochs	[3-5]	3	5	3	3
Text ✓Emo ✓Aud ×					
Dropout	{0.01, 0.05, 0.1, 0.5}	0.1	0.01	0.1	0.05
Learning Rate	{1e-04, 1e-05, 2e-05, 3e-05, 4e-05, 5e-05}	1e-04	3e-05	5e-05	4e-05
Weight Decay	{0, 0.01, 0.1}	0.01	0	0	0
Hidden Layers	{2, 4, 6, 8, 10, 12}	10	10	10	12
Attention Heads	{2, 4, 6, 8, 12}	12	12	6	3
Activation	{“relu”, “gelu”}	gelu	gelu	gelu	gelu
Epochs	[3-5]	4	5	3	5
Text ✓Emo × Aud ✓					
Dropout	{0.01, 0.05, 0.1, 0.5}	0.05	0.05	0.05	0.05
Learning Rate	{1e-04, 1e-05, 2e-05, 3e-05, 4e-05, 5e-05}	1e-04	1e-04	4e-05	5e-05
Weight Decay	{0, 0.01, 0.1}	0	0	0	0
Hidden Layers	{2, 4, 6, 8, 10, 12}	12	4	12	12
Attention Heads	{2, 4, 6, 8, 12}	4	8	12	12
Activation	{“relu”, “gelu”}	gelu	relu	gelu	gelu
Epochs	[3-5]	4	3	4	5
Text ✓Emo × Aud ×					
Dropout	{0.01, 0.05, 0.1, 0.5}	0.1	0.01	0.05	0.05
Learning Rate	{1e-04, 1e-05, 2e-05, 3e-05, 4e-05, 5e-05}	1e-04	4e-05	1e-4	4e-05
Weight Decay	{0, 0.01, 0.1}	0	0	0	0
Hidden Layers	{2, 4, 6, 8, 10, 12}	12	12	4	10
Attention Heads	{2, 4, 6, 8, 12}	8	6	4	12
Activation	{“relu”, “gelu”}	relu	gelu	gelu	relu
Epochs	[3-5]	3	4	5	4

SDCNL and DEPTWEET. Original data splits are retained when provided; otherwise, the data is randomly split (Table 4.1). When no split is established, 10% of the training set is used for validation.

Hyperparameter tuning is done per dataset and model setup using Optuna[17] for 50 trials optimising weighted F1 scores. Table 4.4 and 4.5 enumerates the best-found hyperparameters for the final teacher models and student models, respectively. Text-based teachers are trained using ReLu and a max length of 256. Audio-based teachers are trained for 25 epochs with a 5-epoch early stop, a 512 max length, and using the ReduceLROnPlateau scheduler. All inputs for the AST model are normalised to zero mean and 0.5 standard deviation. Emotion-based teachers are trained for 100 epochs with a 10-epoch early stop and a 256 max length. The student models are tuned and trained with distilled knowledge from the fine-tuned teachers using a max length of 256. All tuning was done using a 90:10 split and was conducted separately from the final model construction. All models are trained using an Adam optimiser on an NVIDIA TITAN RTX machine.

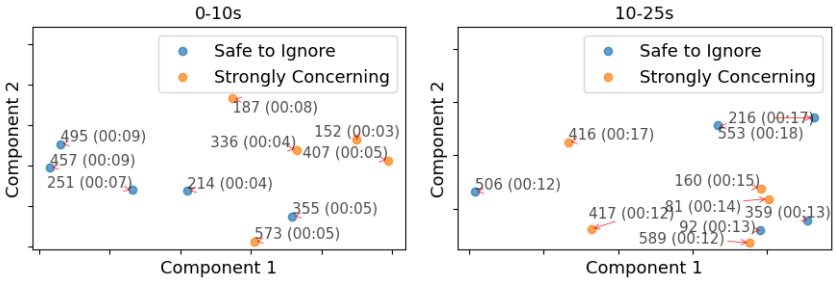
4.5 Audio Representation Analysis

To ensure the feasibility of our audio modality for mental health detection, we give an illustrative visualisation of the audio embeddings, which are generated by input text and learned via the Audio Spectrogram Transformer (AST). We conduct Principal Component Analysis (PCA) to visualise the acquired audio embeddings and their corresponding mental health class labels. In order to emphasise the distinguishability of the embeddings, we select samples from both the least and most concerning labels in each dataset, as shown in Figure 4.4. For each dataset, we group all the generated audio based on durations of 0-to-10-second and 10-to-25-second length[18]. For each of these two audio groups, we generated the corresponding spectrograms and randomly selected ten audio samples for each group to visualise the first two principal components after performing PCA.

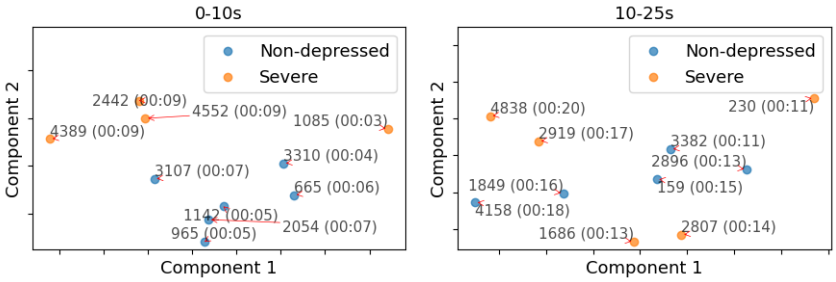
In Figure 4.4, we annotate each sample with the post ID and the audio duration for detailed comparison. Table 4.6 and Table 4.7 contain de-identified and masked post contents for DEPTWEET and SDCNL, respectively. Samples for TwitSuicide and IdenDep may be found in Appendix B. Following the proposed ethical protocols on social media research of [24],

[17] https://optuna.org/

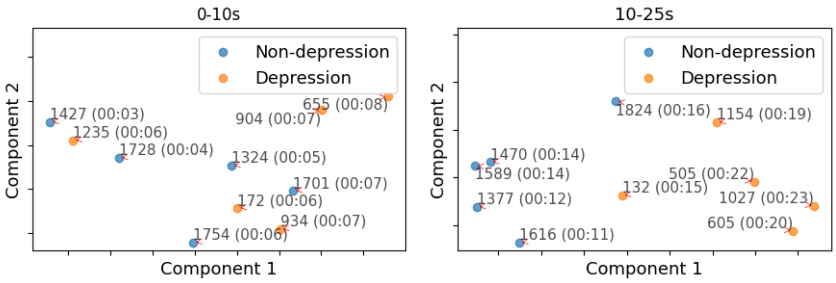
[18] Note that most generated audios are less than 25 seconds.



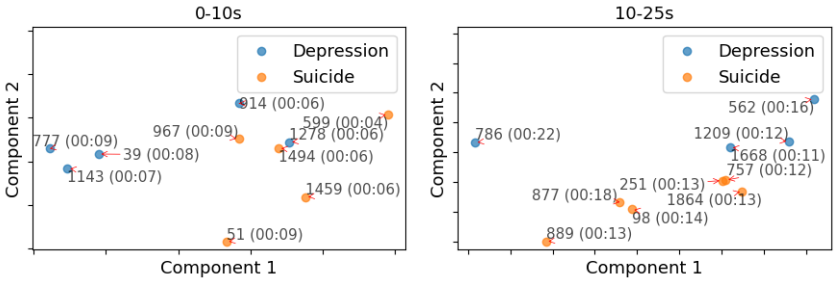
(A) TwitSuicide



(B) DEPTWEET



(C) IdenDep



(D) SDCNL

FIGURE 4.4. Audio analysis using PCA on spectrogram images of audio samples grouped by a maximum of 10s (left) and 10-25s (right). Each sample is labelled with an ID for reference to corresponding texts provided in the Supplementary Material.

usernames and links are masked with special tokens *_USER_* and *_URL_* to protect the identity and privacy of each author. For instance, "@myusername this is the link http://urlsamp.le" is

TABLE 4.6. Samples for the DEPTWEET audio spectrogram analysis. Each sample has been masked to avoid a reverse search of each post. ND: Non-Depressed; SE: Severe.

ID	Class	Text
0-10 seconds		
665	ND	*** miss my sc I'm so depressed without it
965	ND	_USER_ _USER_ _USER_ Frustrated *** fan hai _URL_
1142	ND	Me checking *** I hate *** Continues to check *** and then gets depressed
2054	ND	*** so exhausted *** fighting to stay up until 8pm
3107	ND	Do you feel frustrated *** on the simplest things? _USER_ ...
3310	ND	*** teacher is so tired of *** shit
1085	SE	_USER_ sh000t me it would hurt less ***
2442	SE	*** so lonely. *** going to hurt someone . #depression _USER_
4389	SE	We *** the shit country. *** so depressed. _URL_
4552	SE	*** no reason to live. *** I'll just end it . #depression _USER_
10-25 seconds		
159	ND	_USER_ Neither. *** people who *** clinically depressed are going to be so regardless of their worldview. IMHO
1849	ND	ofc *** days off im taking care of my newphew ... im so tired i work every weekday *** takes forever to get home *** on my days off i babysit... i don't even get paid. im exhausted
2896	ND	Football: *** revive World Cup hopes, *** frustrated by *** _URL_
3382	ND	*** sad of getting old it made us restless... *** so MAD i'm getting old it makes me reckless!!!
4158	ND	I've *** my toenails off and split the nail bed - the pain has progressed over *** days to absolutely excruciating - so bad *** struggling to even walk. This week is going amazing
230	SE	*** first guest: Me. *** self-sabotage and self-destruction.
1686	SE	_USER_ Man, September was so hard *** watched my gma pass away, *** so much other stuff went wrong. I been depressed asf
2807	SE	_USER_ _USER_ _USER_ _USER_ _USER_ I personally can't *** 3 or4 died *** from either trauma or anxiety and *** those who took their own lives because of what happened
2919	SE	*** get the hell out. so I'll just end it . #depression _USER_
4838	SE	*** thinking about suicide more and more *** I don't want to. I don't want *** that trauma on my kid. But it's hard… *** suffering from depression *** 15 years… it's a daily battle… I'm tired

masked as "_USER_ this is the link _URL_". Moreover, we mask parts of the text with "***" to prevent possible reverse searches.

The visualisation shows that our audio embeddings can show a noticeable separation between mental health classes for all four datasets. In datasets derived from Twitter, shorter audio samples display more pronounced distinctions between the most and least concerning classes, whereas, in datasets from Reddit, this separation becomes more evident in longer audio

TABLE 4.7. Samples for the SDCNL audio spectrogram analysis. Each sample has been masked to avoid a reverse search of each post. DEP: Depression; SUI: Suicide

SDCNL		
ID	**Class**	**Text**
0-10 seconds		
39	DEP	*** good qualities in therapist that i can spot early on? ***
777	DEP	I'm *** okay I probably won't, but tonight *** gotta tell myself that. *** until sunrise
914	DEP	Another attempt ***. Hopefully I don't survive
1143	DEP	*** tired of being told it gets better *** never has and never *** will
1278	DEP	How do i know if i have depression *** i just wanna know *** i'm getting close
51	SUI	I'm scared *** alone *** don't know what to do ***
599	SUI	Gave my note to my family *** don't know what to do now
967	SUI	Why do people think they can help *** on ending it all? ***
1459	SUI	*** sleep forever and never want to wakeup again..... ***
1494	SUI	Picking up a gun *** relieved
10-25 seconds		
562	DEP	Songs? *** songs/artist *** feel better/happy when you feel depressed?
786	DEP	*** freaking out constantly *** so lonely Even when I'm around people I'm lonely. I feel crazy. I can't stop looking *** at my awful life, *** such poor judgment
1209	DEP	Anyone want to chat? I'm not in a good place *** Just wondering *** find comfort in each other.
1668	DEP	Any good songs *** to recommend? Any song *** which makes you feel good. Have a good day!
98	SUI	Suicide *** 4 ways I can go... *** overdose *** cop *** bridge *** train
251	SUI	Called off my work , going to end it *** Nice knowing you all I'm finally coming home *** and ***
757	SUI	*** want to have someone. It's all I've *** wanted *** keeps eating away at me and won't change. *** don't want to be alone anymore.
877	SUI	Wanna kill myself tonight. Having suicidal thoughts *** now. I wanna end it *** slit my wrists. After being on this earth *** I'm done.
889	SUI	*** end it all today No more suffering *** humiliation *** worrying about the future, *** such a coward i wanna die so bad
1864	SUI	Please someone help me I need *** an effective and painless method *** tell me I can't last longer.

segments. We assume that this is primarily due to Twitter posts being generally shorter in length, whereas Reddit posts tend to be longer as demonstrated in Table 4.1.

4.6 Results

4.6.1 Overall Performance

We compare our model with fine-tuned PLM baselines and several published baselines that use the same mental health detection datasets and evaluation setup. Note that we select post-only mental health detection models as mentioned in Sections 4.1. We comprehensively evaluate the overall performance and class performance in Table 4.8.

Overall, our model outperforms all baselines on all four benchmark datasets. What should be noted is that our model does not have to be trained with all three different teachers to achieve the best results. As illustrated in Table 4.8, we have four datasets, the initial two originating from Twitter and the latter from Reddit. Our model demonstrates superior performance, even with partial teacher combinations. The datasets from Twitter produce the best results with the combination of Text and Emotion, whereas the Reddit-based datasets perform the best with Text and Audio Knowledge Distillation. The efficacy of each modality teacher combination on different datasets is detailed in Section 4.6.2.

In addition, our model trained with all three modalities still outperforms the other baseline models in most cases and shows greater performance in identifying certain classes. Especially for the *Moderate* (MO) class of the DEPTWEET dataset, our model trained with all three modalities achieves a 34.34 F1 score, while our model trained with partial modalities only achieves a 12.31 F1 score. All the other pre-trained baseline models fail to recognise the MO class. Hence, we can conclude that learning from different modality teachers helps our model achieve much better performances than the baseline models that learned from only textual inputs. Such improvement is more noticeable on datasets with shorter texts. Specifically, our model's best performance is 8.36% and 8.07% higher than the best-performing baseline model on the macro F1 and weighted F1, respectively, on the TwitSuicide dataset. For DEPTWEET, IdenDep, and SDCNL datasets, the best performances of our model are 6.03%, 5.40%, 4.40%, and 2.88%, 4.50%, 4.37% higher than the best-performing baseline model on the macro and weighted F1 scores, respectively.

TABLE 4.8. Overall results using all three teacher modalities (**Ours (All)**) and the best partial teacher combination (**Ours (Best Partial Combination)**) against baselines. Class abbreviation definitions may be found in the Figure 4.2 caption. We present a full teacher combination ablation study in Table 4.9. **Bold** face indicates best score while second best are underlined.

	Overall Performance			**Breakdown F1 Scores**			
TwitSuicide	**Acc**	**F1m**	**F1w**	**(SC)**	**(PC)**	**(SI)**	
Long et al. [138]	56.67	-	-	40.00	50.00	66.00	
BERT	57.58	53.60	57.25	40.00	57.00	64.00	
RoBERTa	55.45	50.61	54.43	37.18	51.63	63.03	
MentalBERT	57.73	52.57	57.39	35.23	56.65	65.84	
MentalRoBERTa	55.91	51.49	55.60	41.62	53.05	62.81	
Ours (Text&Emo)	**65.76**	**61.96**	**65.46**	49.72	**62.34**	**73.81**	
Ours (All)	61.21	59.64	61.23	**54.17**	59.50	65.26	
DEPTWEET	**Acc**	**F1m**	**F1w**	**(SE)**	**(MO)**	**(MI)**	**(ND)**
Kabir et al. [103]†	79.75	38.59	78.89	17.65	21.98	25.43	89.29
BERT	81.89	40.40	80.21	36.36	00.00	34.34	90.90
RoBERTa	82.18	32.04	80.14	00.00	00.00	36.77	91.41
MentalBERT	83.54	36.04	79.72	26.09	00.00	26.67	91.41
MentalRoBERTa	78.48	36.60	78.62	22.22	00.00	34.91	89.28
Ours (Text&Emo)	**84.03**	**46.43**	**83.09**	**41.38**	12.31	**39.07**	**92.95**
Ours (All)	82.77	46.20	82.61	26.09	**34.34**	32.32	92.04
IdenDep	**Acc**	**F1m**	**F1w**	**(DE)**	**(NDE)**		
Tadesse et al. [215]	91.00	-	-	93.00	-		
Ren et al. [190]	91.30	-	-	93.98	-		
BERT	88.65	85.23	88.10	92.34	78.12		
RoBERTa	87.18	82.85	86.34	91.47	74.24		
MentalBERT	89.63	86.71	89.23	92.93	80.49		
MentalRoBERTa	90.11	87.70	89.91	93.15	82.26		
Ours (Text&Audio)	**94.30**	**93.10**	**94.26**	**95.97**	**90.23**		
Ours (All)	93.92	92.58	93.85	95.73	89.43		
SDCNL	**Acc**	**F1m**	**F1w**	**(SUI)**	**(DEP)**		
Haque et al. [84]	72.24	-	-	73.61	-		
Ansari et al. [10]	-	-	-	76.00	-		
BERT	67.02	66.57	66.64	70.45	62.69		
RoBERTa	70.97	70.63	70.69	73.81	67.46		
MentalBERT	69.39	69.21	69.26	71.57	66.86		
MentalRoBERTa	72.30	72.10	72.14	74.45	69.74		
Ours (Text&Audio)	**76.52**	**76.50**	**76.51**	77.12	**75.88**		
Ours (All)	75.20	74.84	74.90	**77.83**	71.86		

† Replicated results.

TABLE 4.9. Ablation study using different combinations of teacher modalities. Class abbreviation definitions may be found in the Figure 4.2 caption. **Bold** face indicates best score while second best are underlined. A ✓indicates the addition of the emotion (Emo) and/or the audio (Aud) teacher/s. Highlighted rows show the best setup.

Text‡	Emo	Aud	Overall Performance			Breakdown F1 Scores			
TwitSuicide			**Acc**	**F1m**	**F1w**	**(SC)**	**(PC)**	**(SI)**	
✓	×	×	58.94	47.50	56.13	16.13	56.55	69.81	
✓	✓	×	**65.76**	**61.96**	**65.46**	49.72	**62.34**	**73.81**	
✓	×	✓	63.79	58.94	63.02	44.30	61.76	70.74	
✓	✓	✓	61.21	59.64	61.23	**54.17**	59.50	65.26	
DEPTWEET			**Acc**	**F1m**	**F1w**	**(SE)**	**(MO)**	**(MI)**	**(ND)**
✓	×	×	**84.52**	44.33	82.40	40.00	13.11	31.28	92.90
✓	✓	×	84.03	**46.43**	**83.09**	**41.38**	12.31	**39.07**	**92.95**
✓	×	✓	83.06	36.00	81.69	00.00	15.38	36.26	92.33
✓	✓	✓	82.77	46.20	82.61	26.09	**34.34**	32.32	92.04
IdenDep			**Acc**	**F1m**	**F1w**	**(DE)**	**(NDE)**		
✓	×	×	92.32	90.67	92.26	94.60	86.74		
✓	✓	×	93.86	92.47	93.78	85.71	89.23		
✓	×	✓	**94.30**	**93.10**	**94.26**	**95.97**	**90.23**		
✓	✓	✓	93.92	92.58	93.85	95.73	89.43		
SDCNL			**Acc**	**F1m**	**F1w**	**(SUI)**	**(DEP)**		
✓	×	×	75.20	75.07	75.10	76.85	73.30		
✓	✓	×	72.82	72.45	72.51	75.65	69.25		
✓	×	✓	**76.52**	**76.50**	**76.51**	77.12	**75.88**		
✓	✓	✓	75.20	74.84	74.90	**77.83**	71.86		

‡We report results from the best performing pre-trained language model for each dataset (Table 4.10): MentalBERT for DEPTWEET and BERT for the others.

4.6.2 Effectiveness of Multimodal Multi-Teachers

To examine the efficacy of each teacher modality and their combinations, we evaluate and explore the framework's performance by the emotion-based teacher, the audio-based teacher, or both, alongside the text-based teacher. The results are presented in Table 4.9.

In general, the multi-teacher structure outperforms the use of a singular text-based teacher, although the effectiveness of different modalities varies across datasets. For the Twitter-based datasets (TwitSuicide and DEPTWEET), applying the emotion-based and text-based teachers together achieves the best results. In contrast, for Reddit-based datasets (IdenDep and SDCNL), the audio- and text-based teachers have better performances. This difference may be attributed to the longer posts in Reddit-based datasets, resulting in longer audio (Table

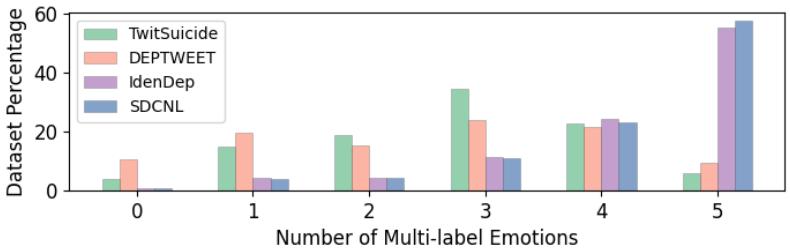


FIGURE 4.5. Distribution of multi-label emotion class labels.

4.1) that contains more acoustic information beneficial for the audio-based teacher. Moreover, due to the lengthier nature of posts in Reddit-based datasets, more emotion lexicon tokens are likely to be matched compared to Twitter-based datasets. This results in a higher number of generated emotion labels during the learning process of the multi-label emotion-based teacher. In Figure 4.5, we compare the number of generated multi-label emotion classes utilised to train the emotion-based teacher across all four datasets. It is evident that a greater proportion of posts in Reddit-based datasets match more emotions from the seven emotion labels (Section 4.3.1.2) compared to Twitter-based datasets. This potential increase in the number of matching emotion labels may present challenges in distinguishing between different emotions during the training of the emotion-based teacher, potentially impacting downstream mental health classification, especially for ambiguous classes such as *Suicide* (SUI) and *Depression* (DEP) in the SDCNL dataset.

We can conclude that using multimodal teachers generally helps detect mental health, and these findings also suggest varying effectiveness of different modalities across datasets with distinct characteristics, offering valuable insights into selecting suitable modalities for improved performance in future scenarios.

4.6.3 Impact of Text-based Teachers

We compare the effectiveness of various PLMs for the text-based teacher. Table 4.10 shows that BERT produces the best weighted F1 across all datasets. However, the domain-specific PLMs perform better for the more concerning classes in a multi-class setup. ClinicalBERT outperforms BERT by 6.66% for *Strongly Concerning* (SC) in the TwitSuicide dataset, while

TABLE 4.10. Ablation study using different PLMs for the text-based teacher. We report results using the best-performing teacher modality combination in Table 4.9 and change only the text-based teacher. Class abbreviation definitions may be found in the Figure 4.2 caption. **Bold** face indicates best score while second best are underlined.

	Overall Performance			Breakdown F1 Scores			
TwitSuicide	**Acc**	**F1m**	**F1w**	**(SC)**	**(PC)**	**(SI)**	
BERT	**65.76**	61.96	**65.46**	49.72	62.34	**73.81**	
RoBERTa	65.15	61.67	64.70	51.19	61.43	72.40	
MentalBERT	63.79	61.08	63.67	51.65	**64.08**	67.52	
ClinicalBERT	65.30	**63.23**	65.25	**56.38**	62.18	71.13	
DEPTWEET	**Acc**	**F1m**	**F1w**	**(SE)**	**(MO)**	**(MI)**	**(ND)**
BERT	83.84	39.80	**83.84**	0.00	**22.22**	**44.05**	92.93
RoBERTa	82.67	35.99	81.80	0.00	14.29	37.23	92.44
MentalBERT	**84.03**	**46.43**	83.09	**41.38**	12.31	39.07	**92.95**
ClinicalBERT	83.54	32.29	80.71	0.00	0.00	37.11	92.05
IdenDep	**Acc**	**F1m**	**F1w**	**(DE)**	**(NDE)**		
BERT	**94.30**	**93.10**	**94.26**	**95.97**	**90.23**		
RoBERTa	94.13	92.89	94.09	95.86	89.93		
MentalBERT	93.21	91.71	93.14	95.24	88.17		
ClinicalBERT	94.24	92.95	94.17	**95.97**	89.94		
SDCNL	**Acc**	**F1m**	**F1w**	**(SUI)**	**(DEP)**		
BERT	**76.52**	**76.50**	**76.51**	**77.12**	**75.88**		
RoBERTa	75.20	75.03	75.07	77.07	72.99		
MentalBERT	73.61	73.61	73.61	73.40	73.82		
ClinicalBERT	75.46	75.46	75.45	75.07	75.84		

MentalBERT achieves 41.38% for *Severe* (SE) in the DEPTWEET dataset, surpassing the other language models which failed to recognise it. To ensure optimal performance, we specifically employ MentalBERT for DEPTWEET, while BERT is used for the other datasets. Nonetheless, the overall performance of the text-based teacher is not significantly impacted by the choice of PLMs. More performance enhancement stems from the inclusion of different modalities, as discussed in the previous sections.

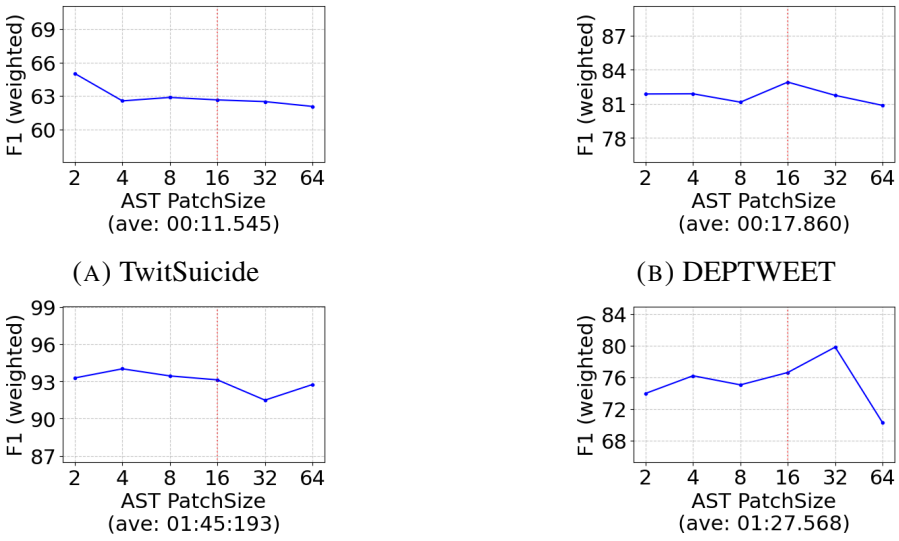
4.6.4 Impact of Student Model Inputs

We examine different combinations of multimodal inputs for the student model in Table 4.11 in order to explore the optimal input for a knowledge distillation for the student model. We concatenate emotion embeddings, audio embeddings, or both with text embeddings from

TABLE 4.11. Ablation study using different combinations of input modalities to the student model. **Bold** face indicates best score while second best are underlined. A ✓indicates the addition of the emotion-based (Emo) and/or the audio-based (Aud) input features. Highlighted rows show our proposed student setup. VT: randomly initialised vanilla transformer.

Text	**Emo**	**Aud**	**Overall Performance**			**Breakdown F1 Scores**			
TwitSuicide			**Acc**	**F1m**	**F1w**	**(SC)**	**(PC)**	**(SI)**	
BERT	✓	✓	51.67	45.43	50.46	26.95	52.00	57.34	
BERT	✓	×	50.91	44.24	48.58	29.58	41.42	61.71	
BERT	×	✓	48.64	33.87	43.31	0.00	41.07	60.55	
BERT	×	×	**65.76**	**61.96**	**65.46**	**49.72**	**62.34**	**73.81**	
VT	×	×	46.36	32.72	41.77	0.00	41.09	57.06	
DEPTWEET			**Acc**	**F1m**	**F1w**	**(SE)**	**(MO)**	**(MI)**	**(ND)**
BERT	✓	✓	**85.49**	34.75	81.24	0.00	**20.59**	25.86	92.54
BERT	✓	×	83.84	32.15	80.81	0.00	0.00	36.36	92.55
BERT	×	✓	84.23	22.86	77.01	0.00	0.00	0.00	91.44
BERT	×	×	84.03	**46.43**	**83.09**	**41.38**	12.31	**39.07**	**92.95**
VT	×	×	84.23	22.86	77.01	0.00	0.00	0.00	91.44
IdenDep			**Acc**	**F1m**	**F1w**	**(DE)**	**(NDE)**		
BERT	✓	✓	91.58	89.99	91.60	93.98	86.00		
BERT	✓	×	91.09	89.21	91.03	93.72	84.70		
BERT	×	✓	91.75	89.85	91.62	94.23	85.47		
BERT	×	×	**94.30**	**93.10**	**94.26**	**95.97**	**90.23**		
VT	×	×	75.77	61.25	70.85	84.97	37.54		
SDCNL			**Acc**	**F1m**	**F1w**	**(SUI)**	**(DEP)**		
BERT	✓	✓	67.55	67.54	67.55	67.89	67.20		
BERT	✓	×	67.81	67.80	67.80	67.38	68.23		
BERT	×	✓	68.34	68.32	68.31	67.57	69.07		
BERT	×	×	**76.52**	**76.50**	**76.51**	**77.12**	**75.88**		
VT	×	×	67.02	66.94	66.97	68.51	65.37		

pre-trained BERT and then pass them to the transformer layer after a linear layer projection. We also test a randomly initialised vanilla transformer compared to the pre-trained BERT. The results indicate that unimodal textual post inputs outperform the concatenation of multimodal inputs for the student model. Moreover, pre-trained BERT yields better results than the randomly initialised vanilla transformer across all datasets. These outcomes underscore the effectiveness of multimodal knowledge acquired from the multi-teachers, efficiently guiding the student to achieve robust performance with only textual inputs.



(A) TwitSuicide (B) DEPTWEET

(C) IdenDep (D) SDCNL

FIGURE 4.6. Parameter study for the audio-based teacher model. Ave: average audio duration for the dataset.

4.6.5 Audio Teacher Parameter Testing

We further investigate the different patch size values for the audio-based teacher model while maintaining a consistent setup for the student model. Figure 4.6 shows each dataset's weighted F1 score for each patch size value. TwitSuicide, DEPTWEET, and IdenDep datasets show a relatively stable performance between 2 to 64 patch sizes; however, for SDCNL, performance improvement may be achieved using a patch size of 32. This may be due to the higher variance in audio duration of the outliers compared to the other three datasets (Figure 4.3). Despite being shorter on average length and duration than the other Reddit-based dataset, SDCNL has some longer audio samples, which may have benefited from a patch size of 32. However, a sharp decline in performance could be expected when the patch size is increased to 64.

4.7 Conclusion

In conclusion, our study introduces a multimodal multi-teacher knowledge distillation model, 3M-Health, designed for mental health detection and presents a comprehensive exploration.

Our experiments demonstrate that the multimodal approach outperforms unimodal counterparts, with the choice of modalities influencing performance across diverse datasets. Notably, the incorporation of audio-based information proves particularly beneficial for social media post-based mental health detection for Reddit-based datasets, emphasising the importance of modality selection based on the nature of the data. Overall, our work contributes valuable insights into the nuanced dynamics of multimodal knowledge distillation for mental health detection, offering a promising avenue for future research in this critical domain.

This chapter, through the introduction and evaluation of 3M-Health, addresses the loss of tonal cues in textual data and semantic representations. It addresses this thesis's research questions by identifying the derived acoustic modality (RQ1), integrating it with other modalities through a multi-teacher distillation framework (RQ2), and applying it to depression and suicide ideation detection tasks in social media (RQ3). In particular, 3M-Health leverages the parallel and complementary affective abstractions from the emotion and acoustic modalities, wherein the acoustic modality derives paralinguistic and prosodic abstractions that capture emotional shifts, intensities, and even psychological distress. It proposes the integration of these modalities, along with the semantic modality, through a distributed multi-teacher knowledge distillation framework, allowing each teacher to specialise with each modality's inductive biases separately before distilling knowledge to a student model. This distillation of expert knowledge effectively enriches the student model's learned representations with multimodal abstractions despite remaining unimodal. Finally, the framework proves effective for mental health detection in social media, particularly for use cases with less explicit emotional words conveying depression and other mental health issues.

CHAPTER 5

Relational and Structural Modality for Information Extraction of Drug-Related Reactions

---

This chapter is the published work **TriG-NER: Triplet Grid-Framework for Discontinuous Named Entity Recognition** [34] published in **WWW 2025**. I am a first author of this paper. I co-developed the research aim and co-designed the methodology. I conducted all of the experiments, analysed the results, and wrote most of the paper.

This work introduces the word-pair modality to address the inherent sequential bias of language models, which limits the extraction of complex information prevalent in medical and mental health-related texts (Table 1.1). Parallel to the emotion and acoustic modalities, the word-pair modality represents a reformulation of information from textual data; however, it incorporates relational and structural abstractions rather than affective and tonal signals, which have diminished significance in objective texts and tasks. This study proposes a grid-based triplet framework to incorporate the word-pair modality, expanding on traditional sequential text processing methods and demonstrating greater effectiveness in extracting complex, disjointed entities (Table 1.2).

Specifically, this framework is applied to a discontinuous named entity recognition (DNER) task. DNER, while not a direct mental health task like disease risk classification, is not merely an NLP objective but a prerequisite for reliable and scalable mental health risk monitoring. Improving DNER enables more accurate analysis of mental health-related factors through the inclusion of disjointed entities prevalent in both user-generated texts and clinical notes; however, these entities are often overlooked and simplified in favour of continuous entities [8].

Neglecting disjointed entities misrepresents and misses crucial insights essential in clinical and mental health analysis.

This work more specifically focuses on improving DNER for adverse drug events and drug-related reactions, which are fundamental for investigating mental health triggers and outcomes. Understanding drug-related reactions is an important facet of substance use disorder (SUD), which in and of itself is a mental health condition and could also be a comorbidity of other mental health problems, complicating diagnosis and treatment. Antidepressants and other prescription drugs are also often abused for non-medical purposes [18], right after illicit drugs, alcohol, and tobacco. Furthermore, researchers have found that 103 prescription drugs are associated with suicidal ideation [106]. This indisputable connection between drugs and mental health makes the extraction of drug-related reactions and adverse events, with specific inclusion of disjointed mentions, a crucial task for reliable mental health systems.

Discontinuous Named Entity Recognition (DNER) presents a challenging problem where entities may be scattered across multiple non-adjacent tokens, making traditional sequence labelling approaches inadequate. Existing methods predominantly rely on custom tagging schemes to handle these discontinuous entities, resulting in models tightly coupled to specific tagging strategies and lacking generalisability across diverse datasets. To address these challenges, we propose TriG-NER, a novel Triplet-Grid Framework that introduces a generalisable approach to learning robust token-level representations for discontinuous entity extraction. Our framework applies triplet loss at the token level, where similarity is defined by word pairs existing within the same entity, effectively pulling together similar and pushing apart dissimilar ones. This approach enhances entity boundary detection and reduces the dependency on specific tagging schemes by focusing on word-pair relationships within a flexible grid structure. We evaluate TriG-NER on three benchmark DNER datasets and demonstrate significant improvements over existing grid-based architectures. These results underscore our framework's effectiveness in capturing complex entity structures and its adaptability to various tagging schemes, setting a new benchmark for discontinuous entity extraction.[1]

[1]Code available at https://github.com/adlnlp/trig_ner.

5.1 Introduction

Named Entity Recognition (NER) is a fundamental task in natural language processing that involves identifying and categorising entities such as person names, locations, or temporal expressions within unstructured text. Traditionally, NER has been approached using sequential labelling techniques like the Begin-Inside-Outside (BIO) scheme, which assigns labels to each token in a sentence. However, while effective for contiguous entities, such schemes struggle to accurately capture discontinuous named entities whose mentions are interrupted by non-entity tokens due to their linear nature and inability to represent complex entity structures.

Recent research in Discontinuous Named Entity Recognition (DNER) has sought to address these limitations by introducing new tagging schemes and model architectures. These include extensions of the BIO scheme like BIOHD [218], span-based methods [234], and grid-based tagging [244], which attempt to represent more complex entity boundaries and relationships. While these methods have shown improvements in extracting discontinuous entities, they often suffer from heavy reliance on task-specific tagging strategies. This makes them highly specialised, limiting their adaptability to new datasets and unseen entity types. Moreover, current solutions primarily focus on sample-based learning objectives, which do not fully capture the token-level dependencies critical for recognising scattered entities. Generative and large language models (LLMs) like ChatGPT have also been explored for DNER, using sequence-to-sequence approaches to generate entity spans. However, these models, optimised for next-word prediction, are not inherently suited for the intricate nature of NER tasks, making them prone to generating incorrect spans and entity boundaries. Grid-tagging methods, on the other hand, have achieved state-of-the-art performance in DNER by modelling word-pair relationships. Nevertheless, they often lack a mechanism to differentiate between similar and dissimilar word-pair representations, particularly for discontinuous entities separated by non-entity tokens.

To address these challenges, we introduce **TriG-NER**, a Triplet-Grid Framework that leverages token-based triplet loss to learn fine-grained word-pair relationships for DNER. Unlike traditional triplet loss, which operates at the sample level by comparing entire sequences,

our method applies triplet loss at the token level, where similarity is defined by word pairs co-occurring within the same entity. This approach enables the model to capture the local dependencies between tokens in discontinuous entities, ensuring that word pairs forming an entity are cohesively represented in the learned feature space. We also propose a grid-based triplet loss that models word-pair relationships within a flexible grid structure, where positive pairs represent tokens within the same entity, and negative pairs include word pairs disrupted by non-entity tokens. The main contributions of this paper are as follows:

1. Token-based Triplet Loss for NER: We introduce a novel token-based triplet loss that learns fine-grained token-level representations for discontinuous entity extraction, contrasting with existing methods that use sample-based triplet loss.

2. Grid-based Triplet Loss Using Word-Pair Relationships: We propose a grid-based triplet loss that defines word-pair similarity based on co-occurrence within the same entity, enhancing the model's ability to capture non-adjacent entity segments.

3. Extensive Evaluations and Qualitative Analysis: We perform extensive evaluations on three widely used DNER benchmark datasets and provide a qualitative analysis that demonstrate the effectiveness of our grid-based triplet framework over existing baselines and prompted large language models.

5.2 Related Works

5.2.1 Discontinuous Named Entity Recognition

Named entity extraction and recognition has traditionally been viewed as a sequence labelling task using the Begin-Inside-Outside (BIO) tags; however, this traditional approach fails for more complex entities such as discontinuous entities. Researchers have recently focused on improving discriminative discontinuous entity recognition through various tagging schemes and methods. Tang et al. (2015) [218] was the first to extend BIO sequential tagging to BIOHD to distinguish inter-entity boundaries, which subsequent studies [153, 217] followed. More

TABLE 5.1. Comparison of NER schemes and losses in recent works in discontinuous named entity recognition.

DNER Models	Core Scheme	Loss
Corro (2024) [48]	Sequence Tagging	NLL
Wang et al. (2019) [234]	Span-based	NLL
Li et al. (2021) [118]	Span-based	NLL
Huang et al. (2023) [90]	Span-based	NLL
Mao et al. (2024) [144]	Span-based	BCE
Dai et al. (2020) [50]	Transition-based	-
Wang et al. (2021) [244]	Grid Tagging	CE
Li et al. (2022) [119]	Grid Tagging	NLL
Liu et al. (2022) [131]	Grid Tagging	CE
Fei et al. (2021) [60]	Seq2Seq	NLL
Yan et al. (2021) [261]	Seq2Seq	NLL
Zhang et al. (2022) [274]	Seq2Seq	-
Xia et al. (2023) [255]	Seq2Seq	MLE
Zhao et al. (2024) [279]	Prompting	-
Zhu et al. (2024) [282]	Prompting	-
Ours	Word-Pair Grid Tagging	Triplet

recently, Corro (2024) [48] proposed a two-layer tagging scheme that uses ten tags; however, these methods fail to capture complex discontinuous entities and suffer from decoding ambiguity. Span-based methods [234, 118, 90] involve the identification of all candidate spans and the merging of disjoint spans. The two-step process, however, is vulnerable to error propagation and identifying all possible span candidates is resource-exhaustive. Other discriminative methods, such as hypergraphs [163, 235] and stack-and-buffer transitions [50], are also explored yet still suffer from error propagation. On the other hand, generative methods [60, 261, 274, 255], leverage sequence-to-sequence language models to directly generate entity spans and types that overcome the challenges presented by different complex entity structures. With the advent of ChatGPT, research in applying large language model (LLM) prompting to discontinuous NER has also seen increased attention [282, 279]. However, generative models are optimised for next-word prediction, not NER, predisposing it to incorrect biases.

Grid tagging [244], another discriminative method, has shown state-of-the-art performance through identifying spans using word pair tags defining word-pair relationships [119, 131]. However, grid tagging approaches are still constrained by their reliance on specific grid tag designs and decoding strategies. Moreover, they tend to treat word pairs independently, failing

to capture the contextual relationships between word pairs that could enhance the recognition of discontinuous entities. This lack of dependency modelling between similar and dissimilar word pairs can result in the misclassification of complex, scattered entity spans. To address these limitations, we propose TriG-NER, a novel Triplet-Grid Framework that integrates token-based triplet loss with grid tagging to model fine-grained word-pair relationships. Unlike existing methods that treat word pairs in isolation, our approach leverages triplet loss to distinguish between similar and dissimilar word pairs, enhancing the model's ability to recognise non-adjacent entity segments.

5.2.2 Triplet Loss

Triplet loss [204] was introduced in the computer vision (CV) area in the field of facial recognition or reidentification [156, 68, 267] for deep metric learning by directly optimising image sample embeddings. Unlike contrastive loss, triplet loss takes three points - an anchor, a positive, and a negative - and ensures that the positive is closer to the anchor than the negative point by a certain margin. This optimisation effectively pulls together images belonging to the same person and pushes away seemingly similar images that do not share the same identity, producing a better feature space. As a result, triplet loss has seen wide adoption and a few variations in other CV fields, such as image segmentation [208], facial synthesis [238], 3D object retrieval [85], and medical image classification [80]. In the area of natural language processing (NLP), researchers have explored the use of triplet loss for text classification [254, 151], relation extraction [205], and spoken language understanding (SLU) [189, 232].

However, traditional triplet loss is typically employed at the sample level, where similarity is defined by class membership, which does not necessarily align with the needs of discontinuous entity extraction. Detecting discontinuous entities requires capturing local dependencies and boundary information within entities scattered across non-adjacent tokens. Our proposed framework addresses these limitations by introducing a grid-based, token-level triplet loss, where word-pair co-occurrence within the same entity defines similarity. This approach ensures that entity tokens are drawn closer together in the feature space, even when interrupted by non-entity tokens that may appear syntactically or semantically similar. To the best of our

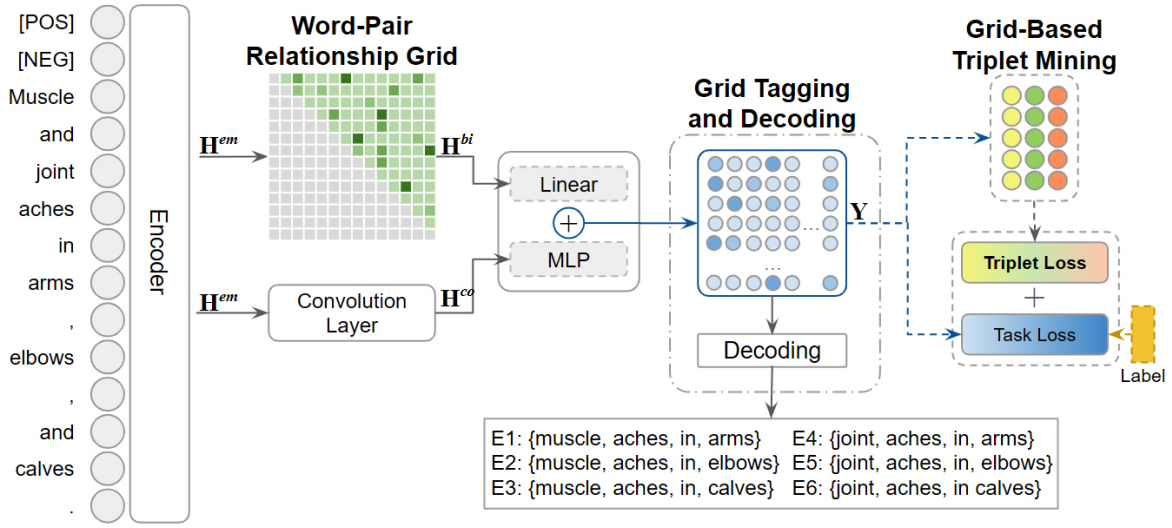


FIGURE 5.1. Overall framework of the proposed TriG-NER

knowledge, no existing work has applied a grid-based, token-level triplet loss for discontinuous named entity recognition, making our approach a novel contribution to this field.

5.3 Methodology

In this study, we propose a new type of DNER architecture that utilises word-pair relationships in a grid structure, along with grid-based triplet mining to improve discontinuous entity extraction. Our framework builds on recent advances in grid tagging and word-to-word relation classification, introducing a novel combination of grid-based tag decoding and triplet loss mechanisms. This section provides an overview of a grid-based NER model, our newly proposed NER model with a word-pair relationship grid, grid tagging and decoding, and grid-based triplet loss.

5.3.1 Grid-based NER Models

Recent studies on Named Entity Recognition (NER) have explored using grid-based tagging schemes to improve discontinuous entity extraction, especially where traditional sequence tagging approaches like the Begin-Inside-Outside (BIO) scheme fall short. In grid-based models, the NER task is treated as a word-to-word relation classification problem, where

a sequence input $X = \{x_1, x_2, ..., x_n\}$ of length n is transformed into a grid output $\mathbf{Y} = \{y_{11}, y_{12}, ..., y_{nn}\} \in \mathbb{R}^{n \times n \times c}$, where c is the number of tag classes. Each element $y_{ij} \in \mathbb{R}^c$ represents the logits used to calculate the probability of a relationship between word i and word j.

Grid-based NER models focus on word-pair relationships, where token pairs, rather than individual tokens, are labelled. This structure allows for representing complex, non-contiguous entity structures, making it a flexible method for DNER. Existing models such as those proposed by [119] and [244] have shown promising results by utilising these word-to-word grids, which map the relationships between tokens, allowing models to handle both contiguous and non-contiguous entities effectively. However, these models treat each word pair independently, which overlooks the inherent relationships between multiple word pairs that can exist within the same entity. This lack of dependency modelling between similar and dissimilar word pairs can result in misclassifications, particularly when dealing with complex, non-adjacent entity structures.

5.3.2 Word-Pair Relationship Grid

Hence, we address this limitation by introducing triplet loss at the word-pair level, which enables the model to explicitly learn the fine-grained distinctions between similar and dissimilar word pairs within the grid. To achieve this, we introduce a word-pair relationship grid to explicitly model the relationships between words within entities. The proposed word-pair relationships are treated as the primary feature for entity extraction, and the overall NER task is transformed into a word-pair classification problem.

The input sentence is first passed through an encoder layer, where we utilise pre-trained language models (PLMs) such as BERT [55], BioClinicalBERT [9], PharmBERT [228], and PubMedBERT [78]. These models generate contextualised word embeddings $\mathbf{H}^{em} \in \mathbb{R}^{n \times d}$, where d is the embedding dimension. A bidirectional LSTM layer is then applied to capture sequential dependencies in the sentence. The embeddings are then passed through two distinct modules: a Convolution Layer and a Biaffine transformation. The Convolution Layer

module [119] generates enhanced word-pair representations $\mathbf{H}^{co} \in \mathbb{R}^{n \times n \times d^{co}}$, where d^{co} is the convolution dimension, while the Biaffine transformation computes word-pair relationships $\mathbf{H}^{bi} \in \mathbb{R}^{n \times n \times d^{bi}}$. These representations are combined in a Co-Predictor Layer, where a linear layer and an MLP map $\mathbf{H}^{bi}$ and $\mathbf{H}^{co}$ to tag relation logits $\mathbf{Y}^{bi}$ and $\mathbf{Y}^{co} \in \mathbb{R}^{n \times n \times c}$. The final grid tag logits are obtained by combining the two: $\mathbf{Y} = \mathbf{Y}^{bi} + \mathbf{Y}^{co}$.

5.3.3 Grid Tagging and Decoding

The grid tagging system, reproduced from [119], classifies word-pair relationships using three tag classes: *None*, *Next-Neighboring-Word* (NNW), and *Tail-Head-Word* (THW). These classes define whether a word pair has no relationship, a neighbouring relationship within an entity, or represents the start and end of an entity, respectively. Once word-pair relationships are classified, the grid decoding process begins, which is crucial for discontinuous entity extraction. The system takes the final grid tag logits $\mathbf{Y}$ and decodes the predicted relationships into entity structures. By focusing on word pairs rather than individual tokens, the grid structure allows our model to flexibly identify discontinuous entity boundaries, which are common in complex entity recognition tasks. The grid tagging and decoding approach enables the model to handle non-contiguous entity spans by considering the relationships between word pairs, making it robust against the limitations of sequential tagging schemes.

5.3.4 Grid-based Triplet Mining

5.3.4.1 Preliminaries

To further optimise the model's performance in capturing discontinuous entities, we introduce a grid-based triplet loss, which enables the model to learn distinctions between similar and dissimilar word pairs more effectively. Triplet loss is a metric learning objective that brings similar word pairs closer while pushing dissimilar pairs farther apart. The loss function is defined as $L_{triplet} = \sum max(f(a,p) - f(a,n) + m, 0)$ where a is an anchor point, p is a positive point similar to the anchor, n is a negative point dissimilar to the anchor, f is a distance function, and m is a margin that ensures a minimum distance between negative pairs

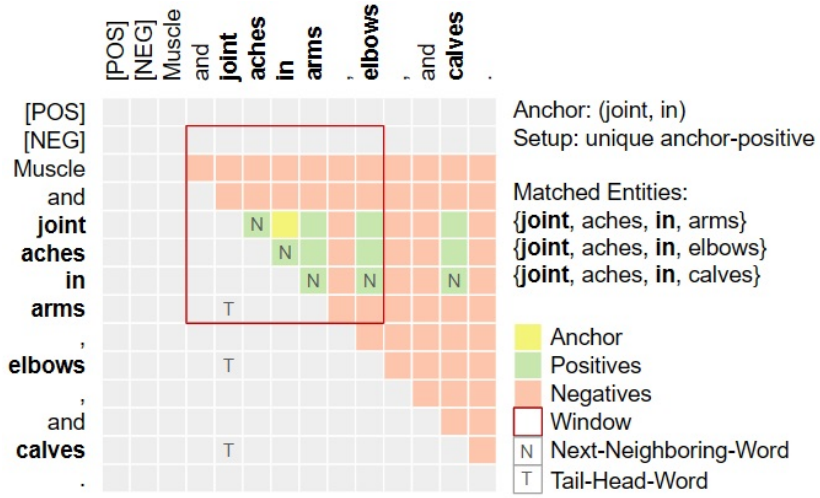


FIGURE 5.2. Example of positive and negative candidates based on the anchor ("joint", "in") with a candidate window of 3.

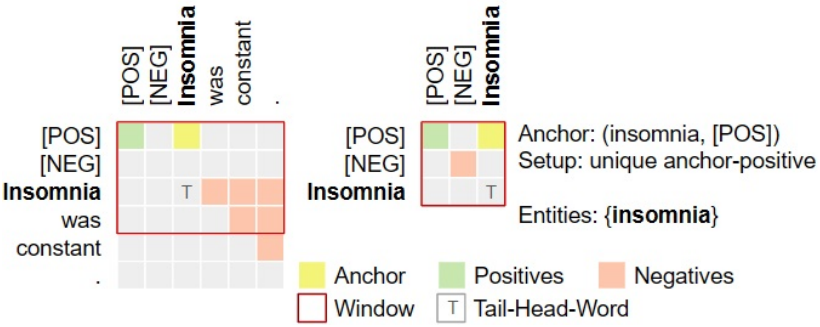


FIGURE 5.3. Example of positive and negative candidates for one-word entities (left) and one-word samples (right).

and positive pairs. We utilise Euclidean distance for our distance function. Our final loss combines the triplet loss with the task loss: $L_{final} = L_{task} + L_{triplet}$.

5.3.4.2 Word-Pair Grid Implementation

We extract our triplets from the word-pair grid representations in our framework. Unlike most sample-based triplet loss implementations that define similarity by sample classes, we define the similarity of our triplet elements based on their existence within entities. For the anchor candidates, we use word-pair grid points that exist in any entity. Each anchor candidate is then matched with positive and negative candidates. Positive candidates are word pairs that

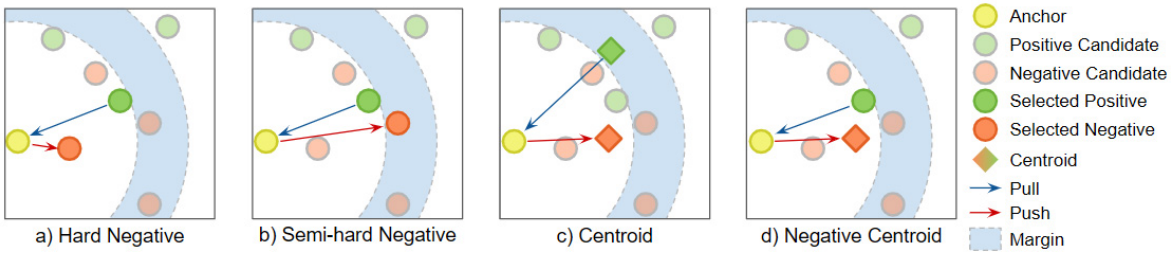


FIGURE 5.4. Triplet Mining Methods

co-exist with the anchor in any entity, while negative candidates are word pairs that don't belong to any entity the anchor is a part of. We illustrate this candidate selection in Figure 5.2.

For special instances, we incorporate two special tokens `[POS]` and `[NEG]` at the start of each sample. These special instances include one-word entities and anchor points that do not have other positive or negative word pairs to match with. For the example in Figure 5.3, the sentence "Insomnia was constant ." with "Insomnia" as an entity uses $cell_{\texttt{insomnia,[POS]}}$ as the anchor point. Since no other positive point could be matched, $cell_{\texttt{[POS],[POS]}}$ is the only positive candidate. In cases where no negative candidates can be used, $cell_{\texttt{[NEG],[NEG]}}$ is used. We experiment with extracting our triplet representations from the Word-Pair Relationship Grid ($\mathbf{H}^{bi}$) or from the final output logits ($\mathbf{Y}$).

5.3.4.3 Triplet Selection

It is crucial to select valid triplets that violate the triplet constraint wherein the positive candidates are farther from the anchor than the negative candidates by a margin [204]. Since generating all possible anchor-positive-negative combinations not only exponentially increases computation time and resources needed but, more importantly, generates uninformative triplets that result in slower convergence during training, we utilise different online triplet selection methods illustrated in Figure 5.4.

(1) **Hard Negative (HN)** selection takes each anchor-positive combination and selects the closest negative candidate from the anchor.

(2) **Semi-hard Negative (SN)** selection takes each anchor-positive combination but, different from the hard negative, selects the negative candidate that is closest to the anchor but farther than the positive point within the set margin.

(3) **Centroid (CE)** takes the mean of all the positive candidates and the mean of all the negative candidates for each anchor as the positive and the negative points.

(4) **Negative Centroid (NC)** utilises all anchor-positive pairs but takes the mean of all the negative candidates as the negative point.

Due to the exponential increase of positive and negative candidates as the sample length increases, we further limit the positive and negative candidate selection by using a candidate window centred on the anchor and by specifically using unique anchor-positive pairs. The unique anchor-positive pair setup utilises only the top half triangle of the grid (Figure 5.2) where an anchor token pair tp_1 is paired with a positive candidate tp_2, but when tp_2 is set as an anchor, tp_1 will not be considered as a positive candidate anymore. This reduces possible redundant information that is not helpful for training, while simultaneously reducing the number of triplets. A comparison of performance between unique and non-unique anchor-positive pairs is provided in Table 5.6.

5.4 Experimental Setup

5.4.1 Datasets

Following previous studies on discontinuous named entity recognition, we use three datasets in the biomedical domain to assess the performance of our proposed system. The CSIRO Adverse Drug Event Corpus (**CADEC**)[2] [107] is a collection of medication consumer posts annotated for entity identification from the public forum AskAPatient. We follow previous literature and use only the adverse drug reaction (ADR) entities. **ShARe13**[3] [183] and **ShARe14**[4] [161] datasets are part of the Shared Annotated Resources used for the CLEF

[2] https://doi.org/10.4225/08/570FB102BDAD2
[3] https://doi.org/10.13026/rxa7-q798
[4] https://doi.org/10.13026/0zgk-9j94

TABLE 5.2. Data statistics

	CADEC	ShARe13	ShARe14
Total Sentences	7,597	18,767	34,618
Total Entities	6,318	11,148	19,073
Continuous Entities	5,639	10,060	17,417
- Percentage	89.25%	90.24%	91.32%
- Number of tokens	1-36	1-9	1-9
Disc. Entities	679	1,088	1,658
- Percentage	10.75%	9.76%	8.68%
- Number of tokens	2-13	2-7	2-7
- Start-End Distance	3-20	3-23	3-23

eHealth Challenge in 2013 and 2014, respectively. They consist of clinical reports annotated for the identification and normalisation of disease disorders. For all datasets, we use the sentence-based, token-level preprocessing script and dataset splits provided by Dai et al. [50] and convert the produced inline format to JSON following Li et al. [119]. A breakdown of relevant statistics for each dataset is presented on Table 5.2 showing CADEC as the smallest dataset with 7,597 sentences while ShARe14 is the largest with 34,618 sentences. All datasets have only around 10% of discontinuous entities.

5.4.2 Baselines and Metrics

We compare our framework with other DNER models. **MAC** [244] first introduced the grid tagging scheme with a segment extractor labelling relative token pairs using the BIS (begin, inside, continuous) scheme and an edge predictor which aligns entity bounds using the *head-to-head* (H2H) and *tail-to-tail* (T2T) tags. **W^{2}NER** [119] introduced a unified NER framework that identifies neighbouring word relationships between non-adjacent entity words using the tags *Next-Neighboring-Word* (NNW) and *Tail-Head-Word* (THW). **TOE** [131] improves upon the W^{2}NER's tagging scheme by adding *Previous-Neighboring-Word* (PNW) and *Head-Tail-Word* (HTW) and incorporating a *Tag Representation Embedding Module* (TREM). **Corro** [48] is a recent model attempting to improve sequence tagging for discontinuous entities through a two-layer tagging system using ten tags. For both W^{2}NER and TOE, we report reproduced results using the published code from each study.

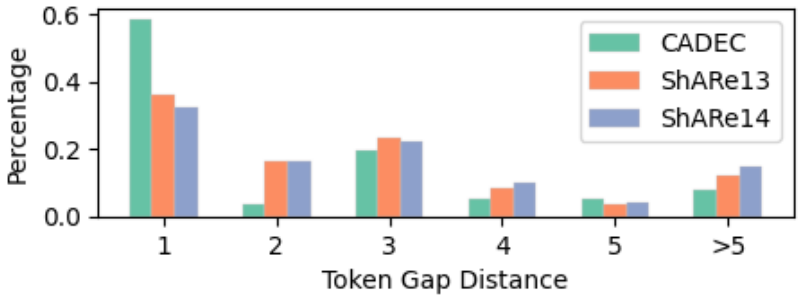


FIGURE 5.5. Distribution of token gaps of discontinuous entities.

Following previous NER studies, we evaluate our framework through exact matching of entities using micro-F1, precision, and recall. We further isolate the effect of our framework on discontinuous entities by reporting F1 scores for sentences with discontinuous entities and for discontinuous entities only (Table 5.3).

5.4.3 Implementation Details

We evaluate our framework using the established training, validation, and test splits by [50]. We list best-performing model setups for each dataset in the Table 5.13. Each model is trained using the AdamW optimiser with a learning rate of 5e-4 for a maximum of 60 epochs and an early stop of 10 epochs. We take the best-performing model on the validation set based on the micro-F1 score. We use a batch size of 12, 6, and 6 for CADEC, ShARe13, and ShARe14, respectively. Our best setup for the CADEC dataset uses a fine-tuned BioBERT, while both ShARe datasets achieve better results with fine-tuned PubMedBERT. More details on these PLMs may be found on Table 2.2. A comparison of PLMs is provided in Table 5.9. All models are trained using an NVIDIA RTX A4500.

5.5 Token Gap Analysis

Figure 5.5 shows the difference in token gaps between CADEC, ShARe13, and ShARe14. CADEC generally shows shorter gaps between spans for discontinuous entities, while the ShARe datasets have wider gaps despite having shorter entities. These differences present

TABLE 5.3. Comparison of performance from our best-performing models for the overall datasets and for discontinuous elements, including sentences containing at least one discontinuous entity (DiscSent) and discontinuous entities only (DiscEnt). **Bold** indicates best scores while underline shows next best.

	Overall			DiscSent	DiscEnt
CADEC	**F1**	**P**	**R**	**F1**	**F1**
MAC [244]	71.50	70.50	72.50	69.80	44.40
W^{2}NER† [119]	72.67	72.02	73.33	69.25	45.78
TOE† [131]	72.24	74.28	70.30	67.98	40.00
Corro [48]	71.90	-	-	-	35.90
Ours	**73.43**	75.35	71.62	**70.59**	**49.71**
ShARe13	**F1**	**P**	**R**	**F1**	**F1**
MAC [244]	81.20	84.30	78.20	68.10	55.90
W^{2}NER† [119]	82.16	84.13	80.29	68.46	57.38
TOE† [131]	81.92	85.05	79.02	67.82	57.06
Corro [48]	82.00	-	-	-	52.10
Ours	**83.22**	86.44	80.24	**69.09**	**60.06**
ShARe14	**F1**	**P**	**R**	**F1**	**F1**
MAC [244]	81.30	78.20	84.70	69.70	54.10
W^{2}NER† [119]	81.31	78.93	83.84	63.08	52.70
TOE† [131]	80.67	78.67	82.78	61.04	49.29
Corro [48]	81.80	-	-	-	49.80
Ours	**82.54**	80.36	84.83	**72.89**	**59.23**

† indicates replicated results.

unique challenges for extracting discontinuous entities in each dataset, highlighting the need for a flexible and adaptable solution like our proposed framework.

5.6 Results

5.6.1 Overall Performance

A comprehensive evaluation of our framework compared to other studies is provided in Table 5.3. The results reflect the performance of our framework on the entire test set, as well as on discontinuous elements, with isolated evaluations on sentences containing at least one discontinuous entity (DiscSent) and on discontinuous entities exclusively (DiscEnt). Our

TABLE 5.4. Complete performance scores from the **best-performing overall model** for sentences with at least one discontinuous entity (DiscSent) and for discontinuous entities only (DiscEnt). **Bold** indicates best scores while underline shows better performance than the best performing baseline scores in Table 5.3.

	Overall			DiscSent			DiscEnt		
Dataset	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**
CADEC	**73.43**	75.35	71.62	70.54	75.52	66.18	48.55	53.16	44.68
ShARe13	**83.22**	86.44	80.24	69.23	79.14	61.53	57.14	71.23	47.71
ShARe14	**82.54**	80.36	84.83	64.82	65.64	64.01	54.40	60.96	49.12

framework demonstrates a clear improvement in both F1 score and precision over W^{2}NER, the best-performing baseline method. The ShARe14 dataset shows the most significant improvement in F1 score, with a 1.23% increase, reaching 82.54. Similarly, the CADEC and ShARe13 datasets show increases of 0.76% (73.43) and 1.06% (83.22), respectively. Furthermore, our framework outperforms the baseline models when focusing on discontinuous elements, with improvements of 0.79%, 0.63%, and 3.19% for DiscSent, and 3.98%, 2.68%, and 5.13% for DiscEnt across the CADEC, ShARe13, and ShARe14 datasets, respectively. Complete performance metrics may be found in Table 5.4. These results underscore the strength of our TriG-NER framework in capturing the complexities of discontinuous entities by leveraging word-pair similarities and dissimilarities. By focusing on token-level relationships within a flexible grid structure, our approach demonstrates superior performance in both overall entity recognition and specifically in handling discontinuous elements, highlighting its adaptability and effectiveness compared to traditional methods.

5.6.2 Triplet Selection

We evaluated the performance of our framework using various triplet selection methods and configuration setups. Table 5.5 shows the performance of our framework under the best-performing model setup for each selection method since different window sizes may affect each method's effectiveness. Among the four strategies, the Centroid strategy consistently shows promising results among the four selection strategies across all datasets, producing

TABLE 5.5. Comparison of different triplet selection methods based on the best-performing setup for each method. **Bold** indicates best scores while underline shows next best. † indicates replicated results from the baseline. HN: Hard Negative; SN: Semi-hard Negative; CE: Centroid; NC: Negative Centroid

Method	CADEC		ShARe13		ShARe14	
	Overall	DiscEnt	Overall	DiscEnt	Overall	DiscEnt
[119]†	72.67	45.75	82.16	**57.38**	81.31	52.70
Hard Negative	71.61	45.41	81.79	54.45	81.87	**57.35**
Soft Negative	72.21	**49.35**	82.56	56.30	82.19	53.79
Centroid	**73.43**	48.55	**83.22**	57.14	82.42	56.22
Negative Centroid	73.33	46.75	82.43	56.22	**82.54**	54.40

TABLE 5.6. Comparison of the anchor-positive pairing and triplet embedding source design setups. **Bold** indicates best scores while underline shows next best.

Setup		CADEC	ShARe13	ShARe14
Pairing	Unique	**73.43**	**83.22**	**82.54**
	Non-unique	71.73	81.82	82.09
Source	Word-Pair Grid ($\mathbf{H}^{bi}$)	71.22	81.19	**82.54**
	Grid tag logits (Y)	**73.43**	**83.22**	82.23

the best scores for overall CADEC and both subsets of ShARe13, while securing the second-best scores for the others. The Negative Centroid strategy also demonstrated encouraging outcomes, having the best score for overall ShARe14 and a competitive second-best for overall CADEC with only a 0.1% disadvantage. On the other hand, the Semi-Negative strategy showed a notably high score for the DiscEnt subset of CADEC. However, it sacrifices overall performance, which falls short of the baseline score, possibly signifying the benefits of a stricter negative candidate selection for the discontinuous entities in the dataset. Similarly, the Hard Negative follows the same trend for ShARe14. Nonetheless, we note that all our triplet selection methods, except Hard Negative, generally outperform and are competitive with the baseline model. This highlights the benefits of leveraging word-pair relationships through our grid-based triplet framework with careful consideration of triplet selection strategies.

In Table 5.6, we compare other design setups for our framework. Using unique anchor-positive pairs through only the top half of the grid sources generally shows superior performance

TABLE 5.7. Complete performance scores from the **best-performing discontinuous entity model** for the overall dataset, for sentences with at least one discontinuous entity (DiscSent), and for discontinuous entities only (DiscEnt). **Bold** indicates best scores while underline shows better performance than the best performing baseline scores in Table 5.3.

	Overall			DiscSent			DiscEnt		
	F1	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**
CADEC	73.22	75.00	71.52	**70.59**	73.81	67.64	**49.71**	54.43	45.74
ShARe13	81.35	85.60	77.50	**69.09**	79.44	61.13	**60.06**	78.52	48.62
ShARe14	82.16	79.78	84.69	**72.89**	74.25	71.59	**59.23**	57.60	60.95

compared to using the entire grid. Utilising only half of the grid lessens uninformative and redundant triplets while also reducing the computational time and resources needed. To highlight the flexibility of our framework, which could be applied to any model with a grid-based component, we further analysed different triplet embedding sources for our framework. Directly applying the triplet loss on the grid tag logits (Y) shows noticeably better performance for CADEC and ShARe13. On the other hand, for ShARe14, the results for both sources are comparable, with a slight improvement from the Word-Pair Relationship Grid ($\mathbf{H}^{bi}$). These findings underscore the effectiveness and versatility of our framework in enhancing discontinuous entity extraction by incorporating word-pair relationships and optimising triplet selection strategies.

5.6.3 Discontinuous Elements Performance

In Table 5.7, we present the performance scores from the model setup that scores highest for the discontinuous entities only (DiscEnt). We observe significantly higher scores for discontinuous entities for the best DiscEnt model with 1.66%, 2.92%, and 4.83% for CADEC, ShARe13, and ShARe14, respectively. However, despite not having the best overall scores in our experiments, the best DiscEnt models still outperform all of the baselines for the CADEC and ShARe14 datasets and are comparable to our overall best model highlighting the ability of our framework to extract discontinuous entities through word-pair triplets.

TABLE 5.8. Comparison of different window sizes. **Bold** indicates best scores while underline shows next best.

Window Size	CADEC	ShARe13	ShARe14
None	71.49	81.74	81.78
1	71.65	81.21	81.91
5	72.77	82.02	**82.54**
10	72.88	**83.22**	81.19
15	70.84	81.26	80.81
20	70.67	81.79	81.33
25	**73.43**	81.83	81.83

5.6.4 Window Size

Given the importance of selecting informative triplets for the triplet loss, we applied a window size centred on the anchor to restrict the positive and negative candidates. In this section, we evaluate the impact of different window sizes on the performance of our best model setups across each dataset.

As shown in Table 5.8, implementing a window significantly improves our framework's performance compared to no window, though the optimal window size varies depending on the dataset. For example, the longer entities in the CADEC dataset benefit from larger window sizes. In contrast, both ShARe datasets achieve optimal performance with smaller window sizes, as the entities in these datasets range from 1 to 9 tokens in length. Removing the window altogether and allowing the framework to select triplets from the entire sequence grid introduces less informative triplets, leading to lower overall performance. Specifically, we observed an improvement of 1.94% for CADEC, 1.48% for ShARe13, and 0.76% for ShARe14.

Our results demonstrate the critical role of window size in enhancing the triplet selection process, ensuring that only the most relevant triplets are used to optimise the learning process. This highlights our framework's adaptability to various dataset characteristics, leading to consistent improvements in performance by effectively leveraging the word-pair relationships within a controlled selection window.

TABLE 5.9. Comparison of different language models used in the encoder with and without our triplet framework based on the best-performing setup for each dataset. **Bold** indicates the overall best scores for each dataset while an underline shows the better score regarding the application of our framework.

PLM	TriG-NER	CADEC	ShARe13	ShARe14
BioBERT [116]	×	72.50	80.25	80.75
	✓	**73.43**	80.72	80.79
BioClinicalBERT [9]	×	71.49	81.78	81.00
	✓	71.42	81.89	80.27
PharmBERT [228]	×	70.78	80.25	80.00
	✓	71.90	80.39	81.11
PubMedBERT [78]	×	70.19	82.00	81.42
	✓	71.39	**83.22**	**82.54**

5.6.5 Encoder Language Models

We evaluated the performance of our framework with different pre-trained language models for the encoder, using the best-performing model setup for each dataset. Table 5.9 presents the results for four biomedical BERT variants, both with and without our grid-based triplet framework. Overall, BioBERT yields the best results for the CADEC dataset, while PubMedBERT outperforms others for both ShARe datasets.

The application of our framework further enhances these scores by 0.93%, 1.22%, and 1.12%, respectively, demonstrating that our framework effectively captures local dependencies via the word-pair triplet implementation. Additionally, our framework consistently improves the performance of most PLMs tested, with the exception of BioClinicalBERT for CADEC and ShARe14.

In Table 5.10, we present the performance improvements achieved by finetuning the pre-trained language models using on a next-word prediction task for each dataset. As expected, finetuning enhances the scores across the board, with more pronounced improvements observed in the ShARe datasets, likely due to the specialised clinical terminology in those datasets compared to the more natural language used in online forums like in CADEC.

TABLE 5.10. Comparison of performance from finetuning the pre-trained language models for the encoder layer. **Bold** indicates best scores while underline shows next best.

Setup	CADEC	ShARe13	ShARe14
Pretrained	72.96	81.35	80.38
Finetuned	**73.43**	**83.22**	**82.54**

TABLE 5.11. Comparison of triplet loss margins. **Bold** indicates best scores while underline shows next best.

Margin	CADEC	ShARe13	ShARe14
0.1	72.58	81.88	82.16
0.5	71.72	81.78	81.86
1	**73.43**	**83.22**	**82.54**
1.5	71.76	81.70	82.18
2	71.41	82.16	80.93

5.6.6 Hyperparameter Testing

We conducted further tests to investigate the impact of different triplet loss margins on the best-performing setup for each dataset. As shown in Table 5.11, using a margin of 1 consistently delivers superior performance across all datasets. In contrast, using a margin of 2 results in a significant performance drop for CADEC and ShARe14, with reductions of 2.02 and 1.61 points, respectively. Similarly, a margin of 1.5 causes a decline of 1.06 points for ShARe13.

These results highlight the sensitivity of our framework to the triplet loss margin and the importance of carefully tuning this hyperparameter. The consistently strong performance with a margin of 1 underscores the robustness of our triplet-based model in capturing word-pair relationships, ensuring optimal performance across different datasets. In Table 5.12, we test different learning rate values for the Adam optimiser and find that the optimal learning rate value for our framework is 5e-04.

Table 5.13 summarises the best hyperparameter combinations for each dataset based on overall performance scores. PubMedBERT provides the best performance for both ShARe datasets while BioBERT works best for CADEC. A centroid triplet selection from the grid tag logits is optimal for CADEC and ShARe13 while a negative centroid selection from the word-pair grid

TABLE 5.12. Comparison of learning rates. **Bold** indicates best scores while underline shows next best.

Learning Rates	CADEC	ShARe13	ShARe14
1e-03	72.40	81.00	81.56
5e-04	**73.43**	**83.22**	**82.54**
3e-04	71.68	81.72	82.08
2e-05	69.53	80.87	81.62

TABLE 5.13. Parameter setup for the best model based on overall performance scores for each dataset.

Setting	CADEC	ShARe13	ShARe14
PLM	BioBERT	PubMedBERT	PubMedBERT
Window Size	25	10	5
Triplet Method	Centroid	Centroid	Neg. Centroid
Learning Rate	5e-04	5e-04	5e-04
Source	Grid Tag Logits	Grid Tag Logits	Word-Pair Grid

is best for ShARe14. The window size varies for each dataset highlighting the importance of tuning these parameters.

5.6.7 Qualitative Analysis

In this section, we demonstrate the effectiveness of our word-pair grid-based triplet framework through a qualitative analysis of the extracted entities, comparing the results with the best-performing baseline model and LLMs, such as Gemini 1.5-flash [220] and GPT-4o [4]. We trained and fine-tuned both our model and the replicated baseline model using tokenised sentences as direct inputs, while the LLMs were not fine-tuned and were provided with task-specific prompts that described the task, input, and expected output format. Table 5.14 provides zero shot and few shot CoT templates. Table 5.15 provides the variables injected in the templates for each dataset. Figure 5.6 presents results for two case studies from CADEC while Figure 5.7 shows a sample from Share13 and Share14 datasets, respectively.

While our framework uses the same tags as W^{2}NER [119], it goes further by leveraging word-pair relationships to accurately recognise multiple non-adjacent entity segments within

TABLE 5.14. Prompt templates used for large language models. One Shot CoT prompt is similar to the Few Shot CoT except that only one example from the training data is provided. Non-CoT prompts remove the last line which asks the LLM to output an explanation.

Prompt Type	Content
Zero Shot CoT	"The task is to find the index of the words from any `{entity_descriptor}` entities from the given text. The text input is already tokenized and is given in a list form where one entry corresponds to a word or punctuation. The word indexes must be based on the list. The entities may be continuous or discontinuous, single-word or multiple words. There may also be no entities in the text. Text: `input` Return the output in a json format followed by a set of steps to explain how the output was generated: "'json [{\"entity\": entity, \"index\":[index1, index2 etc], \"type\": \"`{entity_type}`\"}, {\"entity\": entity, \"index\":[index1, index2, index3 etc], \"type\": \"`{entity_type}`\"}, etc]"' Explanation: explanation"
Few Shot CoT	"The task is to find the index of the words from any `{entity_descriptor}` entities from the given text. The text input is already tokenized and is given in a list form where one entry corresponds to a word or punctuation. The word indexes must be based on the list. The entities may be continuous or discontinuous, single-word or multiple words. There may also be no entities in the text. Below are some examples of input text and output format. Input text: `{input_example_1}` Expected output: `{output_example_1}` Input text: `{input_example_2}` Expected output: `{output_example_2}` Now extract the entities from the text below following the examples above. Text: `{input}` Return the output in a json format followed by a set of steps to explain how the output was generated: "'json [\"entity\": entity, \"index\":[index1, index2 etc], \"type\": \"`{entity_type}`\", \"entity\": entity, \"index\":[index1, index2, index3 etc], \"type\": \"`{entity_type}`\", etc]"' Explanation: explanation"

the input text. In contrast, W^{2}NER processes word pairs in isolation, which limits its ability to recognise entities with more than two disjoint spans, such as "Pain in my lower legs" and "cramping in my lower legs", indexed as [0, 3, 4, 7, 8] and [2, 3, 4, 7, 8], respectively. Furthermore, W^{2}NER struggles to detect uncommon, domain-specific terms and abbreviations, particularly when the entity consists of just one word. For example, in Figure 5.7 (right), our framework successfully extracts the entity "PFO", which stands for "Patent Foramen Ovale", despite the presence of other domain-specific terms. In contrast, W^{2}NER incorrectly extracts "MV", which likely refers to "mitral valve", but is not a disorder.

TABLE 5.15. Variables and examples for each dataset injected in LLM prompts found in Table 5.14.

CADEC	**Value**
`entity_type`	ADR
`entity_descriptor`	adverse drug reaction (ADR)
`input_example_1`	[‘Eczema’, ‘on’, ‘hands’, ‘and’, ‘feet’, ‘,’, ‘rash’, ‘on’, ‘upper’, ‘left’, ‘torso’, ‘,’, ‘depression’, ‘.’]
`output_example_1`	[{‘index’: [0, 1, 4], ‘type’: ‘ADR’}, {‘index’: [0, 1, 2], ‘type’: ‘ADR’}, {‘index’: [6, 7, 8, 9, 10], ‘type’: ‘ADR’}, {‘index’: [12], ‘type’: ‘ADR’}]
`input_example_2`	[‘My’, ‘fingers’, ‘swelled’, ‘up’, ‘and’, ‘hurt’, ‘.’]
`output_example_2`	[{‘index’: [1, 5], ‘type’: ‘ADR’}, {‘index’: [1, 2, 3], ‘type’: ‘ADR’}]
ShARe13	**Value**
`entity_type`	Disorder
`entity_descriptor`	disorder
`input_example_1`	[‘1’, ‘.’, ‘The’, ‘left’, ‘atrium’, ‘is’, ‘mildly’, ‘dilated’, ‘.’, ‘No’, ‘atrial’, ‘septal’, ‘defect’, ‘is’, ‘seen’, ‘by’, ‘2D’, ‘or’, ‘color’, ‘Doppler’, ‘.’]
`output_example_1`	[{‘index’: [3, 4, 7], ‘type’: ’Disorder’}, {’index’: [10, 11, 12], ‘type’: ’Disorder’}]
`input_example_2`	[‘Abd’, ‘:’, ‘She’, ‘had’, ‘an’, ‘ascitic’, ‘abdomen’, ‘that’, ‘was’, ‘very’, ‘large’, ‘,’, ‘round’, ‘,’, ‘and’, ‘soft’, ‘.’]
`output_example_2`	[{‘index’: [5], ‘type’: ‘Disorder’}, {‘index’: [6, 15], ‘type’: ‘Disorder’}]
ShARe14	**Value**
`entity_type`	Disorder
`entity_descriptor`	disorder
`input_example_1`	[‘abd’, ‘soft’, ‘,’, ‘nt’, ‘,’, ‘nd’]
`output_example_1`	[{‘index’: [0, 5], ‘type’: ‘Disorder’}, {‘index’: [0, 3], ‘type’: ‘Disorder’}, {‘index’: [0, 1], ‘type’: ‘Disorder’}]
`input_example_2`	[‘1’, ‘.’, ‘Non’, ‘-’, ‘ST’, ‘-’, ‘elevation’, ‘myocardial’, ‘infarction’, ‘.’]
`output_example_2`	[{‘index’: [2, 3, 4, 5, 6, 7, 8], ‘type’: ‘Disorder’}]

With LLMs’ recent success and popularity for general language generation tasks, we evaluate their performance in extracting entity indexes through zero-shot and few-shot chain-of-thought (CoT) prompting. Because LLMs are optimised for next-word prediction, these models are prone to alignment and indexing problems where, despite clear instructions, the indexes returned do not correspond to the entity words identified. We found that explicitly including the entity words in the return format prompt helps partially but does not entirely resolve the problem. For instance, in Figure 5.6 (right), the entity words "loss of range of motion" are correctly identified; however, the indexes provided are one or two positions off. In some cases, the number of words identified does not equate to the number of indexes returned, such as "*{‘entity’: ‘loss of range of motion’, ‘index’: [32, 36], ‘type’: ‘ADR’}*".

FIGURE 5.6. Case studies for CADEC comparing results from trained models using our framework and a baseline and from zero and few-shot CoT prompt engineering using LLMs. The sample prompt provided follows the few-shot CoT template. All prompt templates are provided in Table 5.14.

Furthermore, LLMs fail to extract discontinuous entities most of the time. In Figure 5.6 (left), both Gemini and GPT-4o completely missed the overlapping continuous and discontinuous entities in the sample despite identifying relevant parts such as "Pain and cramping", "hands", and "lower legs". They cannot effectively split and combine disjoint spans to form discontinuous entities such as "Pain in my hands" and "Pain in my lower legs". GPT-4o Few-shot CoT goes as far as returning the whole input instead of associating the relevant spans together. Lastly, LLMs are prone to extracting entities unrelated to the entity type

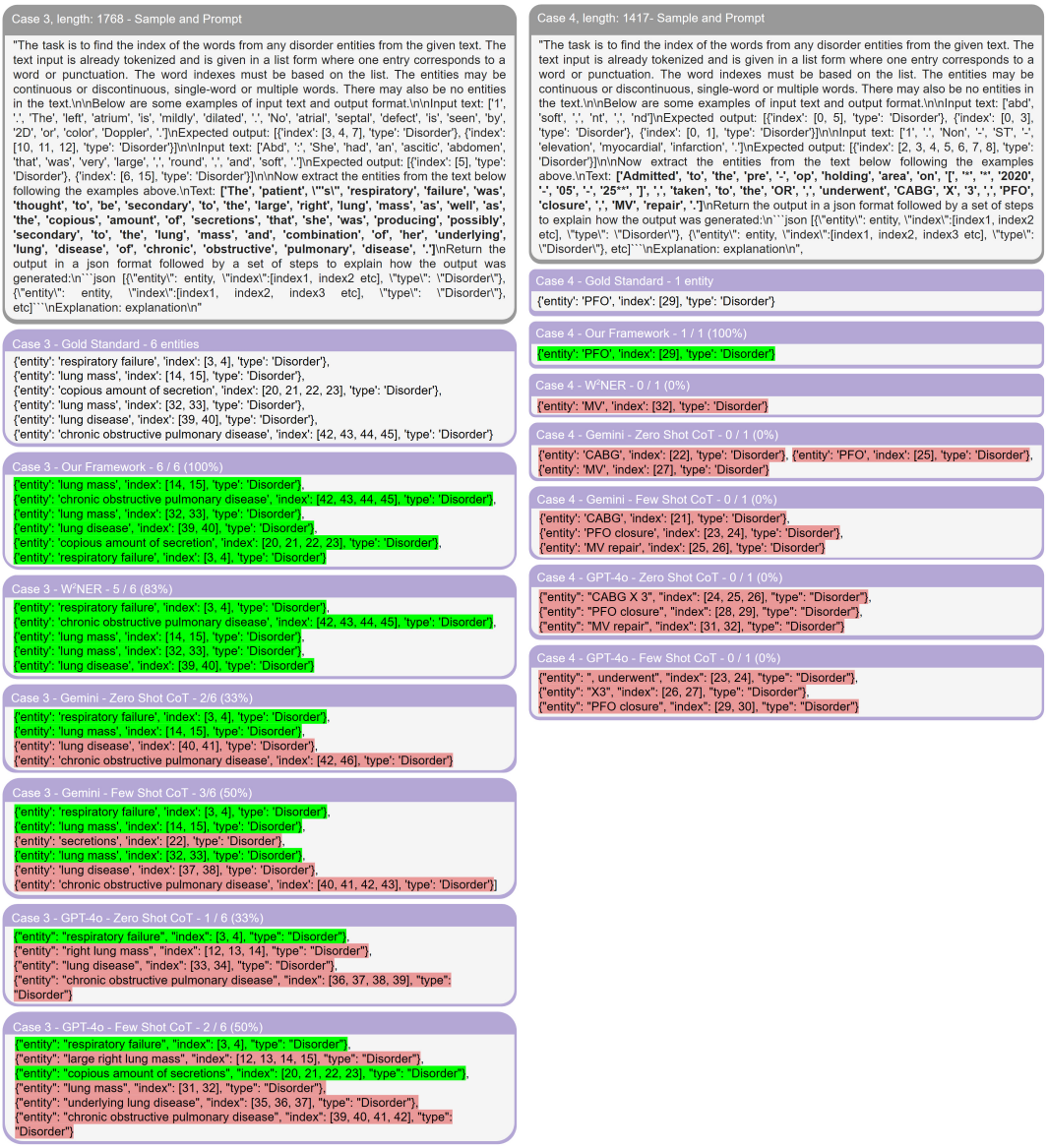


FIGURE 5.7. Case studies for ShARe13 (left) and ShARe14 (right) comparing results from trained models using our framework and a baseline and from zero and few-shot CoT prompt engineering using LLMs. The sample prompt provided follows the few-shot CoT template. All prompt templates are provided in Table 5.14.

provided. For instance, body parts such as "hands" and "lower legs" in Figure 5.6 (left) and medical procedures such as "CABG" (coronary artery bypass graft surgery) in Figure 5.7 (right) are separately identified as ADRs and Disorders, respectively.

While general LLMs have shown significant progress, they still face limitations in specialised tasks like discontinuous entity extraction, unless meticulously designed prompts are used. Trained models continue to outperform current attempts to adopt LLMs for biomedical NER [279]. Our framework, which enhances current trainable DNER models by using token-level, grid-based triplets to account for the similarity and dissimilarity of word pairs, delivers superior performance, especially in handling complex discontinuous entity recognition.

5.7 Conclusion

We introduced TriG-NER, a novel Triplet-Grid Framework designed to improve the extraction of discontinuous named entities by leveraging token-level triplet loss and word-pair relationships. By modelling token pairs within a flexible grid structure, our framework overcomes the limitations of existing tagging schemes, which often struggle to generalise across different datasets.

We evaluated TriG-NER on three benchmark DNER datasets, demonstrating significant improvements over state-of-the-art grid-based architectures. The results validate the effectiveness of our approach in capturing non-adjacent entity segments and underscore the framework's ability to adapt to various tagging schemes, setting a new standard for discontinuous entity extraction. Future work could explore integrating our framework with larger language models and expanding its application to other structured prediction tasks, such as relation extraction and event detection. We hope that our framework, with its innovative grid-based triplet approach, will inspire further research into developing generalisable methods for discontinuous named entity recognition in structured prediction.

This chapter, through TriG-NER, addresses the sequential bias in standard language modelling processes to improve complex entity extraction in free-form, unstructured texts. It addresses the key research questions of this thesis through the introduction of the word-pair modality (RQ1), its integration through a novel grid-based contrastive learning with triplet loss (RQ2), and its application to the extraction of discontinuous adverse drug reaction and disease mentions, fundamental for reliable mental health insights and risk monitoring (RQ3).

The introduced word-pair modality incorporates relational and structural abstractions over the same textual data, complementing and in parallel to conventional semantic abstractions. The grid-based triplet framework further leverages the incorporation of the word-pair modality for discontinuous named entity recognition through simultaneously pulling together representations for similar, co-occurring entity word pairs and pushing apart dissimilar entity and non-entity pairs. The application of this framework to adverse drug events supports the development of improved, reliable, and accurate mental health systems and insights.

CHAPTER 6

Conclusion

This thesis examines various information modalities abstracted from textual data and their multimodal integration to enhance contextualised textual representations for mental healthcare-related tasks. In particular, this research focuses on what abstractive modalities can be derived from textual data (RQ1), how to effectively incorporate different modalities (RQ2), and what mental health-related tasks would benefit from enriched information from text data (RQ3).

To answer these research questions, this thesis first introduced three modalities at three different abstraction levels from the same textual data, namely the emotion modality, the acoustic modality, and the word-pair modality.

The emotion modality is leveraged to explicitly represent affective abstractions derived from textual data, addressing the affective neutrality in the semantic space of language models, particularly for words lacking emotional connotations, thereby improving the representation of implicit emotions. The acoustic modality is introduced to derive paralinguistic and prosodic abstractions from textual data, compensating for the loss of tonal cues in textual data and textual representations. It further complements the emotion modality, enriching the affective knowledge gained from the same textual data. Finally, the word-pair modality is explored to incorporate relational and structural abstractions overlooked by the inherent sequential bias of conventional language modelling, thereby creating more robust representations for complex information extraction tasks.

Further, three different integration methods that leverage each modality's unique information factorisation and inductive biases are explored for mental health-related downstream tasks.

Chapter 3 introduces MM-EMOG, a multi-emotion pretraining framework that incorporates complex human emotions with textual semantics. It leverages the emotion modality through a multi-emotion graph-based pretraining objective, enabling the contextualisation of global and local heterogeneous emotions. It is evaluated through depression and suicide risk detection from social media texts, highlighting the significance of incorporating the emotion modality for emotion-rich mental health detection tasks.

Chapter 4 introduces 3M-Health, a multimodal distillation framework that integrates emotion, acoustic, and semantic modalities for downstream mental health detection tasks. It utilises a multi-teacher knowledge distillation framework that builds modality expertise for each teacher before distilling multimodal knowledge into a text-only student model, thereby enriching the student's semantic representation with multimodal information. The framework demonstrates effective results in social media-based mental health detection tasks, further highlighting the advantages of a multimodal approach over unimodal frameworks.

Chapter 5 introduces TriG-NER, a novel approach to information extraction, particularly for discontinuous entities, incorporating the relational and structural abstractions from the word-pair modality. It enhances textual representations through a grid-based triplet loss approach, which pulls together co-occurring entity word pairs and pushes apart entity and non-entity word pairs. TriG-NER exhibits promising results for the extraction of discontinuous adverse drug reaction entities and diseases, which are typically overlooked and oversimplified, underscoring its potential for more reliable mental health monitoring and insights.

In summary, this thesis explored three modalities from textual data representing different abstraction levels and proposes novel ways of integrating them into text-based mental health downstream tasks. By focusing on ubiquitous texts, each approach remains flexible and scalable even with the integration of other media-based modalities. By treating each abstracted modality as its own representation level, each modality's unique nature is preserved and not lost in the semantic contextualisation of conventional language modelling processes. Finally, while this thesis focuses on the technical feasibility of each proposed approach, each study utilises and compares performance on publicly available real-world data, which future research can easily extended to other mental health-related use cases.

6.1 Future Works

This thesis focused on the multimodality of text-based information and its integration for mental health downstream tasks. Extending research from this work may explore five promising directions.

First is the application of these different textual modalities to other mental health disorders and other mental health-related entities. For disease risk detection, depression and suicidality were of particular interest since they are highly intertwined with negative emotions. However, some mental health disorders do not lean heavily towards negative emotions, such as bipolar disorder, which manifests through extreme mood swings, both positive and negative, and obsessive-compulsive disorders, which may involve more subtle nuances in emotions. For information extraction, drug-related events and disease mentions were primarily investigated; however, other types of entities may also be explored. Treatments, symptoms, stressors, or even the identification and extraction of encounters with mental health professionals could provide clinicians with more information about an individual's mental state. Exploring differences between various mental health diseases and entities could provide fascinating insights into the linguistic nuances for each, just as Chapter 4 showed differences in modality influences between different social media platforms.

Future research may investigate other text-based modalities to further enhance contextualised semantic information from text. Commonsense knowledge, such as the ATOMIC graph [194], has been applied to various downstream tasks, including visual storytelling and general Q&A-based tasks, but has not yet been explored for mental health-related tasks. On the other hand, few researchers have studied causal relations as a separate downstream mental health task [66] but have yet to be incorporated to enhance textual representations in disease detection and information extraction tasks. The addition of these situational-based modalities could prove beneficial to mental health-related tasks to incorporate causes, events, and triggers that affect an individual's mental health state.

With the current prominence of large language models (LLMs), methods of integrating different textual modalities with LLMs must be explored and analysed. Despite appearing

to be capable of reasoning and understanding, LLMs are primarily trained for next-word-prediction with large amounts of data and billions of parameters. Its human-like, cohesive text generation is mainly a process of stringing together the sequence of words with the highest probabilities one after each other. This, however, falls short for nuanced tasks such as mental health detection which involves emotions, motivations, and intents [256, 213]. While LLM-specific prompting techniques, including few-shot prompts, role-based prompts, and chain-of-thought prompts, have made strides for generalised tasks, the incorporation of text-based modalities for mental health related tasks, either through prompts or finetuning methods, have yet to be explored extensively.

With the increase of publicly available multimodal mental health datasets, future research may work towards a general framework for mental health downstream tasks that incorporates different media-based modalities and other abstracted modalities. Generalisability and transferability between data sources must also be explored as current mental health-specific language models are predominantly trained on social media data which is linguistically distinct from clinical data.

Finally, from a broader perspective, a significant challenge in detecting mental health issues on social media texts is the lack of publicly available, clinically validated datasets. As discussed in Chapter 2, these datasets currently dominate text-based data. While social media is a viable and valid data source, the prevailing annotation methods of whether a user is truly diagnosed with a mental health disorder depend heavily on self-disclosure, keyphrase matching, and posting locations (e.g. subreddits or specific topic forums). Even though some datasets employ manual annotation, either by or supervised by experts, these annotation methods mainly assume details regarding an unknown individual's mental state based on the limited information presented to them. Interdisciplinary collaborations between mental health researchers on the medical and technical sides could launch multi-site recruitment studies, much like the RADAR-MDD study [147], with a particular focus on, or at least the inclusion of, de-identified, privacy-preserving social media data.

Recognising the limitations of current multimedia mental health dataset accessibility and the limitations of standard language model semantic representations, this thesis set forth

to exhaust relevant information derived from mental health-related textual data. With the ubiquity of texts in mental health and other related fields, this thesis presents a more accessible and far-reaching approach to mental healthcare. Furthermore, with no dependency on other user metadata, patient information, or other media-based modalities, this thesis demonstrates more reproducible, generalisable, and scalable approaches for more reliable mental health monitoring systems.

Bibliography

[1] Saandeep Aathreya et al. 'Multimodal, context-based dataset of children with Post Traumatic Stress Disorder'. In: *Pattern Recognition Letters* 196 (2025), pp. 228–235. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2025.05.003. URL: https://www.sciencedirect.com/science/article/pii/S0167865525001928.

[2] Muhammad Abdul-Mageed and Lyle Ungar. 'EmoNet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks'. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Regina Barzilay and Min-Yen Kan. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 718–728. DOI: 10.18653/v1/P17-1067. URL: https://aclanthology.org/P17-1067.

[3] Akbobek Abilkaiyrkyzy et al. 'Dialogue System for Early Mental Illness Detection: Toward a Digital Twin Solution'. In: *IEEE Access* 12 (2024), pp. 2007–2024. DOI: 10.1109/ACCESS.2023.3348783.

[4] Josh Achiam et al. 'Gpt-4 technical report'. In: *arXiv preprint arXiv:2303.08774* (2023).

[5] Prakash Adekkanattu et al. 'Deep learning for identifying personal and family history of suicidal thoughts and behaviors from EHRs'. In: *npj Digital Medicine* 7.1 (2024), p. 260. ISSN: 2398-6352. DOI: 10.1038/s41746-024-01266-7. URL: https://doi.org/10.1038/s41746-024-01266-7.

[6] Prottay Kumar Adhikary et al. 'Exploring the Efficacy of Large Language Models in Summarizing Mental Health Counseling Sessions: Benchmark Study'. In: *JMIR Ment Health* 11 (2024), e57306. ISSN: 2368-7959. DOI: 10.2196/57306. URL:

https://mental.jmir.org/2024/1/e57306%20https://doi.org/10.2196/57306.

[7] Aakash Kumar Agarwal et al. 'ReDepress: A Cognitive Framework for Detecting Depression Relapse from Social Media'. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Ed. by Christos Christodoulopoulos et al. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 34652–34670. ISBN: 979-8-89176-332-6. DOI: 10.18653/v1/2025.emnlp-main.1758. URL: https://aclanthology.org/2025.emnlp-main.1758/.

[8] Areej Alhassan et al. 'Discontinuous named entities in clinical text: A systematic literature review'. In: *Journal of Biomedical Informatics* 162 (2025), p. 104783. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2025.104783. URL: https://www.sciencedirect.com/science/article/pii/S1532046425000127.

[9] Emily Alsentzer et al. 'Publicly Available Clinical BERT Embeddings'. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Ed. by Anna Rumshisky et al. Minneapolis, Minnesota, USA: Association for Computational Linguistics, June 2019, pp. 72–78. DOI: 10.18653/v1/W19-1909. URL: https://aclanthology.org/W19-1909.

[10] Gunjan Ansari, Muskan Garg and Chandni Saxena. 'Data Augmentation for Mental Health Classification on Social Media'. In: *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*. National Institute of Technology Silchar, Silchar, India: NLP Association of India (NLPAI), Dec. 2021, pp. 152–161. URL: https://aclanthology.org/2021.icon-main.19.

[11] Ashutosh Anshul et al. '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. DOI: 10.1109/TCSS.2023.3309229.

[12] Junyi Ao et al. 'SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing'. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan,

Preslav Nakov and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 5723–5738. DOI: 10.18653/v1/2022.acl-long.393. URL: https://aclanthology.org/2022.acl-long.393.

[13] Mario Ezra Aragón et al. 'Detecting Depression in Social Media using Fine-Grained Emotions'. 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 1481–1486. DOI: 10.18653/v1/N19-1151. URL: https://aclanthology.org/N19-1151.

[14] Mario Ezra Aragón et al. '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. DOI: 10.1109/TAFFC.2021.3075638.

[15] Mario Ezra Aragón et al. 'DisorBERT: A Double Domain Adaptation Model for Detecting Signs of Mental Disorders in Social Media'. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 15305–15318. DOI: 10.18653/v1/2023.acl-long.853. URL: https://aclanthology.org/2023.acl-long.853/.

[16] Australian Institute of Health and Welfare. *Expenditure on mental health services*. Feb. 2025. URL: https://www.aihw.gov.au/mental-health/topic-areas/facilities-resources/expenditure.

[17] Australian Institute of Health and Welfare. *Mental Health*. URL: https://www.aihw.gov.au/mental-health/overview.

[18] Australian Institute of Health and Welfare. *Mental Health and Substance Abuse*. Aug. 2025. URL: https://www.aihw.gov.au/mental-health/topic-areas/health-wellbeing/mental-illness-and-substance-use.

[19] İrfan Aygün, Buket Kaya and Mehmet Kaya. 'Identifying patients in need of psychological treatment with language representation models'. In: *Multimedia Tools and*

Applications 84.1 (2025), pp. 397–418. ISSN: 1573-7721. DOI: 10.1007/s11042-024-18992-5. URL: https://doi.org/10.1007/s11042-024-18992-5.

[20] Andrew Bailey and Mark D. Plumbley. *Gender Bias in Depression Detection Using Audio Features*. 2021. arXiv: 2010.15120 [cs.SD]. URL: https://arxiv.org/abs/2010.15120.

[21] Shakila Basheer et al. 'Improving mental dysfunction detection from EEG signals: Self-contrastive learning and multitask learning with transformers'. In: *Alexandria Engineering Journal* 106 (2024), pp. 52–59. ISSN: 1110-0168. DOI: https://doi.org/10.1016/j.aej.2024.06.058. URL: https://www.sciencedirect.com/science/article/pii/S1110016824006677.

[22] Iz Beltagy, Matthew E. Peters and Arman Cohan. *Longformer: The Long-Document Transformer*. 2020. arXiv: 2004.05150 [cs.CL]. URL: https://arxiv.org/abs/2004.05150.

[23] Rohit Beniwal and Pavi Saraswat. 'A Hybrid BERT-CNN Approach for Depression Detection on Social Media Using Multimodal Data'. In: *The Computer Journal* 67.7 (Feb. 2024), pp. 2453–2472. ISSN: 0010-4620. DOI: 10.1093/comjnl/bxae018. eprint: https://academic.oup.com/comjnl/article-pdf/67/7/2453/58602510/bxae018.pdf. URL: https://doi.org/10.1093/comjnl/bxae018.

[24] Adrian Benton, Glen Coppersmith and Mark Dredze. 'Ethical Research Protocols for Social Media Health Research'. In: *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Ed. by Dirk Hovy et al. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 94–102. DOI: 10.18653/v1/W17-1612. URL: https://aclanthology.org/W17-1612.

[25] Nicolas Bertagnolli. *Counsel chat: Bootstrapping high-quality therapy data*. 2020.

[26] Yuda Bi et al. 'A multimodal vision transformer for interpretable fusion of functional and structural neuroimaging data'. In: *Human Brain Mapping* 45.17 (2024), e26783. DOI: https://doi.org/10.1002/hbm.26783. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.26783. URL:

https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.26783.

[27] Salah Bouktif, Akib Mohi Ud Din Khanday and Ali Ouni. 'Explainable Predictive Model for Suicidal Ideation During COVID-19: Social Media Discourse Study'. In: *J Med Internet Res* 27 (Jan. 2025), e65434. ISSN: 1438-8871. DOI: 10.2196/65434. URL: https://doi.org/10.2196/65434.

[28] Tom B. Brown et al. 'Language models are few-shot learners'. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Vancouver, BC, Canada: Curran Associates Inc., 2020. ISBN: 9781713829546.

[29] Ana-Maria Bucur et al. 'Datasets for Depression Modeling in Social Media: An Overview'. In: *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*. Ed. by Ayah Zirikly et al. Albuquerque, New Mexico: Association for Computational Linguistics, May 2025, pp. 116–126. ISBN: 979-8-89176-226-8. DOI: 10.18653/v1/2025.clpsych-1.10. URL: https://aclanthology.org/2025.clpsych-1.10/.

[30] Ana-Maria Bucur et al. 'It's Just a Matter of Time: Detecting Depression with Time-Enriched Multimodal Transformers'. In: *Advances in Information Retrieval*. Ed. by Jaap Kamps et al. Cham: Springer Nature Switzerland, 2023, pp. 200–215. ISBN: 978-3-031-28244-7.

[31] Prasadith Buddhitha and Diana Inkpen. 'Multi-task learning to detect suicide ideation and mental disorders among social media users'. In: *Frontiers in Research Metrics and Analytics* 8 (2023). ISSN: 2504-0537. DOI: 10.3389/frma.2023.1152535. URL: https://www.frontiersin.org/articles/10.3389/frma.2023.1152535.

[32] Rina Carines Cabral et al. '3M-Health: Multimodal Multi-Teacher Knowledge Distillation for Mental Health Detection'. In: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. CIKM '24. Boise, ID, USA: Association for Computing Machinery, 2024, pp. 152–162. ISBN: 9798400704369. DOI: 10.1145/3627673.3679635. URL: https://doi.org/10.1145/3627673.3679635.

[33] Rina Carines Cabral et al. ‘MM-EMOG: Multi-Label Emotion Graph Representation for Mental Health Classification on Social Media’. In: *Robotics* 13.3 (2024), p. 53.

[34] Rina Carines Cabral et al. ‘TriG-NER: Triplet-Grid Framework for Discontinuous Named Entity Recognition’. In: *Proceedings of the ACM on Web Conference 2025*. WWW ’25. Sydney NSW, Australia: Association for Computing Machinery, 2025, pp. 2824–2837. ISBN: 9798400712746. DOI: 10.1145/3696410.3714639. URL: https://doi.org/10.1145/3696410.3714639.

[35] Hanshu Cai et al. ‘A multi-modal open dataset for mental-disorder analysis’. In: *Scientific Data* 9.1 (2022), p. 178. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01211-x. URL: https://doi.org/10.1038/s41597-022-01211-x.

[36] Yicheng Cai et al. ‘Depression detection on online social network with multivariate time series feature of user depressive symptoms’. In: *Expert Syst. Appl.* 217.C (May 2023). ISSN: 0957-4174. DOI: 10.1016/j.eswa.2023.119538. URL: https://doi.org/10.1016/j.eswa.2023.119538.

[37] Erik Cambria et al. ‘SenticNet 7: A Commonsense-based Neurosymbolic AI Framework for Explainable Sentiment Analysis’. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, June 2022, pp. 3829–3839. URL: https://aclanthology.org/2022.lrec-1.408.

[38] 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 (2020), pp. 87–102. DOI: 10.1109/TMM.2020.3046867.

[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. DOI: 10.1109/TMM.2020.3046867.

[40] Lei Cao et al. ‘Latent Suicide Risk Detection on Microblog via Suicide-Oriented Word Embeddings and Layered Attention’. 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 1718–1728. DOI:

10.18653/v1/D19-1181. URL: https://aclanthology.org/D19-1181.

[41] Nicholas C Cardamone et al. 'Classifying Unstructured Text in Electronic Health Records for Mental Health Prediction Models: Large Language Model Evaluation Study'. In: *JMIR Med Inform* 13 (Jan. 2025), e65454. ISSN: 2291-9694. DOI: 10.2196/65454. URL: https://doi.org/10.2196/65454.

[42] Defang Chen et al. 'Cross-layer distillation with semantic calibration'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 8. 2021, pp. 7028–7036.

[43] Zhuang Chen et al. 'Depression Detection in Clinical Interviews with LLM-Empowered Structural Element Graph'. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by Kevin Duh, Helena Gomez and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 8181–8194. DOI: 10.18653/v1/2024.naacl-long.452. URL: https://aclanthology.org/2024.naacl-long.452/.

[44] Chun Yueh Chiu et al. 'Multimodal depression detection on instagram considering time interval of posts'. In: *Journal of Intelligent Information Systems* 56 (2021), pp. 25–47.

[45] Kenneth Ward Church and Patrick Hanks. 'Word Association Norms, Mutual Information, and Lexicography'. In: *Computational Linguistics* 16.1 (1990), pp. 22–29. URL: https://aclanthology.org/J90-1003.

[46] Hanne K Collins et al. 'Conveying and detecting listening during live conversation.' In: *Journal of Experimental Psychology: General* (2023).

[47] Glen Coppersmith et al. '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/.

[48] Caio Corro. *A fast and sound tagging method for discontinuous named-entity recognition*. 2024. arXiv: 2409.16243 [cs.CL]. URL: https://arxiv.org/abs/2409.16243.

[49] Hong-Jie Dai et al. 'Deep Learning-Based Natural Language Processing for Screening Psychiatric Patients'. In: *Frontiers in Psychiatry* Volume 11 - 2020 (2021). ISSN: 1664-0640. DOI: 10.3389/fpsyt.2020.533949. URL: https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2020.533949.

[50] Xiang Dai et al. 'An Effective Transition-based Model for Discontinuous NER'. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, July 2020, pp. 5860–5870. DOI: 10.18653/v1/2020.acl-main.520. URL: https://aclanthology.org/2020.acl-main.520.

[51] Ruochen Dang et al. 'Classification of schizophrenia based on RAnet-ET: resnet based attention network for eye-tracking'. In: *Journal of Neural Engineering* 22.2 (Apr. 2025), p. 026053. DOI: 10.1088/1741-2552/adc5a5. URL: https://doi.org/10.1088/1741-2552/adc5a5.

[52] Sagnik de et al. 'SLiTRANet: An EEG-Based Automated Diagnosis Framework for Major Depressive Disorder Monitoring Using a Novel LGCN and Transformer-Based Hybrid Deep Learning Approach'. In: *IEEE Access* 12 (2024), pp. 173109–173126. DOI: 10.1109/ACCESS.2024.3493140.

[53] 'Deciphering language disturbances in schizophrenia: A study using fine-tuned language models'. In: *Schizophrenia Research* 271 (2024), pp. 120–128. ISSN: 0920-9964. DOI: https://doi.org/10.1016/j.schres.2024.07.016. URL: https://www.sciencedirect.com/science/article/pii/S0920996424003219.

[54] David DeVault et al. 'SimSensei kiosk: a virtual human interviewer for healthcare decision support'. In: *Proceedings of the 2014 International Conference on Autonomous*

Agents and Multi-Agent Systems. AAMAS '14. Paris, France: International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 1061–1068. ISBN: 9781450327381.

[55] Jacob Devlin et al. 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. 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. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[56] Jacob Devlin et al. 'Bert: Pre-training of deep bidirectional transformers for language understanding'. In: *arXiv preprint arXiv:1810.04805* (2018).

[57] Hanneke van Dijk et al. 'The two decades brainclinics research archive for insights in neurophysiology (TDBRAIN) database'. In: *Scientific Data* 9.1 (2022), p. 333. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01409-z. URL: https://doi.org/10.1038/s41597-022-01409-z.

[58] Sri Harsha Dumpala et al. 'On combining global and localized self-supervised models of speech'. In: *Interspeech 2022*. ISCA: ISCA, Sept. 2022, pp. 3593–3597.

[59] Zachary Englhardt et al. 'From Classification to Clinical Insights: Towards Analyzing and Reasoning About Mobile and Behavioral Health Data With Large Language Models'. In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8.2 (May 2024). DOI: 10.1145/3659604. URL: https://doi.org/10.1145/3659604.

[60] Hao Fei et al. 'Rethinking boundaries: End-to-end recognition of discontinuous mentions with pointer networks'. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 14. 2021, pp. 12785–12793.

[61] Wei Feng et al. 'Deep learning based prediction of depression and anxiety in patients with type 2 diabetes mellitus using regional electronic health records'. In: *International Journal of Medical Informatics* 196 (2025), p. 105801. ISSN: 1386-5056. DOI: https://doi.org/10.1016/j.ijmedinf.2025.105801. URL: https://www.sciencedirect.com/science/article/pii/S1386505625000188.

[62] I. Fernández-Barrera, S. Bravo-Bustos and M. Vidal. 'Evaluating the Social Media Users' Mental Health Status During COVID-19 Pandemic Using Deep Learning'. In: *International Conference on Biomedical and Health Informatics 2022*. Ed. by Esteban Pino, Ratko Magjarević and Paulo de Carvalho. Cham: Springer Nature Switzerland, 2024, pp. 60–68. ISBN: 978-3-031-59216-4.

[63] Saadia Gabriel et al. 'Can AI Relate: Testing Large Language Model Response for Mental Health Support'. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Ed. by Yaser Al-Onaizan, Mohit Bansal and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 2206–2221. DOI: 10.18653/v1/2024.findings-emnlp.120. URL: https://aclanthology.org/2024.findings-emnlp.120/.

[64] Muskan Garg. 'Mental health analysis in social media posts: a survey'. In: *Archives of Computational Methods in Engineering* 30.3 (2023), pp. 1819–1842.

[65] Muskan Garg. 'Multi-class categorization of reasons behind mental disturbance in long texts'. In: *Knowledge-Based Systems* 276 (2023), p. 110742. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2023.110742. URL: https://www.sciencedirect.com/science/article/pii/S0950705123004926.

[66] Muskan Garg et al. 'CAMS: An Annotated Corpus for Causal Analysis of Mental Health Issues in Social Media Posts'. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, June 2022, pp. 6387–6396. URL: https://aclanthology.org/2022.lrec-1.686/.

[67] Manas Gaur et al. 'Knowledge-Aware Assessment of Severity of Suicide Risk for Early Intervention'. In: *The World Wide Web Conference*. WWW '19. San Francisco, CA, USA: Association for Computing Machinery, 2019, pp. 514–525. ISBN: 9781450366748. DOI: 10.1145/3308558.3313698. URL: https://doi.org/10.1145/3308558.3313698.

[68] Adhiraj Ghosh, Kuruparan Shanmugalingam and Wen-Yan Lin. 'Relation Preserving Triplet Mining for Stabilising the Triplet Loss In re-Identification Systems'. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Jan. 2023, pp. 4840–4849.

[69] Baraa Abou Ghouch and Wael Khreich. 'Reducing false alarms by identifying depression-mimicking expressions'. In: *Neural Computing and Applications* 37.29 (2025), pp. 23863–23882. ISSN: 1433-3058. DOI: 10.1007/s00521-025-11543-5. URL: https://doi.org/10.1007/s00521-025-11543-5.

[70] Sujatha Gollapalli, Beng Ang and See-Kiong Ng. 'Identifying Early Maladaptive Schemas from Mental Health Question Texts'. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 11832–11843. DOI: 10.18653/v1/2023.findings-emnlp.792. URL: https://aclanthology.org/2023.findings-emnlp.792/.

[71] Randy L. Gollub et al. 'The MCIC Collection: A Shared Repository of Multi-Modal, Multi-Site Brain Image Data from a Clinical Investigation of Schizophrenia'. In: *Neuroinformatics* 11.3 (2013), pp. 367–388. ISSN: 1559-0089. DOI: 10.1007/s12021-013-9184-3. URL: https://doi.org/10.1007/s12021-013-9184-3.

[72] Yuan Gong, Yu-An Chung and James Glass. 'AST: Audio Spectrogram Transformer'. In: *Proc. Interspeech 2021*. 2021, pp. 571–575. DOI: 10.21437/Interspeech.2021-698.

[73] Neha Gour et al. 'Transformers for autonomous recognition of psychiatric dysfunction via raw and imbalanced EEG signals'. In: *Brain Informatics* 10.1 (2023), p. 25. ISSN: 2198-4026. DOI: 10.1186/s40708-023-00201-y. URL: https://doi.org/10.1186/s40708-023-00201-y.

[74] Shanky Goyal et al. 'MindLift: AI-powered mental health assessment for students'. In: *Neuroscience Informatics* 5.2 (2025), p. 100208. ISSN: 2772-5286. DOI: https://doi.org/10.1016/j.neuri.2025.100208. URL: https://www.sciencedirect.com/science/article/pii/S2772528625000238.

[75] Jonathan Gratch et al. 'The Distress Analysis Interview Corpus of human and computer interviews'. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Ed. by Nicoletta Calzolari et al. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 3123–3128. URL: https://aclanthology.org/L14-1421/.

[76] Garrett Greiner and Yu Zhang. 'Multi-modal EEG NEO-FFI with Trained Attention Layer (MENTAL) for mental disorder prediction'. In: *Brain Informatics* 11.1 (2024), p. 26. ISSN: 2198-4026. DOI: 10.1186/s40708-024-00240-z. URL: https://doi.org/10.1186/s40708-024-00240-z.

[77] Dongxiao Gu et al. 'An analysis of cognitive change in online mental health communities: A textual data analysis based on post replies of support seekers'. In: *Information Processing Management* 60.2 (2023), p. 103192. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2022.103192. URL: https://www.sciencedirect.com/science/article/pii/S030645732200293X.

[78] Yu Gu et al. 'Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing'. In: *ACM Trans. Comput. Healthcare* 3.1 (Oct. 2021). DOI: 10.1145/3458754. URL: https://doi.org/10.1145/3458754.

[79] Anup Kumar Gupta, Ashutosh Dhamaniya and Puneet Gupta. 'RADIANCE: Reliable and interpretable depression detection from speech using transformer'. In: *Computers in Biology and Medicine* 183 (2024), p. 109325. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2024.109325. URL: https://www.sciencedirect.com/science/article/pii/S0010482524014100.

[80] Faizal Hajamohideen et al. 'Four-way classification of Alzheimer's disease using deep Siamese convolutional neural network with triplet-loss function'. In: *Brain Informatics* 10.1 (2023), p. 5.

[81] Caren Han et al. 'VICTR: Visual Information Captured Text Representation for Text-to-Vision Multimodal Tasks'. In: *Proceedings of the 28th International Conference on Computational Linguistics*. 2020, pp. 3107–3117.

[82] Soyeon Caren Han et al. 'Understanding Graph Convolutional Networks for Text Classification'. In: *arXiv preprint arXiv:2203.16060* (2022).

[83] Shijie Hao et al. 'Interview-based Depression Detection Using LLM-based Text Restatement and Emotion Lexicon'. In: *IEEE Transactions on Affective Computing* (2025), pp. 1–15. DOI: 10.1109/TAFFC.2025.3624419.

[84] Ayaan Haque, Viraaj Reddi and Tyler Giallanza. 'Deep Learning for Suicide and Depression Identification with Unsupervised Label Correction'. In: *Artificial Neural Networks and Machine Learning – ICANN 2021*. Ed. by Igor Farkaš et al. Cham: Springer International Publishing, 2021, pp. 436–447. ISBN: 978-3-030-86383-8.

[85] Xinwei He et al. 'Triplet-Center Loss for Multi-View 3D Object Retrieval'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018.

[86] Amey Hengle et al. 'Still Not Quite There! Evaluating Large Language Models for Comorbid Mental Health Diagnosis'. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 16698–16721. DOI: 10.18653/v1/2024.emnlp-main.931. URL: https://aclanthology.org/2024.emnlp-main.931/.

[87] Geoffrey Hinton, Oriol Vinyals and Jeff Dean. 'Distilling the knowledge in a neural network'. In: *arXiv preprint arXiv:1503.02531* (2015).

[88] Anne de Hond et al. 'Predicting Depression Risk in Patients With Cancer Using Multimodal Data: Algorithm Development Study'. In: *JMIR Med Inform* 12 (Jan. 2024), e51925. ISSN: 2291-9694. DOI: 10.2196/51925. URL: https://doi.org/10.2196/51925.

[89] Derek Howard et al. 'Transfer Learning for Risk Classification of Social Media Posts: Model Evaluation Study'. In: *J Med Internet Res* 22.5 (May 2020), e15371. ISSN: 1438-8871. DOI: 10.2196/15371. URL: https://doi.org/10.2196/15371.

[90] Peixin Huang et al. 'T 2 -NER: A Two-Stage Span-Based Framework for Unified Named Entity Recognition with Templates'. In: *Transactions of the Association for Computational Linguistics* 11 (2023), pp. 1265–1282. DOI: 10.1162/tacl_a_00602. URL: https://aclanthology.org/2023.tacl-1.72.

[91] Jihyun K. Hur et al. 'Language sentiment predicts changes in depressive symptoms'. In: *Proceedings of the National Academy of Sciences* 121.39 (2024), e2321321121. DOI: 10.1073/pnas.2321321121. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2321321121. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2321321121.

[92] Keith Ito and Linda Johnson. *The LJ Speech Dataset*. https://keithito.com/LJ-Speech-Dataset/. 2017.

[93] Ravi Iyer, Maja Nedeljkovic and Denny Meyer. 'Using Vocal Characteristics To Classify Psychological Distress in Adult Helpline Callers: Retrospective Observational Study'. In: *JMIR Formative Research* 6.12 (Dec. 2022), e42249. DOI: 10.2196/42249. URL: https://doi.org/10.2196/42249.

[94] Ali Akbar Jamali, Corinne Berger and Raymond J Spiteri. 'Momentary Depressive Feeling Detection Using X (Formerly Twitter) Data: Contextual Language Approach'. In: *JMIR AI* 2 (Nov. 2023), e49531. ISSN: 2817-1705. DOI: 10.2196/49531. URL: https://doi.org/10.2196/49531.

[95] Sadari Jayawardena, Julien Epps and Eliathamby Ambikairajah. 'Ordinal Logistic Regression With Partial Proportional Odds for Depression Prediction'. In: *IEEE Transactions on Affective Computing* 14.1 (2023), pp. 563–577. DOI: 10.1109/TAFFC.2020.3031300.

[96] Shaoxiong Ji et al. *Domain-specific Continued Pretraining of Language Models for Capturing Long Context in Mental Health*. 2023. arXiv: 2304.10447 [cs.CL]. URL: https://arxiv.org/abs/2304.10447.

[97] Shaoxiong Ji et al. 'MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare'. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, June 2022, pp. 7184–7190. URL: https://aclanthology.org/2022.lrec-1.778.

[98] Shaoxiong Ji et al. 'Mentalbert: Publicly available pretrained language models for mental healthcare'. In: *arXiv preprint arXiv:2110.15621* (2021).

[99] Shaoxiong Ji et al. 'Suicidal ideation and mental disorder detection with attentive relation networks'. en. In: *Neural Comput. Appl.* (June 2021).

[100] Xiaoqi Jiao et al. 'TinyBERT: Distilling BERT for Natural Language Understanding'. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Ed. by Trevor Cohn, Yulan He and Yang Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 4163–4174. DOI: 10.18653/v1/2020.findings-emnlp.372. URL: https://aclanthology.org/2020.findings-emnlp.372.

[101] Alistair E. W. Johnson et al. 'MIMIC-III, a freely accessible critical care database'. In: *Scientific Data* 3.1 (2016), p. 160035. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.35. URL: https://doi.org/10.1038/sdata.2016.35.

[102] Alistair E. W. Johnson et al. 'MIMIC-IV, a freely accessible electronic health record dataset'. In: *Scientific Data* 10.1 (2023), p. 1. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01899-x. URL: https://doi.org/10.1038/s41597-022-01899-x.

[103] Mohsinul Kabir et al. 'DEPTWEET: A typology for social media texts to detect depression severities'. In: *Computers in Human Behavior* 139 (2023), p. 107503. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2022.107503. URL: https://www.sciencedirect.com/science/article/pii/S0747563222003235.

[104] Boyoung Kang and Munpyo Hong. 'Development and Evaluation of a Mental Health Chatbot Using ChatGPT 4.0: Mixed Methods User Experience Study With Korean Users'. In: *JMIR Med Inform* 13 (Jan. 2025), e63538. ISSN: 2291-9694. DOI: 10.2196/63538. URL: https://doi.org/10.2196/63538.

[105] Migyeong Kang et al. 'CURE: Context- and Uncertainty-Aware Mental Disorder Detection'. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 17924–17940. DOI: 10.18653/v1/2024.emnlp-main.994. URL: https://aclanthology.org/2024.emnlp-main.994/.

[106] Karina Karapetian et al. 'Supervised Relation Extraction Between Suicide-Related Entities and Drugs: Development and Usability Study of an Annotated PubMed Corpus'. In: *J Med Internet Res* 25 (Mar. 2023), e41100. ISSN: 1438-8871. DOI: 10.2196/41100. URL: https://doi.org/10.2196/41100.

[107] Sarvnaz Karimi et al. 'Cadec: A corpus of adverse drug event annotations'. In: *Journal of Biomedical Informatics* 55 (2015), pp. 73–81. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2015.03.010. URL: https://www.sciencedirect.com/science/article/pii/S1532046415000532.

[108] Parisa Khodabakhshi, Masoud Mahootchi and Hadi Mosadegh. 'A novel hybrid clustering and classification framework for multi-level depression detection in social media texts'. In: *Engineering Applications of Artificial Intelligence* 160 (2025), p. 111952. ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2025.111952. URL: https://www.sciencedirect.com/science/article/pii/S0952197625019608.

[109] Byeongchang Kim, Hyunwoo Kim and Gunhee Kim. 'Abstractive Summarization of Reddit Posts with Multi-level Memory Networks'. 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. 2519–2531. DOI: 10.18653/v1/N19-1260. URL: https://aclanthology.org/N19-1260/.

[110] Thomas N Kipf and Max Welling. 'Semi-supervised classification with graph convolutional networks'. In: *arXiv preprint arXiv:1609.02907* (2016). URL: https://doi.org/10.48550/arXiv.1609.02907.

[111] Kurt Kroenke et al. 'The PHQ-8 as a measure of current depression in the general population'. In: *Journal of Affective Disorders* 114.1 (2009), pp. 163–173. ISSN: 0165-0327. DOI: https://doi.org/10.1016/j.jad.2008.06.026. URL: https://www.sciencedirect.com/science/article/pii/S0165032708002826.

[112] Puneet Kumar et al. 'Multimodal Interpretable Depression Analysis Using Visual, Physiological, Audio and Textual Data'. In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2025, pp. 5305–5315. DOI: 10.1109/WACV61041.2025.00518.

[113] Juan S. Lara et al. 'Deep Bag-of-Sub-Emotions for Depression Detection in Social Media'. In: *Text, Speech, and Dialogue: 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings*. Olomouc, Czech Republic: Springer-Verlag, 2021, pp. 60–72. ISBN: 978-3-030-83526-2. DOI: 10.1007/978-3-030-83527-9_5. URL: https://doi.org/10.1007/978-3-030-83527-9_5.

[114] Erik Larsen et al. 'Validating the efficacy and value proposition of mental fitness vocal biomarkers in a psychiatric population: prospective cohort study'. In: *Frontiers in Psychiatry* Volume 15 - 2024 (2024). ISSN: 1664-0640. DOI: 10.3389/fpsyt.2024.1342835. URL: https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2024.1342835.

[115] Clinton Lau, Xiaodan Zhu and Wai-Yip Chan. 'Automatic depression severity assessment with deep learning using parameter-efficient tuning'. en. In: *Front. Psychiatry* 14 (June 2023), p. 1160291.

[116] Jinhyuk Lee et al. 'BioBERT: a pre-trained biomedical language representation model for biomedical text mining'. In: *Bioinformatics* 36.4 (Sept. 2019), pp. 1234–1240. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz682. eprint: https://academic.oup.com/bioinformatics/article-pdf/36/4/1234/48983216/bioinformatics_36_4_1234.pdf. URL: https://doi.org/10.1093/bioinformatics/btz682.

[117] Seungeun Lee et al. 'Multimodal integration of neuroimaging and genetic data for the diagnosis of mood disorders based on computer vision models'. In: *Journal of Psychiatric Research* 172 (2024), pp. 144–155. ISSN: 0022-3956. DOI: https://doi.org/10.1016/j.jpsychires.2024.02.036. URL: https://www.sciencedirect.com/science/article/pii/S0022395624001006.

[118] Fei Li et al. 'A Span-Based Model for Joint Overlapped and Discontinuous Named Entity Recognition'. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong et al. Online: Association for Computational Linguistics, Aug. 2021, pp. 4814–4828. DOI: 10.18653/v1/2021.acl-long.372. URL: https://aclanthology.org/2021.acl-long.372.

[119] Jingye Li et al. 'Unified Named Entity Recognition as Word-Word Relation Classification'. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.10 (June 2022), pp. 10965–10973. DOI: 10.1609/aaai.v36i10.21344. URL: https://ojs.aaai.org/index.php/AAAI/article/view/21344.

[120] Kang Li et al. 'Towards cross-modality medical image segmentation with online mutual knowledge distillation'. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 01. 2020, pp. 775–783.

[121] Qiong Li et al. 'Spatio-temporal Multi-granularity for Skeleton-based Depression Risk Recognition'. In: *IEEE Journal of Biomedical and Health Informatics* (2025), pp. 1–12. DOI: 10.1109/JBHI.2025.3587401.

[122] Wentao Li et al. 'Simple action for depression detection: using kinect-recorded human kinematic skeletal data'. In: *BMC Psychiatry* 21.1 (2021), p. 205. ISSN: 1471-244X. DOI: 10.1186/s12888-021-03184-4. URL: https://doi.org/10.1186/s12888-021-03184-4.

[123] Wenyu Li et al. 'Zero-shot Explainable Mental Health Analysis on Social Media by Incorporating Mental Scales'. In: *Companion Proceedings of the ACM Web Conference 2024*. WWW '24. Singapore, Singapore: Association for Computing Machinery, 2024, pp. 959–962. ISBN: 9798400701726. DOI: 10.1145/3589335.3651584. URL: https://doi.org/10.1145/3589335.3651584.

[124] Xin Li et al. 'EMo Transformer: Transformer-Based Depression Detection via Eye Movements'. In: *2024 IEEE International Conference on Multimedia and Expo (ICME)*. 2024, pp. 1–6. DOI: 10.1109/ICME57554.2024.10688241.

[125] Yichun Li, Rajesh Nair and Syed Mohsen Naqvi. 'Video-Based Skeleton Data Analysis for ADHD Detection'. In: *2023 IEEE Symposium Series on Computational Intelligence (SSCI)*. 2023, pp. 1–6. DOI: 10.1109/SSCI52147.2023.10372062.

[126] Yutong Li et al. 'A Visually Interpretable Convolutional-Transformer Model for Assessing Depression from Facial Images'. In: *2023 IEEE International Conference on Multimedia and Expo (ICME)*. 2023, pp. 252–257. DOI: 10.1109/ICME55011.2023.00051.

[127] Bin Liang et al. 'Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks'. In: *Knowledge-Based Systems* 235 (2022), p. 107643. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2021.107643. URL: https://www.sciencedirect.com/science/article/pii/S0950705121009059.

[128] Chenhao Lin et al. 'SenseMood: Depression Detection on Social Media'. In: *Proceedings of the 2020 International Conference on Multimedia Retrieval*. ICMR '20. Dublin, Ireland: Association for Computing Machinery, 2020, pp. 407–411. ISBN: 9781450370875. DOI: 10.1145/3372278.3391932. URL: https://doi.org/10.1145/3372278.3391932.

[129] Kaiying Lin et al. 'Aiding Large Language Models Using Clinical Scoresheets for Neurobehavioral Diagnostic Classification From Text: Algorithm Development and Validation'. In: *JMIR AI* 4 (Oct. 2025), e75030. ISSN: 2817-1705. DOI: 10.2196/75030. URL: https://doi.org/10.2196/75030.

[130] Hu Linmei et al. 'Heterogeneous Graph Attention Networks for Semi-supervised Short Text Classification'. 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4821–4830. DOI: 10.18653/v1/D19-1488. URL: https://aclanthology.org/D19-1488.

[131] Jiang Liu et al. 'TOE: A Grid-Tagging Discontinuous NER Model Enhanced by Embedding Tag/Word Relations and More Fine-Grained Tags'. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), pp. 177–187. DOI: 10.1109/TASLP.2022.3221009.

[132] Tengfei Liu et al. 'Hierarchical Graph Convolutional Networks for Structured Long Document Classification'. In: *IEEE Transactions on Neural Networks and Learning Systems* (2022), pp. 1–15. DOI: 10.1109/TNNLS.2022.3185295.

[133] Xien Liu et al. 'Tensor Graph Convolutional Networks for Text Classification'. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05 (Apr. 2020), pp. 8409–8416. DOI: 10.1609/aaai.v34i05.6359. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6359.

[134] Yang Liu et al. 'Leveraging ChatGPT to optimize depression intervention through explainable deep learning'. In: *Frontiers in Psychiatry* Volume 15 - 2024 (2024). ISSN: 1664-0640. DOI: 10.3389/fpsyt.2024.1383648. URL: https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2024.1383648.

[135] Yinhan Liu et al. 'Roberta: A robustly optimized bert pretraining approach'. In: *arXiv preprint arXiv:1907.11692* (2019). DOI: 10.48550/arXiv.1907.11692.

[136] Zhaolu Liu et al. 'Listening to Mental Health Crisis Needs at Scale: Using Natural Language Processing to Understand and Evaluate a Mental Health Crisis Text Messaging Service'. In: *Frontiers in Digital Health* Volume 3 - 2021 (2021). ISSN: 2673-253X. DOI: 10.3389/fdgth.2021.779091. URL: https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2021.779091.

[137] Zixuan Liu, Guofang Zhang and Yanguang Shen. 'Psychomedical named entity recognition method based on multi-level feature extraction and multi-granularity embedding fusion'. In: *Scientific Reports* 15.1 (2025), p. 16927. ISSN: 2045-2322. DOI: 10.1038/s41598-025-90939-8. URL: https://doi.org/10.1038/s41598-025-90939-8.

[138] Siqu Long et al. *A Quantitative and Qualitative Analysis of Suicide Ideation Detection using Deep Learning*. 2022. DOI: 10.48550/ARXIV.2206.08673. URL: https://arxiv.org/abs/2206.08673.

[139] Isabelle Lorge et al. 'Detecting the clinical features of difficult-to-treat depression using synthetic data from large language models'. In: *Computers in Biology and Medicine* 194 (2025), p. 110246. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2025.110246. URL: https://www.sciencedirect.com/science/article/pii/S0010482525005979.

[140] David E. Losada and Fabio Crestani. 'A Test Collection for Research on Depression and Language Use'. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Ed. by Norbert Fuhr et al. Cham: Springer International Publishing, 2016, pp. 28–39. ISBN: 978-3-319-44564-9.

[141] Kuan-Chieh Lu, Syauki Aulia Thamrin and Arbee L. P. Chen. 'Depression detection via conversation turn classification'. In: *Multimedia Tools and Applications* 82.25 (2023), pp. 39393–39413. ISSN: 1573-7721. DOI: 10.1007/s11042-023-15103-8. URL: https://doi.org/10.1007/s11042-023-15103-8.

[142] Sean MacAvaney et al. '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*. 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.

[143] Paulo Mann, Aline Paes and Elton H. Matsushima. 'See and Read: Detecting Depression Symptoms in Higher Education Students Using Multimodal Social Media Data'. In: *Proceedings of the International AAAI Conference on Web and Social Media* 14.1 (May 2020), pp. 440–451. DOI: 10.1609/icwsm.v14i1.7313. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/7313.

[144] Tingyun Mao et al. 'A simple but effective span-level tagging method for discontinuous named entity recognition'. In: *Neural Computing and Applications* 36.13 (2024),

pp. 7187–7201. ISSN: 1433-3058. DOI: 10.1007/s00521-024-09454-y. URL: https://doi.org/10.1007/s00521-024-09454-y.

[145] Juan Martinez-Romo, Lourdes Araujo and Blanca Reneses. 'Guardian-BERT: Early detection of self-injury and suicidal signs with language technologies in electronic health reports'. In: *Computers in Biology and Medicine* 186 (2025), p. 109701. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2025.109701. URL: https://www.sciencedirect.com/science/article/pii/S0010482525000514.

[146] Gazi Hasan Al Masud et al. 'Effective depression detection and interpretation: Integrating machine learning, deep learning, language models, and explainable AI'. In: *Array* 25 (2025), p. 100375. ISSN: 2590-0056. DOI: https://doi.org/10.1016/j.array.2025.100375. URL: https://www.sciencedirect.com/science/article/pii/S2590005625000025.

[147] F. Matcham et al. 'Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): a multi-centre prospective cohort study protocol'. In: *BMC Psychiatry* 19.1 (2019), p. 72. ISSN: 1471-244X. DOI: 10.1186/s12888-019-2049-z. URL: https://doi.org/10.1186/s12888-019-2049-z.

[148] Puneet Mathur et al. 'Utilizing Temporal Psycholinguistic Cues for Suicidal Intent Estimation'. In: *Advances in Information Retrieval*. Ed. by Joemon M. Jose et al. Cham: Springer International Publishing, 2020, pp. 265–271. ISBN: 978-3-030-45442-5. DOI: 10.1007/978-3-030-45442-5_33. URL: https://doi.org/10.1007%2F978-3-030-45442-5_33.

[149] Abdullah Mazhar et al. 'Figurative-cum-Commonsense Knowledge Infusion for Multimodal Mental Health Meme Classification'. In: *Proceedings of the ACM on Web Conference 2025*. WWW '25. Sydney NSW, Australia: Association for Computing Machinery, 2025, pp. 637–648. ISBN: 9798400712746. DOI: 10.1145/3696410.3714778. URL: https://doi.org/10.1145/3696410.3714778.

[150] Thomas H. McCoy, Victor M. Castro and Roy H. Perlis. 'Estimating depression severity in narrative clinical notes using large language models'. In: *Journal of Affective Disorders* 381 (2025), pp. 270–274. ISSN: 0165-0327. DOI: https://doi.org/

10.1016/j.jad.2025.04.014. URL: https://www.sciencedirect.com/science/article/pii/S016503272500566X.

[151] Johannes Melsbach et al. 'Triplet transformer network for multi-label document classification'. In: *Proceedings of the 22nd ACM Symposium on Document Engineering*. DocEng '22. San Jose, California: Association for Computing Machinery, 2022. ISBN: 9781450395441. DOI: 10.1145/3558100.3563843. URL: https://doi.org/10.1145/3558100.3563843.

[152] Yiwen Meng et al. 'Bidirectional Representation Learning From Transformers Using Multimodal Electronic Health Record Data to Predict Depression'. In: *IEEE Journal of Biomedical and Health Informatics* 25.8 (2021), pp. 3121–3129. DOI: 10.1109/JBHI.2021.3063721.

[153] Alejandro Metke-Jimenez and Sarvnaz Karimi. 'Concept Identification and Normalisation for Adverse Drug Event Discovery in Medical Forums.' In: *BMDID@ ISWC*. 2016.

[154] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. 2013. arXiv: 1301.3781 [cs.CL]. URL: https://arxiv.org/abs/1301.3781.

[155] David N. Milne et al. 'CLPsych 2016 Shared Task: Triaging content in online peer-support forums'. In: *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*. Ed. by Kristy Hollingshead and Lyle Ungar. San Diego, CA, USA: Association for Computational Linguistics, June 2016, pp. 118–127. DOI: 10.18653/v1/W16-0312. URL: https://aclanthology.org/W16-0312/.

[156] Zuheng Ming et al. 'Simple Triplet Loss Based on Intra/Inter-Class Metric Learning for Face Verification'. In: *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*. 2017, pp. 1656–1664. DOI: 10.1109/ICCVW.2017.194.

[157] Seyed Iman Mirzadeh et al. 'Improved knowledge distillation via teacher assistant'. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 04. 2020, pp. 5191–5198.

[158] Rohan Mishra et al. '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 Research Workshop*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 147–156. DOI: 10.18653/v1/N19-3019. URL: https://aclanthology.org/N19-3019.

[159] Saif M Mohammad and Peter D Turney. 'Crowdsourcing a word–emotion association lexicon'. In: *Computational intelligence* 29.3 (2013), pp. 436–465. DOI: 10.1111/j.1467-8640.2012.00460.x.

[160] Saif M. Mohammad and Svetlana Kiritchenko. 'Using Hashtags to Capture Fine Emotion Categories from Tweets'. In: *Computational Intelligence* 31.2 (2015), pp. 301–326. DOI: 10.1111/coin.12024. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/coin.12024. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/coin.12024.

[161] Danielle L. Mowery et al. 'Task 2: ShARe/CLEF eHealth Evaluation Lab 2014'. In: *Proceedings of CLEF 2014*. Sheffield, United Kingdom, Sept. 2014. URL: https://hal.science/hal-01086544.

[162] Danielle L. Mowery et al. 'Towards Automatically Classifying Depressive Symptoms from Twitter Data for Population Health'. In: *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 182–191. URL: https://aclanthology.org/W16-4320.

[163] Aldrian Obaja Muis and Wei Lu. 'Learning to Recognize Discontiguous Entities'. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Ed. by Jian Su, Kevin Duh and Xavier Carreras. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 75–84. DOI: 10.18653/v1/D16-1008. URL: https://aclanthology.org/D16-1008.

[164] Wajid Mumtaz and Abdul Qayyum. 'A deep learning framework for automatic diagnosis of unipolar depression'. In: *International Journal of Medical Informatics* 132 (2019), p. 103983. ISSN: 1386-5056. DOI: https://doi.org/10.1016/j.

ijmedinf.2019.103983. URL: https://www.sciencedirect.com/science/article/pii/S1386505619307154.

[165] Jibon Naher. 'Can ChatGPT provide a better support: a comparative analysis of ChatGPT and dataset responses in mental health dialogues'. In: *Current Psychology* 43.28 (2024), pp. 23837–23845. ISSN: 1936-4733. DOI: 10.1007/s12144-024-06140-z. URL: https://doi.org/10.1007/s12144-024-06140-z.

[166] Usman Naseem et al. 'Early Identification of Depression Severity Levels on Reddit Using Ordinal Classification'. In: *Proceedings of the ACM Web Conference 2022*. WWW '22. Virtual Event, Lyon, France: Association for Computing Machinery, 2022, pp. 2563–2572. ISBN: 9781450390965. DOI: 10.1145/3485447.3512128. URL: https://doi.org/10.1145/3485447.3512128.

[167] Fateme Nateghi Haredasht et al. 'Predicting treatment retention in medication for opioid use disorder: a machine learning approach using NLP and LLM-derived clinical features'. In: *Journal of the American Medical Informatics Association* 32.12 (Sept. 2025), pp. 1865–1876. ISSN: 1527-974X. DOI: 10.1093/jamia/ocaf157. eprint: https://academic.oup.com/jamia/article-pdf/32/12/1865/64341326/ocaf157.pdf. URL: https://doi.org/10.1093/jamia/ocaf157.

[168] Dat Quoc Nguyen, Thanh Vu and Anh Tuan Nguyen. *BERTweet: A pre-trained language model for English Tweets*. 2020. arXiv: 2005.10200 [cs.CL]. URL: https://arxiv.org/abs/2005.10200.

[169] Jianyuan Ni et al. 'Cross-modal knowledge distillation for Vision-to-Sensor action recognition'. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 4448–4452.

[170] Luiz Henrique Pereira Niero et al. 'PsyBERTpt: A Clinical Entity Recognition Model for Psychiatric Narratives'. In: *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*. 2023, pp. 672–677. DOI: 10.1109/CBMS58004.2023.00298.

[171] Bridianne O'Dea et al. 'Detecting suicidality on Twitter'. In: *Internet Interventions* 2.2 (2015), pp. 183–188. ISSN: 2214-7829. DOI: https://doi.org/10.1016/

j.invent.2015.03.005. URL: https://www.sciencedirect.com/science/article/pii/S2214782915000160.

[172] Julia Ohse et al. 'GPT-4 shows potential for identifying social anxiety from clinical interview data'. In: *Scientific Reports* 14.1 (2024), p. 30498. ISSN: 2045-2322. DOI: 10.1038/s41598-024-82192-2. URL: https://doi.org/10.1038/s41598-024-82192-2.

[173] Josue Ortega Caro et al. 'BrainLM: A foundation model for brain activity recordings'. In: *bioRxiv* (2023), pp. 2023–09.

[174] Tianjian Ouyang et al. 'Health CLIP: Depression Rate Prediction Using Health Related Features in Satellite and Street View Images'. In: *Companion Proceedings of the ACM Web Conference 2024*. WWW '24. Singapore, Singapore: Association for Computing Machinery, 2024, pp. 1142–1145. ISBN: 9798400701726. DOI: 10.1145/3589335.3651451. URL: https://doi.org/10.1145/3589335.3651451.

[175] Wei Pan et al. 'Exploring the ability of vocal biomarkers in distinguishing depression from bipolar disorder, schizophrenia, and healthy controls'. In: *Frontiers in Psychiatry* Volume 14 - 2023 (2023). ISSN: 1664-0640. DOI: 10.3389/fpsyt.2023.1079448. URL: https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2023.1079448.

[176] Sungjoon Park et al. 'Suicidal Risk Detection for Military Personnel'. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, Nov. 2020, pp. 2523–2531. DOI: 10.18653/v1/2020.emnlp-main.198. URL: https://aclanthology.org/2020.emnlp-main.198/.

[177] Rashmi Patel et al. 'NeuroBlu, an electronic health record (EHR) trusted research environment (TRE) to support mental healthcare analytics with real-world data'. In: *BMJ Open* 12.4 (2022). ISSN: 2044-6055. DOI: 10.1136/bmjopen-2021-057227. eprint: https://bmjopen.bmj.com/content/12/4/e057227.full.pdf. URL: https://bmjopen.bmj.com/content/12/4/e057227.

[178] Dipti Pawar and Shraddha Phansalkar. 'Adapting Deep Learning Transformer Model for Depression-Specific Named Entity Recognition: A Transfer Learning Approach'.

In: *2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS)*. 2024, pp. 1098–1105. DOI: 10.1109/ICACRS62842.2024.10841739.

[179] Jeffrey Pennington, Richard Socher and Christopher Manning. 'GloVe: Global Vectors for Word Representation'. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162.

[180] Anxo Perez et al. 'DepreSym: A Depression Symptom Annotated Corpus and the Role of Large Language Models as Assessors of Psychological Markers'. In: *Language Resources and Evaluation* (2025), pp. 1–26.

[181] Cuong Pham, Tuan Hoang and Thanh-Toan Do. 'Collaborative Multi-Teacher Knowledge Distillation for Learning Low Bit-width Deep Neural Networks'. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 6435–6443.

[182] 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*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 9–12. DOI: 10.18653/v1/W18-5903. URL: https://aclanthology.org/W18-5903.

[183] Sameer Pradhan et al. 'Task 1: ShARe/CLEF eHealth Evaluation Lab 2013.' In: *CLEF (working notes)* 1179 (2013).

[184] Wei Qin et al. 'Explainable and Interactive LLMs-Augmented Depression Detection in Social Media'. In: *IEEE Transactions on Computational Social Systems* (2025), pp. 1–12. DOI: 10.1109/TCSS.2025.3591755.

[185] Saima Rani, Khandakar Ahmed and Sudha Subramani. 'From Posts to Knowledge: Annotating a Pandemic-Era Reddit Dataset to Navigate Mental Health Narratives'. In: *Applied Sciences* 14.4 (2024). ISSN: 2076-3417. DOI: 10.3390/app14041547. URL: https://www.mdpi.com/2076-3417/14/4/1547.

[186] Mirco Ravanelli et al. ‘SpeechBrain: A general-purpose speech toolkit’. In: *arXiv preprint arXiv:2106.04624* (2021).

[187] Kavita Rawat and Trapti Sharma. ‘An enhanced CNN-Bi-transformer based framework for detection of neurological illnesses through neurocardiac data fusion’. In: *Scientific Reports* 15.1 (2025), p. 11379. ISSN: 2045-2322. DOI: 10.1038/s41598-025-96052-0. URL: https://doi.org/10.1038/s41598-025-96052-0.

[188] Darrel A Regier, Emily A Kuhl and David J Kupfer. ‘The DSM-5: Classification and criteria changes’. en. In: *World Psychiatry* 12.2 (June 2013), pp. 92–98.

[189] Fuji Ren and Siyuan Xue. ‘Intention Detection Based on Siamese Neural Network With Triplet Loss’. In: *IEEE Access* 8 (2020), pp. 82242–82254. DOI: 10.1109/ACCESS.2020.2991484.

[190] Lu Ren et al. ‘Depression Detection on Reddit With an Emotion-Based Attention Network: Algorithm Development and Validation’. In: *JMIR Med Inform* 9.7 (July 2021), e28754. ISSN: 2291-9694. DOI: 10.2196/28754. URL: http://www.ncbi.nlm.nih.gov/pubmed/34269683.

[191] Fabien Ringeval et al. ‘AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition’. In: *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. AVEC ’19. Nice, France: Association for Computing Machinery, 2019, pp. 3–12. ISBN: 9781450369138. DOI: 10.1145/3347320.3357688. URL: https://doi.org/10.1145/3347320.3357688.

[192] Misha Sadeghi et al. ‘Exploring the Capabilities of a Language Model-Only Approach for Depression Detection in Text Data’. In: *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. 2023, pp. 1–5. DOI: 10.1109/BHI58575.2023.10313367.

[193] Salim Salmi et al. ‘The Most Effective Interventions for Classification Model Development to Predict Chat Outcomes Based on the Conversation Content in Online Suicide Prevention Chats: Machine Learning Approach’. In: *JMIR Ment Health* 11 (Sept. 2024), e57362. ISSN: 2368-7959. DOI: 10.2196/57362. URL: https://doi.org/10.2196/57362.

[194] Maarten Sap et al. 'ATOMIC: an atlas of machine commonsense for if-then reasoning'. In: AAAI'19/IAAI'19/EAAI'19. Honolulu, Hawaii, USA: AAAI Press, 2019. ISBN: 978-1-57735-809-1. DOI: 10.1609/aaai.v33i01.33013027. URL: https://doi.org/10.1609/aaai.v33i01.33013027.

[195] Sara Sardari et al. 'Audio based depression detection using Convolutional Autoencoder'. In: *Expert Systems with Applications* 189 (2022), p. 116076. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2021.116076. URL: https://www.sciencedirect.com/science/article/pii/S0957417421014147.

[196] Tamilarasi Sarveswaran and Vijayarajan Rajangam. 'Schizophrenia Segmentation From sMRI Data: A Fusion Approach Using U-Net and Vision Transformer-Based Grad-CAM'. In: *IEEE Access* 13 (2025), pp. 184376–184397. DOI: 10.1109/ACCESS.2025.3625766.

[197] Justyna Sarzynska-Wawer et al. 'Detecting formal thought disorder by deep contextualized word representations'. In: *Psychiatry Research* 304 (2021), p. 114135. ISSN: 0165-1781. DOI: https://doi.org/10.1016/j.psychres.2021.114135. URL: https://www.sciencedirect.com/science/article/pii/S0165178121004315.

[198] Ramit Sawhney et al. 'A Time-Aware Transformer Based Model for Suicide Ideation Detection on Social Media'. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 7685–7697. DOI: 10.18653/v1/2020.emnlp-main.619. URL: https://aclanthology.org/2020.emnlp-main.619.

[199] Ramit Sawhney et al. 'PHASE: Learning Emotional Phase-aware Representations for Suicide Ideation Detection on Social Media'. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 2415–2428. DOI: 10.18653/v1/2021.eacl-main.205. URL: https://aclanthology.org/2021.eacl-main.205.

[200] Ramit Sawhney et al. ‘Robust suicide risk assessment on social media via deep adversarial learning’. In: *Journal of the American Medical Informatics Association* 28.7 (Mar. 2021), pp. 1497–1506. ISSN: 1527-974X. DOI: 10.1093/jamia/ocab031. eprint: https://academic.oup.com/jamia/article-pdf/28/7/1497/38983084/ocab031.pdf. URL: https://doi.org/10.1093/jamia/ocab031.

[201] Ramit Sawhney et al. ‘Suicide Ideation Detection via Social and Temporal User Representations using Hyperbolic Learning’. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 2176–2190. DOI: 10.18653/v1/2021.naacl-main.176. URL: https://aclanthology.org/2021.naacl-main.176.

[202] Ramit Sawhney et al. ‘Towards Ordinal Suicide Ideation Detection on Social Media’. In: *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. WSDM ’21. Virtual Event, Israel: Association for Computing Machinery, 2021, pp. 22–30. ISBN: 9781450382977. DOI: 10.1145/3437963.3441805. URL: https://doi.org/10.1145/3437963.3441805.

[203] Matthias Schmidmaier et al. ‘Using Nonverbal Cues in Empathic Multi-Modal LLM-Driven Chatbots for Mental Health Support’. In: *Proc. ACM Hum.-Comput. Interact.* 9.5 (Sept. 2025). DOI: 10.1145/3743724. URL: https://doi.org/10.1145/3743724.

[204] Florian Schroff, Dmitry Kalenichenko and James Philbin. ‘Facenet: A unified embedding for face recognition and clustering’. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 815–823.

[205] Seungmin Seo et al. ‘Active Learning on Pre-trained Language Model with Task-Independent Triplet Loss’. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.10 (June 2022), pp. 11276–11284. DOI: 10.1609/aaai.v36i10.21378. URL: https://ojs.aaai.org/index.php/AAAI/article/view/21378.

[206] Ashish Sharma et al. 'Towards Facilitating Empathic Conversations in Online Mental Health Support: A Reinforcement Learning Approach'. In: *Proceedings of the Web Conference 2021*. WWW '21. Ljubljana, Slovenia: Association for Computing Machinery, 2021, pp. 194–205. ISBN: 9781450383127. DOI: 10.1145/3442381.3450097. URL: https://doi.org/10.1145/3442381.3450097.

[207] Jonathan Shen et al. 'Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions'. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, pp. 4779–4783. DOI: 10.1109/ICASSP.2018.8461368.

[208] Daming Shi, Maysam Orouskhani and Yasin Orouskhani. 'A conditional Triplet loss for few-shot learning and its application to image co-segmentation'. In: *Neural Networks* 137 (2021), pp. 54–62. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2021.01.002. URL: https://www.sciencedirect.com/science/article/pii/S0893608021000022.

[209] Han-Chin Shing et al. '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*. New Orleans, LA: Association for Computational Linguistics, June 2018, pp. 25–36. DOI: 10.18653/v1/W18-0603. URL: https://aclanthology.org/W18-0603.

[210] Gopendra Vikram Singh et al. 'Deciphering Cognitive Distortions in Patient-Doctor Mental Health Conversations: A Multimodal LLM-Based Detection and Reasoning Framework'. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 22546–22570. DOI: 10.18653/v1/2024.emnlp-main.1256. URL: https://aclanthology.org/2024.emnlp-main.1256/.

[211] Pradyumna Prakhar Sinha et al. '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.

[212] Aseem Srivastava et al. 'Knowledge Planning in Large Language Models for Domain-Aligned Counseling Summarization'. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 17775–17789. DOI: 10.18653/v1/2024.emnlp-main.984. URL: https://aclanthology.org/2024.emnlp-main.984/.

[213] Jane K Stallman and Gerald J Haeffel. 'A paper-and-pencil questionnaire outperforms GPT for measuring cognitive vulnerability to depression and predicting depressive symptoms'. en. In: *Sci. Rep.* 15.1 (June 2025), p. 19782.

[214] Akshay Swaminathan et al. 'Natural language processing system for rapid detection and intervention of mental health crisis chat messages'. In: *npj Digital Medicine* 6.1 (2023), p. 213. ISSN: 2398-6352. DOI: 10.1038/s41746-023-00951-3. URL: https://doi.org/10.1038/s41746-023-00951-3.

[215] Michael M. Tadesse et al. 'Detection of Depression-Related Posts in Reddit Social Media Forum'. In: *IEEE Access* 7 (2019), pp. 44883–44893. DOI: 10.1109/ACCESS.2019.2909180.

[216] Xu Tan et al. 'Multilingual Neural Machine Translation with Knowledge Distillation'. In: *International Conference on Learning Representations*. 2018.

[217] Buzhou Tang et al. 'Recognizing Continuous and Discontinuous Adverse Drug Reaction Mentions from Social Media Using LSTM-CRF'. In: *Wireless Communications and Mobile Computing* 2018.1 (2018), p. 2379208.

[218] Buzhou Tang et al. 'Recognizing disjoint clinical concepts in clinical text using machine learning-based methods'. In: *AMIA Annual Symposium Proceedings*. Vol. 2015. American Medical Informatics Association. Nov. 2015, pp. 1184–1193. eprint: https://pubmed.ncbi.nlm.nih.gov/26958258.

[219] Hao Tang et al. 'Dependency Graph Enhanced Dual-transformer Structure for Aspect-based Sentiment Classification'. In: *Proceedings of the 58th Annual Meeting of the*

Association for Computational Linguistics. Online: Association for Computational Linguistics, July 2020, pp. 6578–6588. DOI: 10.18653/v1/2020.acl-main.588. URL: https://aclanthology.org/2020.acl-main.588.

[220] Gemini Team et al. 'Gemini: a family of highly capable multimodal models'. In: *arXiv preprint arXiv:2312.11805* (2023).

[221] Bazen Gashaw Teferra et al. 'Leveraging large language models for automated depression screening'. en. In: *PLOS Digit. Health* 4.7 (July 2025), e0000943.

[222] Julia Thomas et al. 'An Explainable Artificial Intelligence Text Classifier for Suicidality Prediction in Youth Crisis Text Line Users: Development and Validation Study'. In: *JMIR Public Health Surveill* 11 (Jan. 2025), e63809. ISSN: 2369-2960. DOI: 10.2196/63809. URL: https://doi.org/10.2196/63809.

[223] Hugo Touvron et al. *LLaMA: Open and Efficient Foundation Language Models*. 2023. arXiv: 2302.13971 [cs.CL]. URL: https://arxiv.org/abs/2302.13971.

[224] Suppawong Tuarob et al. 'Forecasting National-Level Self-Harm Trends With Social Networks'. In: *IEEE Access* 11 (2023), pp. 64796–64814. DOI: 10.1109/ACCESS.2023.3289295.

[225] Elsbeth Turcan and Kathy McKeown. 'Dreaddit: A Reddit Dataset for Stress Analysis in Social Media'. In: *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*. Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 97–107. DOI: 10.18653/v1/D19-6213. URL: https://aclanthology.org/D19-6213.

[226] 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. ISSN: 0167-739X. DOI: https://doi.org/10.1016/j.future.2021.05.032. URL: https://www.sciencedirect.com/science/article/pii/S0167739X21001825.

[227] Ana-Sabina Uban, Berta Chulvi and Paolo Rosso. 'Explainability of Depression Detection on Social Media: From Deep Learning Models to Psychological Interpretations and Multimodality'. In: *Early Detection of Mental Health Disorders by Social Media*

Monitoring: The First Five Years of the eRisk Project. Ed. by Fabio Crestani, David E. Losada and Javier Parapar. Cham: Springer International Publishing, 2022, pp. 289–320. ISBN: 978-3-031-04431-1. DOI: 10.1007/978-3-031-04431-1_13. URL: https://doi.org/10.1007/978-3-031-04431-1_13.

[228] Taha ValizadehAslani et al. 'PharmBERT: a domain-specific BERT model for drug labels'. In: *Briefings in Bioinformatics* 24.4 (June 2023), bbad226. ISSN: 1477-4054. DOI: 10.1093/bib/bbad226. eprint: https://academic.oup.com/bib/article-pdf/24/4/bbad226/50917371/bbad226.pdf. URL: https://doi.org/10.1093/bib/bbad226.

[229] L. Alexander Vance et al. 'Natural language processing to identify suicidal ideation and anhedonia in major depressive disorder'. In: *BMC Medical Informatics and Decision Making* 25.1 (2025), p. 20. ISSN: 1472-6947. DOI: 10.1186/s12911-025-02851-w. URL: https://doi.org/10.1186/s12911-025-02851-w.

[230] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

[231] Jayakorn Vongkulbhisal, Phongtharin Vinayavekhin and Marco Visentini-Scarzanella. 'Unifying heterogeneous classifiers with distillation'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3175–3184.

[232] Roman Vygon and Nikolay Mikhaylovskiy. 'Learning Efficient Representations for Keyword Spotting with Triplet Loss'. In: *Speech and Computer*. Ed. by Alexey Karpov and Rodmonga Potapova. Cham: Springer International Publishing, 2021, pp. 773–785. ISBN: 978-3-030-87802-3.

[233] Cheng Wan et al. 'Association between depressive symptoms and diagnosis of diabetes and its complications: A network analysis in electronic health records'. In: *Frontiers in Psychiatry* Volume 13 - 2022 (2022). ISSN: 1664-0640. DOI: 10.3389/fpsyt.2022.966758. URL: https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2022.966758.

[234] Bailin Wang and Wei Lu. 'Combining Spans into Entities: A Neural Two-Stage Approach for Recognizing Discontiguous Entities'. 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 et al. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6216–6224. DOI: 10.18653/v1/D19-1644. URL: https://aclanthology.org/D19-1644.

[235] Bailin Wang and Wei Lu. 'Neural Segmental Hypergraphs for Overlapping Mention Recognition'. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff et al. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 204–214. DOI: 10.18653/v1/D18-1019. URL: https://aclanthology.org/D18-1019.

[236] Guangyu Wang et al. 'Optimized glycemic control of type 2 diabetes with reinforcement learning: A proof-of-concept trial'. In: *Nature Medicine* 29.10 (2023), pp. 2633–2642. DOI: 10.1038/s41591-023-02552-9.

[237] Hao Wang et al. 'MFE-Former: Disentangling Emotion-Identity Dynamics via Self-Supervised Learning for Enhancing Speech-Driven Depression Detection'. In: *IEEE Journal of Biomedical and Health Informatics* (2025), pp. 1–12. DOI: 10.1109/JBHI.2025.3594166.

[238] Haoyi Wang, Victor Sanchez and Chang-Tsun Li. 'Age-Oriented Face Synthesis With Conditional Discriminator Pool and Adversarial Triplet Loss'. In: *IEEE Transactions on Image Processing* 30 (2021), pp. 5413–5425. DOI: 10.1109/TIP.2021.3084106.

[239] Kunze Wang et al. *ME-GCN: Multi-dimensional Edge-Embedded Graph Convolutional Networks for Semi-supervised Text Classification*. 2022. arXiv: 2204.04618 [cs.CL].

[240] Song Wang et al. 'A multi-stage large language model framework for extracting suicide-related social determinants of health'. In: *Communications Medicine* 5.1 (2025), p. 404. ISSN: 2730-664X. DOI: 10.1038/s43856-025-01114-z. URL: https://doi.org/10.1038/s43856-025-01114-z.

[241] Song Wang et al. 'An NLP approach to identify SDoH-related circumstance and suicide crisis from death investigation narratives'. In: *Journal of the American Medical*

Informatics Association 30.8 (Apr. 2023), pp. 1408–1417. ISSN: 1527-974X. DOI: 10.1093/jamia/ocad068. eprint: https://academic.oup.com/jamia/article-pdf/30/8/1408/50908641/ocad068.pdf. URL: https://doi.org/10.1093/jamia/ocad068.

[242] Yaqi Wang et al. 'A hybrid model using multimodal feature perception and multiple cross-attention fusion for depressive episodes detection'. In: *Information Fusion* 124 (2025), p. 103354. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2025.103354. URL: https://www.sciencedirect.com/science/article/pii/S1566253525004270.

[243] Yiding Wang et al. 'A Multimodal Feature Fusion-Based Method for Individual Depression Detection on Sina Weibo'. In: *2020 IEEE 39th International Performance Computing and Communications Conference (IPCCC)*. 2020, pp. 1–8. DOI: 10.1109/IPCCC50635.2020.9391501.

[244] Yucheng Wang et al. 'Discontinuous Named Entity Recognition as Maximal Clique Discovery'. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong et al. Online: Association for Computational Linguistics, Aug. 2021, pp. 764–774. DOI: 10.18653/v1/2021.acl-long.63. URL: https://aclanthology.org/2021.acl-long.63.

[245] Zhong-Ling Wang et al. 'Self-Adapted Utterance Selection for Suicidal Ideation Detection in Lifeline Conversations'. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by Andreas Vlachos and Isabelle Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 1436–1446. DOI: 10.18653/v1/2023.eacl-main.105. URL: https://aclanthology.org/2023.eacl-main.105/.

[246] Samantha Weber et al. 'Using a fine-tuned large language model for symptom-based depression evaluation'. In: *npj Digital Medicine* 8.1 (2025), p. 598. ISSN: 2398-6352. DOI: 10.1038/s41746-025-01982-8. URL: https://doi.org/10.1038/s41746-025-01982-8.

[247] Jason Wei et al. *Finetuned Language Models Are Zero-Shot Learners*. 2022. arXiv: 2109.01652 [cs.CL]. URL: https://arxiv.org/abs/2109.01652.

[248] Ziru Weng et al. 'Multi-Scale Temporal-Frequency Attention Network Based on Ocular Imaging for Depression Detection'. In: *IEEE Journal of Biomedical and Health Informatics* (2025), pp. 1–11. DOI: 10.1109/JBHI.2025.3604064.

[249] Isabella Catharina Wiest et al. 'Detection of suicidality from medical text using privacy-preserving large language models'. In: *The British Journal of Psychiatry* 225.6 (2024), pp. 532–537. DOI: 10.1192/bjp.2024.134.

[250] Isabella Catharina Wiest et al. 'Detection of suicidality from medical text using privacy-preserving large language models'. In: *The British Journal of Psychiatry* 225.6 (2024), pp. 532–537. DOI: 10.1192/bjp.2024.134.

[251] World Health Organization. *Mental Disorders*. June 2022. URL: https://www.who.int/news-room/fact-sheets/detail/mental-disorders.

[252] World Health Organization. 'One in 100 deaths is by suicide'. In: *World Health Organization news release, June* 17 (2021). URL: https://www.who.int/news/item/17-06-2021-one-in-100-deaths-is-by-suicide.

[253] Chuhan Wu, Fangzhao Wu and Yongfeng Huang. 'One Teacher is Enough? Pre-trained Language Model Distillation from Multiple Teachers'. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Ed. by Chengqing Zong et al. Online: Association for Computational Linguistics, Aug. 2021, pp. 4408–4413. DOI: 10.18653/v1/2021.findings-acl.387. URL: https://aclanthology.org/2021.findings-acl.387.

[254] Yujia Wu et al. 'Text Classification using Triplet Capsule Networks'. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. 2020, pp. 1–7. DOI: 10.1109/IJCNN48605.2020.9207201.

[255] Yu Xia et al. 'Debiasing Generative Named Entity Recognition by Calibrating Sequence Likelihood'. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber and Naoaki Okazaki. Toronto, Canada: Association for Computational

Linguistics, July 2023, pp. 1137–1148. DOI: 10.18653/v1/2023.acl-short.98. URL: https://aclanthology.org/2023.acl-short.98.

[256] Xuhai Xu et al. 'Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data'. In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8.1 (Mar. 2024). DOI: 10.1145/3643540. URL: https://doi.org/10.1145/3643540.

[257] Zhentao Xu, Verónica Pérez-Rosas and Rada Mihalcea. 'Inferring Social Media Users' Mental Health Status from Multimodal Information'. eng. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, May 2020, pp. 6292–6299. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.772/.

[258] Zhongzhi Xu et al. 'Detecting suicide risk using knowledge-aware natural language processing and counseling service data'. In: *Social Science Medicine* 283 (2021), p. 114176. ISSN: 0277-9536. DOI: https://doi.org/10.1016/j.socscimed.2021.114176. URL: https://www.sciencedirect.com/science/article/pii/S0277953621005086.

[259] Shweta Yadav et al. 'Towards Identifying Fine-Grained Depression Symptoms from Memes'. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 8890–8905. DOI: 10.18653/v1/2023.acl-long.495. URL: https://aclanthology.org/2023.acl-long.495/.

[260] Yuta Yamauchi et al. 'Development and Evaluation of an Auditory VR Generative System via Natural Language Interaction to Aid Exposure Therapy for PTSD Patients'. In: *ACM Trans. Comput. Healthcare* (Mar. 2025). Just Accepted. DOI: 10.1145/3723048. URL: https://doi.org/10.1145/3723048.

[261] Hang Yan et al. 'A Unified Generative Framework for Various NER Subtasks'. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume*

1: Long Papers). Ed. by Chengqing Zong et al. Online: Association for Computational Linguistics, Aug. 2021, pp. 5808–5822. DOI: 10.18653/v1/2021.acl-long.451. URL: https://aclanthology.org/2021.acl-long.451.

[262] Kailai Yang et al. '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.

[263] Lehan Yang and Kele Xu. 'Cross modality knowledge distillation for multi-modal aerial view object classification'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 382–387.

[264] Liang Yao, Chengsheng Mao and Yuan Luo. 'Graph Convolutional Networks for Text Classification'. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019), pp. 7370–7377. DOI: 10.1609/aaai.v33i01.33017370. URL: https://ojs.aaai.org/index.php/AAAI/article/view/4725.

[265] Jeewoo Yoon et al. 'D-vlog: Multimodal Vlog Dataset for Depression Detection'. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.11 (June 2022), pp. 12226–12234. DOI: 10.1609/aaai.v36i11.21483. URL: https://ojs.aaai.org/index.php/AAAI/article/view/21483.

[266] Yanhong Yu et al. 'Depression and Severity Detection Based on Body Kinematic Features: Using Kinect Recorded Skeleton Data of Simple Action'. In: *Frontiers in Neurology* Volume 13 - 2022 (2022). ISSN: 1664-2295. DOI: 10.3389/fneur.2022.905917. URL: https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2022.905917.

[267] Ye Yuan et al. 'In Defense of the Triplet Loss Again: Learning Robust Person Re-Identification With Fast Approximated Triplet Loss and Label Distillation'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2020.

[268] Kimia Zandbiglari et al. 'Enhancing suicidal behavior detection in EHRs: A multi-label NLP framework with transformer models and semantic retrieval-based annotation'.

In: *Journal of Biomedical Informatics* 161 (2025), p. 104755. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2024.104755. URL: https://www.sciencedirect.com/science/article/pii/S1532046424001734.

[269] H. Zen et al. 'LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech'. In: *Proc. Interspeech*. Sept. 2019. DOI: 10.21437/Interspeech.2019-2441.

[270] Wei Zhai et al. 'Chinese MentalBERT: Domain-Adaptive Pre-training on Social Media for Chinese Mental Health Text Analysis'. In: *Findings of the Association for Computational Linguistics: ACL 2024*. Ed. by Lun-Wei Ku, Andre Martins and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 10574–10585. DOI: 10.18653/v1/2024.findings-acl.629. URL: https://aclanthology.org/2024.findings-acl.629/.

[271] Chen Zhang, Qiuchi Li and Dawei Song. *Aspect-based Sentiment Classification with Aspect-specific Graph Convolutional Networks*. 2019. arXiv: 1909.03477 [cs.CL].

[272] Jun Zhang and Yanrong Guo. 'Multilevel depression status detection based on fine-grained prompt learning'. In: *Pattern Recognition Letters* 178 (2024), pp. 167–173. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2024.01.005. URL: https://www.sciencedirect.com/science/article/pii/S0167865524000059.

[273] Pingyue Zhang et al. 'DEPA: Self-Supervised Audio Embedding for Depression Detection'. In: *Proceedings of the 29th ACM International Conference on Multimedia*. MM '21. Virtual Event, China: Association for Computing Machinery, 2021, pp. 135–143. ISBN: 9781450386517. DOI: 10.1145/3474085.3479236. URL: https://doi.org/10.1145/3474085.3479236.

[274] Shuai Zhang et al. 'De-Bias for Generative Extraction in Unified NER Task'. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 808–818. DOI: 10.18653/v1/2022.acl-long.59. URL: https://aclanthology.org/2022.acl-long.59.

[275] Shuning Zhang et al. 'IntervEEG-LLM: Exploring EEG-Based Multimodal Data for Customized Mental Health Interventions'. In: *Companion Proceedings of the ACM on Web Conference 2025*. WWW '25. Sydney NSW, Australia: Association for Computing Machinery, 2025, pp. 2320–2326. ISBN: 9798400713316. DOI: 10.1145/3701716.3717550. URL: https://doi.org/10.1145/3701716.3717550.

[276] Wenjing Zhang et al. 'Detecting individuals with severe mental illness using artificial intelligence applied to magnetic resonance imaging'. In: *eBioMedicine* 90 (2023). doi: 10.1016/j.ebiom.2023.104541. ISSN: 2352-3964. DOI: 10.1016/j.ebiom.2023.104541. URL: https://doi.org/10.1016/j.ebiom.2023.104541.

[277] Yufeng Zhang et al. *Every Document Owns Its Structure: Inductive Text Classification via Graph Neural Networks*. 2020. arXiv: 2004.13826 [cs.CL].

[278] Zhenwen Zhang et al. 'Natural Language Processing for Depression Prediction on Sina Weibo: Method Study and Analysis'. In: *JMIR Ment Health* 11 (Sept. 2024), e58259. ISSN: 2368-7959. DOI: 10.2196/58259. URL: https://doi.org/10.2196/58259.

[279] Jin Zhao et al. 'A Novel Cascade Instruction Tuning Method for Biomedical NER'. In: *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2024, pp. 11701–11705. DOI: 10.1109/ICASSP48485.2024.10446885.

[280] Weipeng Zhou et al. 'Identifying Rare Circumstances Preceding Female Firearm Suicides: Validating A Large Language Model Approach'. In: *JMIR Ment Health* 10 (Oct. 2023), e49359. ISSN: 2368-7959. DOI: 10.2196/49359. URL: https://doi.org/10.2196/49359.

[281] Feiyu Zhu et al. 'MTNet: Multimodal transformer network for mild depression detection through fusion of EEG and eye tracking'. In: *Biomedical Signal Processing and Control* 100 (2025), p. 106996. ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2024.106996. URL: https://www.sciencedirect.com/science/article/pii/S1746809424010541.

[282] Xingyu Zhu et al. ‘GL-NER: Generation-Aware Large Language Models for Few-Shot Named Entity Recognition’. In: *Artificial Neural Networks and Machine Learning – ICANN 2024*. Ed. by Michael Wand et al. Cham: Springer Nature Switzerland, 2024, pp. 433–448. ISBN: 978-3-031-72350-6.

[283] Hamad Zogan et al. ‘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. DOI: 10.1007/s11280-021-00992-2.

[284] Hamad Zogan et al. ‘Hierarchical Convolutional Attention Network for Depression Detection on Social Media and Its Impact During Pandemic’. In: *IEEE Journal of Biomedical and Health Informatics* (2023), pp. 1–9. DOI: 10.1109/JBHI.2023.3243249.

[285] Bochao Zou et al. ‘Semi-Structural Interview-Based Chinese Multimodal Depression Corpus Towards Automatic Preliminary Screening of Depressive Disorders’. In: *IEEE Transactions on Affective Computing* 14.4 (2023), pp. 2823–2838. DOI: 10.1109/TAFFC.2022.3181210.

Appendix A Search Methods

The statistics provided in Chapter 2 are preliminary insights on a survey being conducted on media-based multimodal literature. Details of the data collection and initial review done is outlined here.

A comprehensive search of research databases was conducted to retrieve mental health-related literature published between 01 January 2020 and 31 October 2025[1]. Six databases, namely Google Scholar, Scopus, ACM Digital Library, PubMed, IEEE Xplore, and Web of Science, were searched using search terms outlined in Table A1. We identify four focus areas: model, mental health, modality or source, and objective. The search terms are applied to identify matches within titles and abstracts. When possible, the search is also limited by language and publication type, only including studies written in English and that are research articles, excluding other types of publications such as letters, editorials, and book chapters.

TABLE A1. Summary of search terms applied for collecting relevant literature.

Focus	**Search terms**
Mental Health†	"mental health" OR "mental illness*" OR "mental disorder*" OR "suicid*" OR "depress*" OR "anxiety" OR "bipolar" OR "PTSD" OR "stress disorder*" OR "schizophre*" OR "autis*" OR "anorexia"
Framework	"language model*" OR "LLM*" OR "PLM*" OR "transformer*"
Modality/Source	"multimodal*" OR "multi-modal" OR "visual" OR "image*" OR "video*" OR "audio" OR "speech*" OR "sound*" OR "voice*" OR "text*" OR "language" OR "sensor*" OR "time?series" OR "social media" OR "interview*" OR "wearable*"
Objective	detect* OR generat* OR predict* OR classif* OR explain* OR explanation

† The search was conducted with the initial inclusion of autism, however, it is excluded in subsequent steps along with other neurodevelopmental conditions. It is still reported here for reproducibility.

A total of 3,268 articles were retrieved from all six databases, of which 1,731 were unique based on titles matching. A first screening pass was done based on source quality and title information. To ensure high quality analysis, only articles published in journals in the Q1

[1]This may include electronically published material that is due for physical publication by publishers in future dates.

quantile based on the Scientific Journal Rankings (SJR) and in top tier conferences ranking A/A* based on the International Computing Research & Education (ICORE) Ranking[2] are included. Based on the title, articles that are discussion types, protocols, reviews, surveys, workshop overviews, cross-sectional studies, observational studies, and longitudinal studies are excluded. Non-mental health-based studies are also excluded. For this search, neurodevelopmental conditions such as Autism Spectrum Disorder (ASD) and Attention Deficit Hyperactivity Disorder (ADHD), neurocognitive disorders such as Dementia and Alzheimer's Disease, and other brain development-related conditions are excluded. While also largely considered as mental health disorders and are as equally important, this search focused on internalised and externalised mental health conditions.

After a first pass, a total of 335 articles were considered for further review of the abstract, further removing articles that does not focus on mental health. This resulted in the 263 articles used for the statistics shown in Figure 2.1.

[2] https://www.core.edu.au/icore-portal

Appendix B 3M-Health Audio Representation Analysis

This appendix section shows more supplementary examples for the audio analysis in Chapter 4 Section 4.5, illustrating the differences in post content for the most and least concerning classes among different audio durations. Table B1 and Table B2 shows samples from the TwitSuicide and IdenDep datasets, respectively.

TABLE B1. Samples for the TwitSuicide audio spectrogram analysis. Each sample has been masked to avoid a reverse search of each post. SI: Safe to Ignore; SC: Seriously Concerning.

ID	Class	Text
0-10 seconds		
214	SI	_USER_ *** i can't get that link to work
251	SI	_USER_ or, *** anyone from the *** Rookies all-female racers team
355	SI	*** kill myself.. watching Drag Me To Hell
457	SI	i'm afraid my ups might be dead *** making ticking noises
495	SI	It's too early to be awake *** got up 3 1/2 hours ago! *** never wake up before 8.
152	SC	_USER_ *** never wanted to be dead til now...
187	SC	_USER_ thanks...now *** me kill myself
336	SC	feeling like death *** want to die
407	SC	I *** die right now no one loves me
573	SC	*** hate my life sometimes i just want to die
10-25 seconds		
92	SI	_USER_ Its a story about how success as *** columnist, *** helped create, returned him to alcohol & suicidal thoughts
216	SI	_USER_ Thx for *your* part ***! Any time you reblog *** I instantly get 3-5x the activity on it I normally do!
359	SI	Gosh back to *** from *** with a rather large BANG. Really don't want to be here, much rather be sailing *** for lunch
506	SI	Last paper ***! Lucky I've bought *** or I will be dead.
553	SI	oh noesss *** is dieing im ganna kill myself!!! *** room with no phone fml.
81	SC	_USER_ ugh. *** that one was awful. suicide always *** gets to me...
160	SC	_USER_ *** so stupid! *** think I still can? I want to kill myself!
416	SC	I need to go on suicide watch. ***, The Fumble, ***, Jose Mesa, ***, and now this... Where's my razor blade?
417	SC	I never have any one to talk to *** i hate my self *** kill myself if no one *** say anything to me on ***
589	SC	So ***. I'm in pain. Sucks. That was *** the point. suicide an option?

TABLE B2. Samples for the IdenDep audio spectrogram analysis. Each sample has been masked to avoid a reverse search of each post. NDE: Non-Depressive; DE: Depressive.

ID	Class	Text
0-10 seconds		
1324	NDE	*** Sympathy gift ideas + ***
1427	NDE	*** friend vlog
1701	NDE	does anyone want to hear the story about *** 'beef' between ***
1728	NDE	TRUST by " THE HIPSTERS " (ft. *** and ***)
1754	NDE	How to have a strong family *** products or services have helped *** family stay strong together?
172	DE	It's *** easier to fall back in than to fight *** it
655	DE	*** friends are throwing a LAN party *** I wasn't invited. *** only one who didn't get an invitation.
904	DE	*** feel bitter about everything. *** bitter about being bitter.
934	DE	I'm sad I feel sad. *** I feel something.
1235	DE	... I just want to crawl in a whole and cry ***
10-25 seconds		
1377	NDE	Having friends *** opposite sex *** in a relationship _URL_
1470	NDE	*** introduce my girlfriend [18F] to my family My girlfriend lives *** I live in *** *** introduce her to my mom but I don't know how
1589	NDE	365 New Ways To Hug Your Love *** discover and post videos or pictures of New Ways To Hug in the new subreddit ***
1616	NDE	With family being a main interest in your lives, what *** would you purchase *** to help the family to grow?
1824	NDE	What Would You Do? Would you move away from your family *** to somewhere far where your kids would have a better education *** provide for your family better, like buying a house; *** moving from *** to the *** or ***?
132	DE	Anyone else feel like everyone hates them? *** paranoia? *** the dark cloud over my head just gives off a shitty vibe *** people think I don't like them and vice Versa.
505	DE	That feeling when you hate who you *** but can't *** change because you are so used to being like this for *** years. *** a shitty person. The thought of change seems impossible *** at this point.
605	DE	Fuck me When you're *** a piece of shit *** look at other girls and lie to you, while lying *** next to you. I'll never be enough. Ever. For anyone. *** want to ducking die.
1027	DE	Addicted to depression *** when I feel like *** self-loathing and depressive *** becoming less, I feel shit *** don't feel depressed anymore. *** I want it to go away *** part of me wants to stay depressive and feel suicidal.
1154	DE	Is it depression *** don't want to build memories anymore? *** I get really nostalgic. *** I don't want to get too attached to people *** just end up hurting in the future.