Research on Early Detection of Depression Based on Neuro-Symbolic AI Approach via Social Media Analysis

Dou Rongyu

A Thesis submitted to Tokushima University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

March, 2024

Department of Information Science and Intelligent Systems
Graduate School of Advanced Technology and Science
Tokushima University, Japan

Table of Contents

List of Tables

List of Figures

Abstract

Depression, a pervasive mental health disorder, extends its impact globally, leaving profound consequences on individuals and society at large. Characterized by persistent feelings of despair, diminished interest, low energy levels, and heightened self-evaluation and guilt, depression manifests through various debilitating symptoms. Individuals grappling with this condition often exhibit disinterest in daily activities, disturbances in sleep patterns, alterations in appetite, compromised concentration, and, at times, even inclinations toward self-harm.

The accurate diagnosis and timely identification of early-stage depression represent formidable challenges within clinical practice. The subjective nature of depressive symptoms, which varies markedly from person to person and across diverse cultures, further complicates the diagnostic process. Current diagnostic approaches heavily rely on the experiential knowledge of clinicians and subjective accounts provided by patients. Unfortunately, this reliance introduces potential subjective biases and inconsistencies. Moreover, patients may conceal symptoms or inadvertently overlook signs of depression, thereby adding layers of complexity to achieving an accurate diagnosis. In response to these challenges, the integration of cutting-edge technologies, particularly artificial intelligence (AI), has emerged as an imperative for facilitating early depression detection.

Meanwhile, increasing evidence suggests that specific language and emotions expressed on social media platforms may provide clues about depression. In this context, machine learning is gradually being applied to depression detection based on social media text data. Methods based on traditional machine learning can perform automatic, objective, and effective assessments, but their performance largely depends on feature construction and selection, with generalization limited by the features and algorithms used. In contrast, deep learning, with the goal of understanding the context of complex natural language sentences, has fundamentally transformed the potential feature extraction process. Existing depression detection systems based on deep learning can execute continuous processes such as preprocessing, feature extraction, and depression detection, achieving end-to-end automated depression detection, which holds significant implications for the prevention and treatment of depression.

Nevertheless, a notable drawback in the realm of AI for early depression detection, compounded by challenges in data acquisition, labeling, and model generalization, is the inherent uninterpretability of the models employed. Many contemporary deep learning models, notably those rooted in deep neural networks, are often deemed "black box" models, signifying the opacity of their internal decision-making processes and logic. This lack of transparency poses a significant obstacle in elucidating how the model arrives at conclusions regarding an individual's early-stage depression. Given the importance of clear explanations in clinical decision-making, the interpretability challenge becomes a potential impediment to the widespread clinical application and acceptance of such models among doctors and patients.

Enter TAM-SenticNet, a pioneering Neuro-Symbolic AI framework meticulously crafted for the early detection of depression through an in-depth analysis of social media content. In a concerted effort to overcome the constraints of traditional diagnostic tools, TAM-SenticNet seamlessly integrates neural networks for adept feature extraction and sentiment analysis with symbolic reasoning for intricate logical inference. This fusion significantly enhances the model's explainability, addressing a critical gap in current AI applications for mental health.

Empirical evaluations reveal that TAM-SenticNet excels beyond existing models in performance metrics, achieving a Precision of 0.665, Recall of 0.881, and F_1-score of 0.758, coupled with superior latency metrics, including ERDE_5 and ERDE_{50} at 0.025, $\mathrm{Latency}_{TP}$ at 1.0, and $\mathrm{F}_{latency}$ at 0.675. These achievements highlight TAM-SenticNet's cutting-edge approach to early depression detection, making it a pioneering tool in the application of AI for mental health informatics.

In conclusion, this paper delves into the intricate facets of depression, emphasizing the pivotal role of advanced technologies, specifically Neuro-Symbolic AI, in achieving early detection. TAM-SenticNet not only presents an innovative approach to addressing early depression but also lays the groundwork for the extensive integration of AI into mental health research and practice.

Keywords: Neuro-Symbolic AI, Depression Detection, Social Media Analysis, Early Intervention, Sentiment Analysis, Explainability

1 Introduction

1.1 Research Background

Depression, a significant public health challenge, impacts individuals and society at large [1, 2, 3]. Early detection is vital for effective intervention [4]. Traditional diagnostic methods like the Hamilton Depression Rating Scale (HAMD) and Beck Depression Inventory (BDI) are crucial but have limitations, including episodic assessment and reliance on self-reporting, which may miss early or subtle signs [5]. In contrast, social media offers a continuous, real-time mental health monitoring platform, providing insights into emotional states overlooked by traditional methods [6, 7, 8].

Recent AI-driven approaches in depression detection have evolved across neural, symbolic, and hybrid methodologies. Neural models excel in pattern recognition but often lack interpretability and struggle with nuanced emotional analysis [9, 10]. Symbolic models offer structured reasoning and interpretability but can lack adaptability to complex datasets [11, 12]. Hybrid approaches attempt to balance these aspects but still face challenges in real-time processing and comprehensive emotional understanding [13, 14].

TAM-SenticNet, our Neuro-Symbolic AI framework, addresses these challenges. It combines neural networks' sentiment analysis capabilities with symbolic reasoning's logical inference, providing nuanced interpretations of emotional expressions in social media data. This approach not only overcomes the limitations of traditional methods but also harnesses the dynamic capabilities of social media analytics, setting a new standard in mental health research.

1.2 Research Motivation

1.2.1 Objective of the Study

The overarching objective is to significantly enhance the efficiency and precision of early depression detection by synergizing the strengths of neural networks and symbolic reasoning. The TAM-SENTICNET model, developed through the application of Neuro-Symbolic AI, represents an innovative approach that seeks to redefine the state-of-the-art in the field.

1.2.2 Benefits of Neural Networks and Symbolic Reasoning Integration

The synergy of neural networks, adept at processing vast datasets and extracting intricate features, with symbolic reasoning, known for facilitating a deeper understanding of the inference process, not only enables the model to process complex data but also enhances the interpretability of its decisions.

1.2.3 Innovation in AI Technology Application

Through the seamless amalgamation of neural networks and symbolic reasoning, our aspiration is not only to advance the current landscape of early depression detection but also to introduce an innovative AI-powered method that sets a new standard in the field.

1.2.4 Primary Goal for Early Depression Detection

The primary goal of this research is to present a more dependable, robust, and efficient solution for the early identification and intervention of depression, ultimately seeking to ameliorate the quality of life for individuals grappling with depression and lessen the societal burden associated with this prevalent mental health condition.

1.2.5 Advocacy for Neuro-Symbolic AI Integration in Mental Health

By advocating for and exemplifying the integration of Neuro-Symbolic AI in the realm of mental health, we aim to set a pioneering precedent. This initiative not only offers inspiration for interdisciplinary research but also strives to propel the profound integration of artificial intelligence technology with the multifaceted domains of psychology, psychiatry, and related fields.

1.2.6 Inspiration and Theoretical Foundation

The inspiration and theoretical foundation for this research are deeply rooted in the prevailing trend of multidisciplinary integration. By weaving together insights and methodologies from the dynamic fields of artificial intelligence, neuroscience, and psychology, this study stands as a testament to the power of cross-disciplinary collaboration. The theoretical innovations and empirical findings emerging from this study are poised to not only advance our understanding of early depression detection but also provide novel perspectives and technical support that can reshape the landscape of mental health research.

1.3 Research Objectives

The overarching aim of this study is multifaceted, encompassing the following specific research objectives:

1.3.1 Development of TAM-SENTICNET Model

- Expand the Neuro-Symbolic AI framework to construct the TAM-SENTICNET model.
- Integrate Neural Network and Symbolic Reasoning components effectively.

1.3.2 Enhancement of Early Depression Identification

- Investigate the potential of TAM-SENTICNET in improving the accuracy of early depression detection.
- Assess the model's capability to identify nuanced signs of early depression.

1.3.3 Utilization of TAM Model and SenticNet Library

- Incorporate the TAM model and leverage the SenticNet sentiment analysis library in the model architecture.

- Evaluate the synergistic impact of combining these elements on the overall performance.

1.3.4 Automation and Objectivity in Detection

- Aim for the automation of the detection process using TAM-SENTICNET.
- Assess the model's objectivity in identifying early signs, minimizing subjective biases.

1.3.5 Comprehensive Evaluation Metrics

- Develop and apply comprehensive evaluation metrics to gauge the effectiveness of TAM-SENTICNET.
- Include metrics such as precision, recall, F1-score, and latency measures for a holistic assessment.

1.3.6 Validation of Automated Detection

- Validate the automated detection capabilities of TAM-SENTICNET through empirical studies.
- Compare the model's performance against existing methods and benchmarks.

1.3.7 Exploration of Interdisciplinary Applications

- Explore potential interdisciplinary applications of TAM-SENTICNET beyond early depression detection.
- Investigate the model's adaptability to broader mental health informatics scenarios.

By delineating these specific research objectives, this study endeavors to contribute insights and advancements to the field of early depression detection, leveraging the innovative TAM-SENTICNET model grounded in Neuro-Symbolic AI principles.

1.4 Research Contributions

This study makes the following important contributions to the field of early depression detection, highlighting the interpretability of the models:

1.4.1 Explainable model design

A highly explainable TAM-SENTICNET model is proposed, which fully integrates neural network and symbolic reasoning, so that the decision-making process of the model has a clear explanation, which is helpful to understand the basis for the identification of early depression.

1.4.2 Application of emotion analysis

The TAM model is used for emotion analysis, which links emotional patterns with the characteristics of early depression, emphasizes the important role of emotion analysis in explaining the model, and provides a new perspective for depression research.

1.4.3 Integration of SenticNet library

The successful integration of SenticNet sentiment analysis library enriched the emotion information of the model, improved the explainability of the model, and contributed substantial content to the explainability of the model.

1.4.4 Empirical verification and practical analysis

The interpretability of TAM-SENTICNET model was verified through fully designed experimental proof, and its practicability for the detection of early depression was analyzed, providing a new idea and method for the automatic detection of early depression.

1.5 Thesis Organizations

In this chapter, we introduce the background of the research. We propose a TAM-SENTICNET model for the detection of early onset depression, focusing on solving the unexplainability of the general model. The organization of the paper is structured as follows:

Chapter 2:Literature review

Chapter 2:Research method

Chapter 4: Eexperiment

Chapter 5: Conclusion and Future work

2 Literature review

Depression detection as a subset of sentiment analysis. In this chapter, we will specifically review the development of sentiment analysis in the context of social media, including social media context sentiment tnalysis task, methods for sentiment analysis, and detection methods for depression, etc.

2.1 Social Media Context Sentiment Analysis Task

2.1.1 Emotion Classification

Sentiment classification, also known as sentiment polarity classification, is one of the most common tasks in sentiment analysis. It is based on the assumption that the opinions in the target text about an entity or aspect can be simply categorized into one of two opposite sentiment polarities, or positioned on a continuous variable between these two sentiment polarities[15]. Therefore, sentiments are generally divided into three main categories: positive, negative, or neutral. To express the intensity of sentiment, various measures can be used, such as the commonly used measurement range of ［-1, 1］, where -1 represents the maximum negative emotion, 1 represents the maximum positive emotion, and 0 represents a neutral attitude[16, 17]. Some studies categorize sentiment ratings into 5 levels, setting 0 as the maximum negative and 4 as the maximum positive emotion[18]. However, Thelwall et al. argue that positive and negative emotions can co-exist. They propose an algorithm that simultaneously measures both sentiment polarities, meaning that the sentiment classification result for a sentence can have both positive and negative values to express the intensity of emotion[19].

2.1.2 Emotion Analysis

According to research in affective psychology, although positive and negative emotions are crucial dimensions, there are many other types of emotions and criteria for measuring emotional intensity. Positive and negative polarities alone may not fulfill the requirements of emotion classification[20]. Tasks that involve subdividing emotion types are known as emotion analysis. Bollen et al. analyzed public emotions based on the Profile of Mood States, a psychological scale measuring mood, using six dimensions: tension, depression, anger, vigor, fatigue, and confusion[21] . Another study, based on Plutchik's emotion development psychology theory[22], mapped eight emotions—anger, fear, sadness, disgust, surprise, anticipation, trust, and joy—into four pairs of emotional polarities. This study discerned changes in the emotional states of Twitter users during significant events[23].

2.1.3 Temporal Analysis of Emotions

Time is a crucial dimension in defining problems related to sentiment analysisliu[24]. Over time, people may persist or change their opinions, and even introduce new viewpoints. Therefore, predicting future sentiments or events is essential in sentiment analysis. This task involves determining the sentiment expressed in the text, identifying and forecasting changes in sentiment trends over time, very much akin to sentiment prediction[25]. In a study, Twitter topics were categorized into different time periods such as peak periods, pre-peak periods, and post-peak periods based on time series analysis. This research tested the volume of topic discussions in different time periods, confirming the relationship between the popularity of Twitter topics and the intensity of positive and negative sentiments[26].

2.1.4 Subjective Detection

The task of subjective detection is to identify whether a given sentence is subjective. Objective sentences convey factual information, while subjective sentences can express personal thoughts, such as opinions, evaluations, emotions, and beliefs. These sentences may contain positive or negative sentiments, unlike objective sentences. This task can be seen as a preliminary step in sentiment classification, as effective subjective detection ensures more accurate sentiment classification[27]. Assessing the subjectivity of a sentence is even considered a more challenging process than distinguishing positive, negative, or neutral emotions[28].

2.1.5 Opinion Summarization

Beineke et al. introduced the concept of opinion summarization when analyzing movie reviews on "Rotten Tomatoes" to concisely express reviewers' evaluations of key aspects of a film[29]. While similar to text summarization, opinion summarization places a greater emphasis on extracting entity features commonly mentioned in one or more texts and their associated sentiments. Consequently, opinion summarization tasks can be divided into two aspects: single-text and multi-text opinion summarization. Single-text opinion summarization involves analyzing facts present in the text, such as changes in sentiment direction and discovering connections between different entities or features, while extracting more coherent text segments[30]. On the other hand, in multi-text opinion summarization, once entities or features are detected, it becomes necessary to group or rank multiple sentences expressing sentiments related to those entities or features to extract meaningful statements. The final form of the summary can be in text[31], numerical[32], or graphical formats, describing key entities or features and quantifying sentiments related to each entity or feature in some way[33]. For instance, Hu et al. tallied the number of positive or negative sentences related to each product feature in user reviews, extracting a summary of user feedback by quantifying the number of these sentences[32].

2.1.6 Opinion Retrieval

Opinion retrieval aims to retrieve documents containing opinions, perspectives, or viewpoints based on given query terms. This was a primary task in the TREC Blog Track from 2006 to 2010[34]. In opinion retrieval systems, it is common to calculate two scores for each document: a relevance score to the query and an opinion score regarding the query. The final ranking of documents is then determined based on the combined scores[35].

2.1.7 Opinion Holder Extractionl

Opinion holder extraction is the task of identifying who holds an opinion (or the source of the opinion)[36]. Many text analysis tasks focus on finding more expressive and influential opinions, extracting information related to different perspectives. Recognizing opinion holders is crucial for distinguishing opinions from various viewpoints[37]. For example, in the sentence "What is Miss Universe's opinion on world peace?" the opinion holder is "Miss Universe". To answer this question, both the opinion and the opinion holder need to be extracted simultaneously. This task is based on fine-grained opinion mining. It is important to note that the opinion holder can be explicit (coming from named entities or noun phrases in the sentence) or implicit (coming from the author of the text being analyzed)[38].

2.1.8 Irony and Sarcasm Detectionl

Irony and sarcasm detection focus on identifying statements that contain ironic or sarcastic content. Discovering sentences with irony and sarcasm can significantly enhance the performance of sentiment analysis but is also one of the most challenging tasks in the field of natural language processing. This difficulty arises, in part, from the lack of consensus among researchers (linguists, psychologists, computer scientists) on how to formally define irony or sarcasm and their structures[39]. A widely accepted view is that a key feature of sarcastic sentences involves the use of positive words to express negative opinions, closely related to the context of the surrounding text[39, 40, 41].

2.1.9 Cross-Domain Sentiment Analysis

A critical drawback in sentiment analysis is its strong dependency on the domain. In other words, methods that perform well in one domain may not generalize well to another. This issue hinders the potential sharing of valuable information across domains.

Consequently, some scholars have initiated research on cross-domain sentiment analysis tasks to address this problem[42, 43, 44].

2.1.10 Multimodal Sentiment Analysis

While text has consistently been a hot topic in sentiment analysis research, social media data is not limited to a single text mode. For instance, users expressing their opinions on Twitter may often accompany their tweets with photos uploaded on Instagram and relevant videos on YouTube. Video-based social media platforms alone provide two modalities: sound and visuals. Consequently, multimodal sentiment analysis has emerged as a new research area. Scholars aim to identify emotions expressed by individuals on social multimedia platforms, considering visual, audio, and text information[45].

Multimodal sentiment analysis can involve two modalities, combining different pairs, or three modalities altogether. For example, Xu et al. proposed a bimodal sentiment analysis model based on merged neural networks, extracting sentiment features separately from text and images and then combining them[46]. Poria et al. utilized combined feature vectors from text, visuals, and audio to train a classifier, presenting a parallel data fusion approach[47].

Other tasks related to sentiment analysis include multilingual sentiment analysis[48], geolocation-based sentiment monitoring[49], and fake opinion detection. Fake opinion detection focuses on identifying opinions or comments containing untrustworthy content that distorts the public's perception of events, companies, or products[50].

2.2 Social Media Contextual Sentiment Analysis Techniques

In the contemporary landscape, as natural language processing technologies mature, sentiment analysis has experienced unprecedented development and updates. However, this study does not delve into the step-by-step processes of each algorithm, as numerous research efforts have extensively investigated and analyzed cutting-edge technologies in sentiment analysis from a technical perspective[51]. Scholars have provided detailed

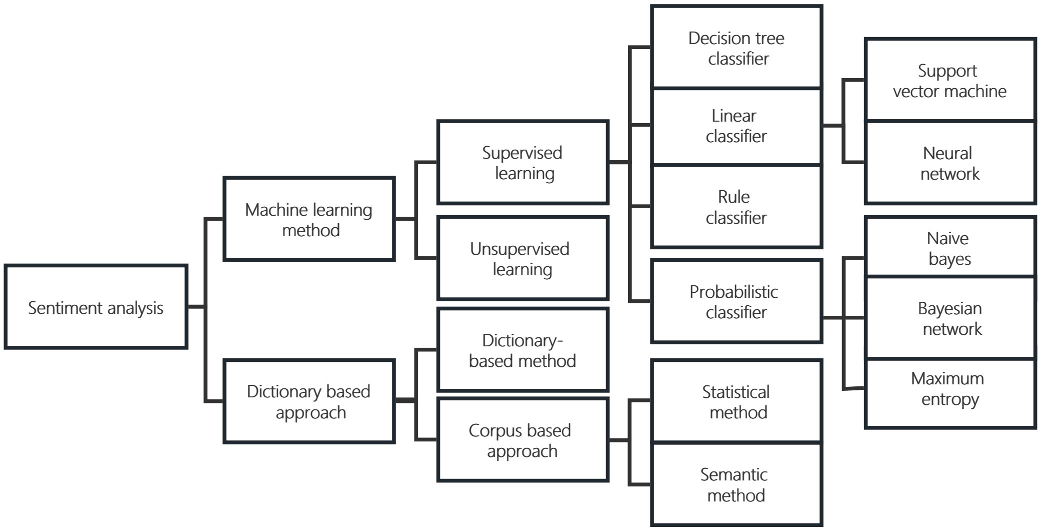

Fig 1. Sentiment Classification Techniques

introductions, evaluations, and comparisons of commonly used sentiment analysis algorithms, including natural language processing techniques[52], machine learning[53], deep learning[54], and more. To identify common themes from these studies, some scholars have classified sentiment analysis research from various angles. Feldman categorizes all sentiment analysis research into five types: document-level, sentence-level, aspect-level, comparative sentiment analysis, and sentiment lexicon construction[55]. Another study outlines more refined classification criteria for sentiment classification techniques[56], including categorization based on dictionaries and machine learning methods, as depicted in Fig. 1. This article draws inspiration from the latter's classification framework, focusing on exploring improvements in commonly used sentiment analysis algorithms for the characteristics of social media, providing technical insights for future relevant research.

2.2.1 Method Based on Dictionary

This method relies on a sentiment dictionary, which is a curated collection of sentiment words, phrases, and even idioms. The construction of sentiment dictionaries can be classified into two categories: dictionary-based and corpus-based methods. The former typically involves manually collecting and annotating initial sentiment words (seeds) and expanding this collection by searching for synonyms and antonyms in dictionaries. One common example is SentiWordNet[57], developed from the well-known WordNet dictionary. The primary drawback of this method is its inability to adapt to domain-specific features and consider contextual nuances. However, even so, it provides a simple and effective solution for sentiment polarity analysis in the context of social media. Corpus-based techniques aim to provide word lists relevant to specific domains. These lists start with a set of seed sentiment words and then utilize statistical methods such as latent semantic analysis to expand the word list by searching for words related to the seeds. The emotional connotations of words often change with their contextual usage, and new words not covered by dictionaries may emerge, especially in the rapidly evolving environment of social media. Therefore, many studies focus on updating dictionaries (e.g., collecting new words, adding emoticons) or dynamically constructing emotional scores for words to enhance sentiment analysis results. Saif et al. use context and semantic information extracted from DBpedia to update the weighted emotional orientation of words and add new words to the dictionary[58]. Another study combines the semantic features of nouns, using information gain and cosine similarity to modify the emotional scores defined in SentiWordNet, thereby improving sentiment analysis performance[59]. Hung argues that high-quality information has a stronger impact on consumer behavior than low-quality information. To apply context information to the domain, preferences of sentiment dictionaries and text quality classification vectors are combined[60]. Emoticons in the context of social media serve as natural emotional labels and are also used to reinforce and construct sentiment dictionaries along with words[61].

2.2.2 Overview of machine learning processes

Machine learning methods can be broadly classified into two categories based on the nature of the data used for learning: supervised and unsupervised learning techniques. Both of these methods rely on the selection and extraction of an appropriate feature set for sentiment analysis. In the feature set, natural language processing techniques play a crucial role, and typical features include N-grams, POS (part-of-speech) features, sentiment word features, syntactic patterns, positional features, conceptual features, and rhetorical features, among others [61]. Among supervised learning techniques, support vector machines, naive Bayes, and maximum entropy are some of the commonly used algorithms[62]. However, due to the lack of fully annotated corpora, researchers have proposed semi-supervised and unsupervised learning methods[63]. Additionally, combining supervised and unsupervised techniques or hybrid methods with dictionaries has been widely applied in sentiment classification, often outperforming the use of dictionaries or machine learning methods alone. For example, Er et al. combined dictionary and machine learning methods to create user profiles, extracting personal preferences to analyze typing habits and emotional fluctuations[64]. Moreover, the depth of the model structure can be used to categorize machine learning into traditional machine learning and deep learning. The general process of utilizing machine learning methods to detect depression in social media text data is depicted in Fig. 2 and involves steps such as data collection, data preprocessing (basic preprocessing and feature engineering), learning text representations using machine learning algorithms, and evaluating the learned model using test data.

Common evaluation metrics for measuring the performance of depression detection algorithms include Accuracy,Precision, Recall, and F_1-score. However, these metrics do not account for the time factor. In response to this, Losada et al. [15] proposed the Early Risk Detection Error (ERDE) metric. This metric takes into account both the correctness of binary decisions and the delay in making decisions by the model. The delay is measured by the number of input texts (posts or comments) the model receives before providing a prediction (k).

Social media content is often short and concise, with a plethora of personal feelings and comments on daily life events. The nature of short texts combined with noise

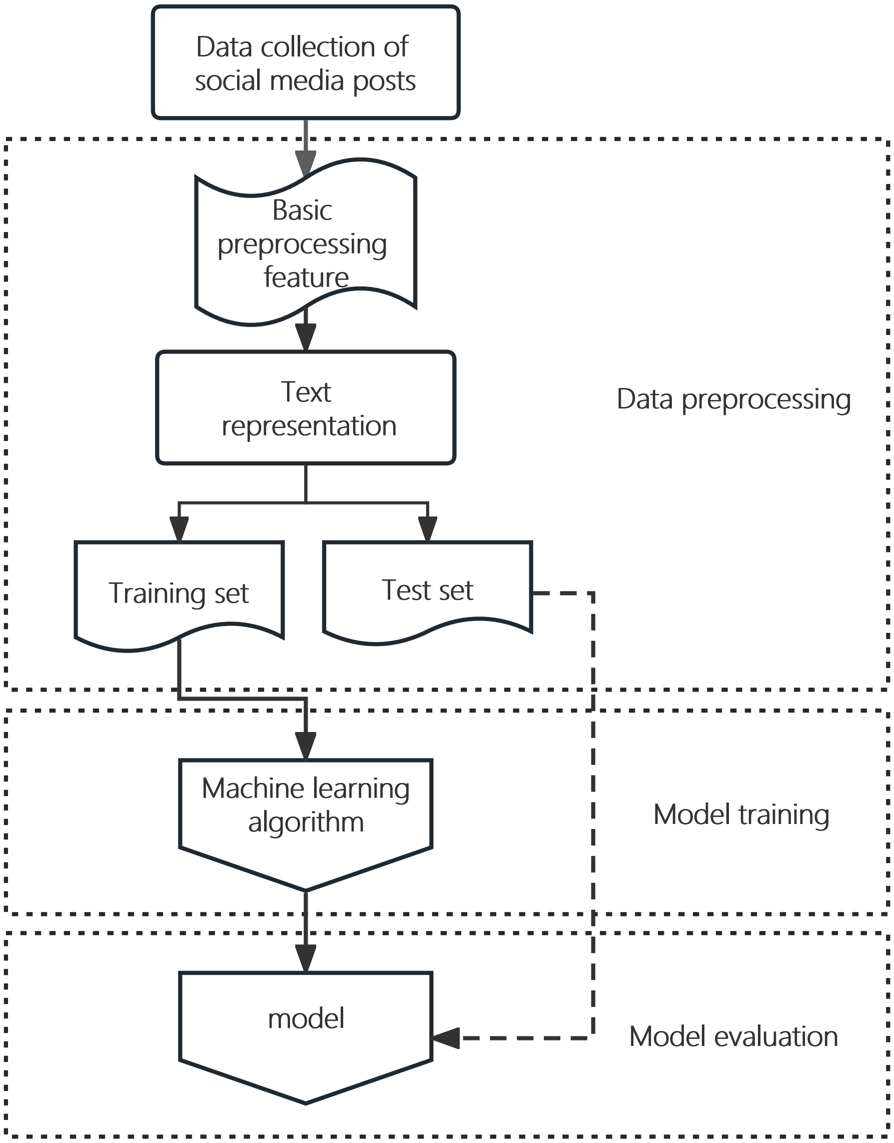

Fig 2. General process of detecting depression using machine learning

poses many challenges for machine learning methods. To compare the performance differences of machine learning algorithms on various datasets, Choi et al. tested several machine learning algorithms on four different social media datasets (IMDB, Twitter, hotel reviews, and Amazon reviews). The results showed that for sentiment analysis to achieve optimal performance, it is necessary to meet the following criteria: ① The training set data should be at least 2%of the dataset; ② The optimal training text length should be between 50-150 characters; ③ Documents with higher subjectivity are more suitable for the training set[65]. These results also suggest that it is crucial to choose more appropriate algorithms based on the different characteristics of social media platforms, rather than simply selecting sentiment analysis techniques based on performance comparisons.

In addition to selecting algorithms based on research goals, combinations of different machine learning algorithms are often used to overcome the issues of classification imbalance and low recall rates in existing single machine learning algorithms. In the performance evaluation of existing Twitter sentiment analysis systems, Zimbra et al. found that three out of the top four systems, which exhibited the best performance, used ensembles of machine learning classifiers[51]. Among them, BPEF (Bootstrapping Ensemble Frame)[66]combines combinations of parameter sets, different classifiers, and feature sets, with average accuracy exceeding 70%for sentiment classification, even outperforming state-of-the-art deep learning methods.

Deep learning methods, inspired by the neural systems of the human brain, have had a significant impact on a range of applications, including natural language processing, speech recognition, and computer vision. They have also been successfully applied in sentiment analysis research. Unlike machine learning, deep learning models do not rely on feature extractors because these features are learned directly during the training process. The main idea behind this work is to use word embedding tools similar to Word2Vec[67]to embed words into neural network models as learned features for sentiment training and classification.

Shirani-Mehr, based on the Stanford Sentiment Treebank, investigated the semantic analysis capabilities of different deep learning models for movie reviews, demonstrating that deep learning networks can automatically extract features and achieve higher performance when learning complex decision boundaries[68]. After confirming the applicability of deep neural networks in extracting sentiment, Panthati et al. used convolutional neural networks and long short-term memory architecture to extract features from customer reviews. The results showed that these two deep learning methods outperformed standalone Naive Bayes and support vector machine classifiers in terms of accuracy[69]. With the growing research interest in deep learning, this technology has rapidly been applied to sentiment analysis, outperforming traditional methods. The advantages of deep learning models include high accuracy, but they also have some notable drawbacks, such as training time consumption and the inability to interpret the semantics of the final decision.

2.3 Data acquisition and preprocessing

2.3.1 Social media text data collection

Social media text data mainly comes from posts and comments posted by users on various social media platforms. Researchers typically obtain data for depression detection from platforms such as Reddit, Twitter, and Sina Weibo by crawling or using APIs. Currently, there are few commonly used public datasets, including the RSDD (Reddit self-reported depression diagnosis) dataset [70], the depression early detection datasets ERiskD[71]and ERiskD 2018[72]from the ERisk (early risk prediction on the Internet) task, the CLPsych 2015 (computational linguistics and clinical psychology shared task dataset CLPD [73]for depression detection, and the MDDL dataset[74]for depression detection created by Shen et al. using the Twitter API. These datasets consist of collections of posts published by users and are typically labeled based on users' self-reported diagnoses (such as "I have been diagnosed with depression") and manual review. The statistical information for each dataset is shown in1

Tab 1. Statistics of common public datasets

Data Set	Number of depressed users	Number of comparison users	Number of posts by depressed users	Number of posts by comparison users
RSDD	9210	107274	8924490	103948506
ERiskD 2017 (Test set/Training set)	83 / 52	403 / 349	30851 / 18706	26417 / 217665
ERiskD 2018 (Test set/Training set)	135 / 79	752 / 741	49557 / 40665	481837 / 504523
CLPD (Test set/Training set)	327 / 150	573 / 300	-	-
MDDL	1402	>300000000	292564	100 million

2.3.2 Data Preprocessing

Raw data undergoes basic preprocessing and feature engineering to generate a structured text table, which is then fed into machine learning models for classification and detection. Basic preprocessing typically involves steps such as data cleaning, tokenization, and standardization, aiming to reduce the interference caused by vocabulary size and non-essential information.

Feature engineering aims to transform text data, either raw or preprocessed through basic steps, into numerical data that computers can comprehend. In natural language processing, text representation can be categorized into basic feature representation, static word embeddings, and contextual word embeddings, as illustrated in Fig. 3. Basic feature representation requires manually constructing features to represent text, commonly used in conjunction with traditional machine learning methods, or as input for deep learning. On the other hand, static word embeddings and contextual word embeddings are typically employed in combination with deep learning approaches.

Basic Feature Representation can extract key information from the text and even consider the order of word occurrences. However, it cannot integrate contextual semantic information, which is crucial in natural language understanding. Static word embeddings express the original meaning of words, word similarity, and even contextual relationships. They are often used in conjunction with deep neural networks and show good performance in natural language processing.

Methods based on contextual embeddings strive to learn the contextual semantics of words as much as possible. Their effectiveness in various natural language processing tasks is attributed to their extensive data, intensive training, model capacity, and the use of unsupervised training methods. This gives them powerful language representation and feature extraction capabilities, leading to excellent performance across multiple natural language processing tasks.

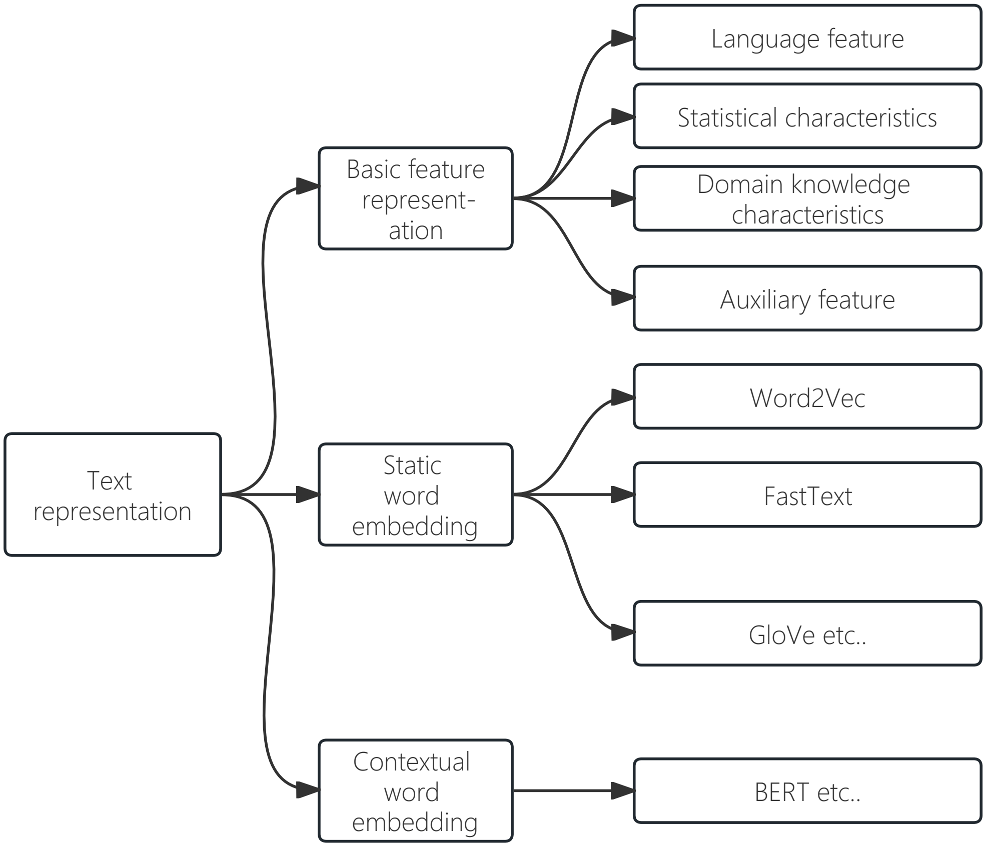
Text representation
Basic feature represent-ation
Static word embedding
Contextual word embedding
Language feature
Statistical characteristics
Domain knowledge characteristics
Auxiliary feature
Word2Vec
FastText
GloVe etc..
BERT etc..

Fig 3. Classification of text representation

2.4 Depression Detection Based on Traditional Machine Learning

The application of traditional machine learning to depression detection using social media text data is primarily divided into two research directions: studies based on different features and studies based on different machine learning algorithms. Research focused on different features for depression detection aims to explore diverse and reliable features, often employing classic algorithms such as support vector machines. On the other hand, research based on different machine learning algorithms emphasizes the construction of more complex and integrated algorithms.

2.4.1 Detection Based on Different Fundamental Features

Before applying traditional machine learning for depression detection, it is necessary to manually construct features from user posts. Different fundamental features and their characteristics are shown in Table 2 Among them, language features can display distinct language styles between individuals with depression and those with mental well-being, thereby revealing different psychological processes. Commonly used language features include Linguistic Inquiry and Word Count (LIWC). LIWC compares words in the text with specific dictionaries, outputting word categories and frequencies. Nguyen et al. [75]demonstrated the strong indicative power of LIWC in predicting depression at the post level. Fatima et al. [76]achieved good discrimination between depression and non-depression posts using LIWC.

Tab 2. Various basic features and their characteristics

Basic Feature	**Representative Use Case**	**Characteristics**
Language feature	LIWC etc.	Having the ability to explain depression, it is user-friendly, but not suitable for casual documents such as social media.
Statistical characteristics	BOW; TF-IDF; *N*-Gram etc.	Capable of fully utilizing the original meanings of keywords, with strong versatility; however, it can only represent text based on frequency and word order, lacking effective utilization of contextual information.
Domain knowledge characteristics	Themes and emotions, etc.	Strongly related to the field of depression, capable of effectively explaining the differences between patients and healthy users.
Auxiliary feature	User behavior and generationActive mode	Often used as supplementary information with other characteristic junctions Combined use

The language features provide the ability to explain depression, and they can be used solely by analyzing word semantics. However, they are more suitable for formal documents such as news articles rather than informal or colloquial documents like social media posts. Compared to language-pattern-based methods, statistical features such as bag-of-words (BOW) and term frequency-inverse document frequency (TF-IDF) make better use of the original meanings of keywords by counting word frequencies[77], and they are more versatile. Prieto et al.[78]used a simple bag-of-words model, extracted N-gram features, and applied correlation-based feature selection for depression detection, achieving good classification accuracy and speed improvement. Dos Santos et al.[79]found that TF-IDF can make potentially useful predictions from very small datasets.

For the detection of mental disorders, knowledge features in areas such as themes and emotions show good effectiveness. Typically, depression patients have different interests in topics compared to mentally healthy users, allowing effective differentiation based on the differences in discussed themes. For example, Nguyen et al. [80]found that thematic and language psychological features are highly effective predictors, achieving good results in post-level depression detection by combining both features. Emotion-based features can provide information from more abstract emotional aspects and are more relevant, effectively revealing differences between depression patients and mentally healthy users. For instance, Chen et al. [81], building upon LIWC, introduced a set of fine-grained emotion features, demonstrating the effectiveness of emotion features. Leiva et al. [82], while incorporating TF-IDF, also introduced three emotion polarity features (positive, neutral, negative emotions), proving that methods incorporating sentiment analysis are more accurate than those relying solely on TF-IDF.

In addition to utilizing language, statistical, and domain knowledge features, many scholars have explored auxiliary features. Auxiliary features, such as user behavior and lifestyle patterns, often serve as supplements to the above features, allowing a more realistic and detailed comparison between depression users and healthy users, with more comprehensive information available. Hu et al. [83], building on language features, incorporated behavioral features and compared the classification accuracy of models under different observation windows, finding that language and behavioral features can accurately identify whether a user is depressed, with the best performance observed at a 2-month observation time. Chen et al. [81]combined LIWC with lifestyle features, demonstrating the effectiveness of combined features.

Overall, in depression detection based on social media text data, the most primitive single feature often lacks sufficient information, leading to the continuous exploration and addition of more features. Under comprehensive features, various user information can be utilized, but too many or even redundant features can decrease model efficiency. Therefore, in the field of depression detection using traditional machine learning methods, determining which features to construct and how to select representative features remains an important issue. Additionally, considering how to build suitable learning algorithms to match the selected features, thereby allowing the model to perform better, is also worth considering.

2.4.2 Detection based on different types of algorithms

In machine learning, the construction and selection of features are crucial, and the choice and improvement of learning algorithms are equally important; the two complement each other. In the detection of depression based on social media text data, researchers aim to match various features to improve detection performance, address practical issues such as limited labeled data and lack of support for incremental learning, and enable early detection of depression.

Comprehensive features can encompass information about depression users, but not all learning algorithms can effectively match them to produce good results. To address this, many researchers have explored various approaches. For example, Peng et al. [84] proposed using a multi-kernel support vector machine for depression text classification based on user profile features, user behavior features, and post text features. The multi-kernel support vector machine can adaptively select the optimal kernel for different features, resulting in better performance compared to a single-kernel support vector machine. Although the multi-kernel support vector machine performs well, it still has some limitations, such as being unsuitable for larger datasets and being more sensitive to missing data. Ensemble learning can overcome the limitations of a single classifier, thereby improving detection performance and generalization. For instance, Liu et al. used feature selection methods, treating multiple single classifiers as base learners and employing logistic regression as a combination strategy to build a stacked model. The proposed model not only reduces data dimensionality, improving model efficiency, but also overcomes the limitations of individual models, enhancing model generalization, achieving an accuracy rate of 90.27% in identifying depression patients.

Classical machine learning for depression identification on social media either requires sufficient historical data or does not support incremental learning. To address these issues, Tariq et al. [85] introduced a semi-supervised joint training model that combines random forest, support vector machine, and naive Bayes. The proposed model requires only a small amount of labeled data to label a large amount of unlabeled data, saving significant human resources. Burdisso et al. [86]proposed the SS3 model, which supports incremental training on text streams and has achieved advanced performance in early detection of depression. Although the SS3 model performs well, one drawback is that the model's input section uses a bag-of-words approach, making it unable to consider issues such as text word order.

Classical depression detection methods lack timeliness because detecting depression requires patients to first recognize their mental issues and then overcome shame to seek medical help, a process that often takes a long time. Typically, when patients are diagnosed with depression, it has already reached a severe level or even includes suicidal tendencies. Considering these issues, many researchers have investigated early detection of depression. Briand et al. [87] suggested that if posts from new users are semantically close to posts from at-risk users, then new users may also be at risk of depression. To achieve this, they built an information retrieval subsystem and a supervised learning subsystem, with the predictions from each subsystem merged using a decision algorithm. The proposed model can not only detect the condition of existing users but also early detect depression in new users. Cacheda et al. [88] proposed the twin-case method for early detection of depression. The twin-case method uses two independent random forest classifiers, one for detecting depressed individuals and the other for identifying non-depressed individuals. The two options (depressed and non-depressed) are independently predicted, avoiding the delays caused by the mutual competition of the two options in single-case methods. The results indicate that the twin-case method's performance is significantly better than that of single-case methods, improving current state-of-the-art model detection performance by over 10% .

In summary, in the application of traditional machine learning for depression detection, feature construction and selection have become comprehensive and mature, and algorithms that match multiple features have achieved good results. However, current research has relatively fewer explorations into practical issues such as limited labeled data, which should be strengthened in the future. Additionally, some researchers have explored early detection of depression and proposed novel methods, but overall, there is still room for improvement in the effectiveness of such algorithms.

The summary of traditional machine learning algorithms in depression detection is shown in Table 3.

Tab 3. Summary of traditional machine learning algorithms for depression detection

Main Algorithm	Year	Data Source	Advantage	Limitation
Multi-core SVM	2019	Sina Weibo	Ability to explain depression; User-friendly; Not suitable for casual documents like social media	Changes in depression levels and dynamic status over time were not considered
Feature selection and stacking integration strategy	2021	Weibo	Eliminate redundant features; Can make up for the shortcomings of multiple single classifiers	Small data set; Depressed users are young and not representative of the general population
Semi-supervised joint training	2019	Reddit	The model is more robust with less dependence on annotated data	The influence of mood change on depressive symptoms was not considered; Text word order was not considered; Parameter optimization was not performed
SS3	2019	ERiskD 2017	Supports incremental learning, interpretable, and has low computational cost	The influence of mood change on depressive symptoms was not considered; Text word order was not considered
Information retrieval and supervised learning	2018	ERiskD 2017	Can detect semantic connections between posts from new users and risk users	Search engines do not add indexes to features related to depression; The classifier is limited
Bidirectional learning	2019	ERiskD 2017	The reward and punishment detection efficiency is better than the singleton method	Emotional semantic features are not considered

2.5 Depression detection based on deep learning

Traditional machine learning requires the manual construction of a large number of features. However, building effective features often consumes a significant amount of time and effort for researchers. In contrast, deep learning can automatically extract features based on raw text vectors and has the ability to abstract and generalize information. In many cases, especially when dealing with large datasets, deep learning demonstrates excellent performance. In the context of depression detection based on social media text data, common deep learning algorithms include Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), algorithms incorporating attention mechanisms, and Transformer-based models like BERT.

2.5.1 Depression detection based on CNN

In depression detection based on social media text data, CNN has been studied and used due to its strong feature extraction ability. The basic framework of depression detection using CNN is shown in Fig. 4. Text data is transformed into numerical data by word embedding technology to form word embedding matrix. Then, multiple convolution kernels of different sizes are used for convolution operation. Finally, the binary classification results are output through the pooling layer and the fully connected layer.

In their application, Trotzek et al. [89] utilized FastText pre-trained word embeddings based on Wikipedia as input for CNN. Simultaneously, they employed logistic regression to handle user-level language metadata. The outputs from both components were then fused for classification. The results showed that the constructed model exhibited the best overall performance in early detection of depression. Considering the common issue of class imbalance in real-world data, Kim et al. [90] introduced SMOTE (synthetic minority oversampling technique) on top of CNN to overcome performance loss caused by imbalanced data categories.

During the feature extraction process using CNN, gating units play a crucial role

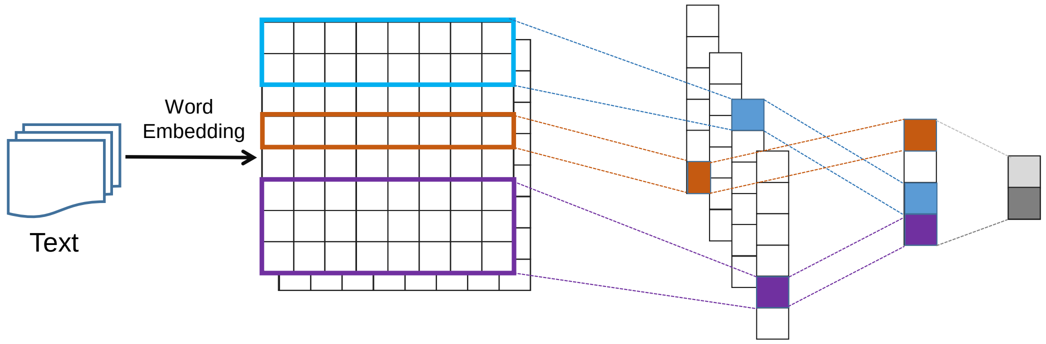

Fig 4. Depression detection framework based on CNN

in highlighting important information while filtering out less relevant details. This helps identify key influencing factors and reduces the model's parameter count, leading to further performance improvement. Rao et al. [91] incorporated gating units into CNN, combining the strong feature extraction capability of CNN with the ability to selectively capture crucial emotional information from user posts. This integration enhances the model's detection performance and stability by filtering out less important information.

2.5.2 Depression detection based on RNN

While CNN can extract local information from text and has excellent parallel computing capabilities, it struggles to capture long-distance textual semantic information. In comparison, RNN, due to the introduction of memory units, has an advantage in processing textual data by retaining previous information. The basic framework of RNN is illustrated in Fig. 5. RNN units sequentially read the word embedding information of each word, where hi represents the output unit of the hidden layer containing information from the previous time step, h(i-1). Traditional RNNs face the issue of gradient vanish-

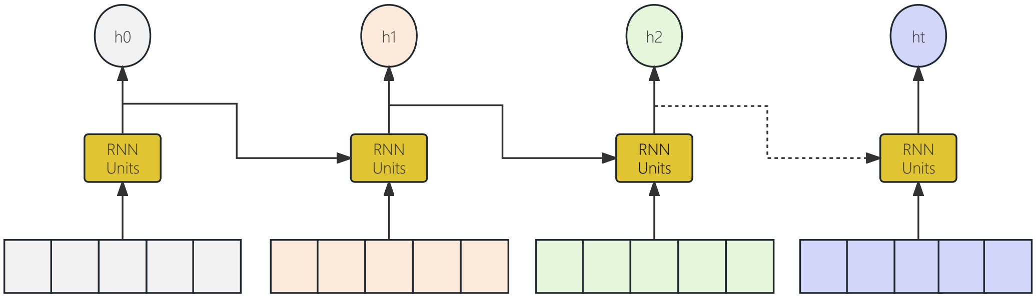

Fig 5. Basic framework of RNN

ing, leading researchers to propose variant models like LSTM (long short-term memory) and GRU (gated recurrent unit) to address this problem.

In the application of RNN and its variants, such as LSTM, for depression detection, Amanat et al. [92] developed the RNN-LSTM model, demonstrating superior performance compared to CNN. BiLSTM, in contrast to LSTM, incorporates training for both preceding and succeeding contexts, effectively utilizing semantic information from both directions and enhancing the model's performance in sequence classification problems. Ahmad et al. [93] proposed the use of BiLSTM for depression detection, showing that BiLSTM outperforms LSTM in various metrics, although they did not address the issue of imbalanced data categories. Cong et al. [94] constructed the X-A-BiLSTM model, discovering that employing XGBoost on top of BiLSTM helps alleviate the problem of data imbalance.

2.5.3 Depression Detection Based on CNN-RNN and Attention Mechanism

In theory, the CNN-RNN architecture combines the excellent feature extraction capabilities of CNN with the sequence modeling capabilities of RNN. In depression detection based on social media text data, researchers have explored this architecture. Aragón et al. [95] transformed the content of user posts into sub-emotion sequences. After feature extraction using CNN, a bidirectional gated recurrent unit (BiGRU) captured the context of sub-emotion sequences. Finally, an attention mechanism was employed to extract important sub-emotions from sentences. The proposed model showed an accuracy improvement of 7% and 12% compared to standalone CNN and RNN, respectively. Moreover, when dealing with smaller datasets, standard CNN and RNN performance was inferior to traditional machine learning methods.

Zogan et al. [96] constructed the DepressionNet framework, which combines stacked BiGRU for handling user behavior features and a combination of CNN with attention-enhanced BiGRU for extracting summaries of user posts. By fusing user behavior and posting history, this framework automatically detects depression. Experimental results indicated that the CNN+BiGRU model achieved good accuracy, and the proposed model outperformed CNN+BiGRU by at least 2% in various metrics.

In depression detection, attention mechanisms allocate weights to information, prioritizing important information related to depression. Given that many mental health patients express their feelings and emotions indirectly through metaphors on social media [97, 98], Zhang et al. [99] introduced the Metaphor-Based Attention Model (MAM). The MAM model utilized the Recurrent Neural Network Multi-Head Contextual Attention (RNN_MHCA) [100, 101] module to acquire sentence and text metaphor features, calculating attention weights based on metaphor features. Experimental results indicated that the attention-based MAM model effectively learned implicit emotional information from users, confirming the effectiveness of metaphorical information in depression detection. Similarly, Almars [102] proposed using attention mechanisms to analyze Arabic text data related to depression. By adding an attention mechanism to BiLSTM, the model learned crucial hidden features of depression, outperforming BiLSTM by 3% in accuracy. Ren et al. [103] introduced the Emotion-Based Attention Network (EAN) model, incorporating attention mechanisms. Through model comparisons, Ren et al. demonstrated that attention mechanisms effectively improved model performance, confirming the effectiveness of emotional semantic information in depression detection.

Attention mechanisms not only enhance model performance but also, through visualizing their weight scores, analyze words and sentences strongly associated with depression. Song et al. [104] proposed the Feature Attention Network (FAN), which integrated features such as depression symptoms, emotions, rumination, and writing style to simulate the diagnostic process of experts. The FAN model generated interpretability by analyzing attention weights, confirming the crucial role of emotional information in depression detection, although the overall model performance was not outstanding. Uban et al. [105] , combining emotional information, applied the Hierarchical Attention Network (HAN) to depression detection. By analyzing the abstract representations of data in the network layers, the model provided comprehensive explanations of predictions. However, the HAN model focused more on language-related information and neglected the modeling of user behavior, time, and other features. Zogan et al. [106] introduced the Multi-Aspect Depression Detection Hierarchical Attention Network (MDHAN), a hybrid model based on HAN. The model combined features from text, behavior, time, and semantics, improving predictive performance. Through the analysis of attention weights, the model explained its prediction method. However, the MDHAN model lacks an analysis of emotional aspects.

2.5.4 Depression Detection Based on BERT

The Transformer model utilizes self-attention encoders to autonomously explore correlations between words within the same sentence, thereby obtaining more profound encoding information. Additionally, Transformer completely discards the use of structures resembling recurrent neural networks, significantly enhancing computational speed and the ability to process long sentences. Built upon a two-layer bidirectional Transformer structure, BERT, a pre-trained language model, possesses a robust capability for modeling semantic information, as illustrated in Fig. 6. BERT requires the addition of identifiers before and after the sentence as separators, followed by using word position information, segment information, and word embeddings as inputs to the two-layer Transformer encoders. BERT can function both as a word embedding technique and, when

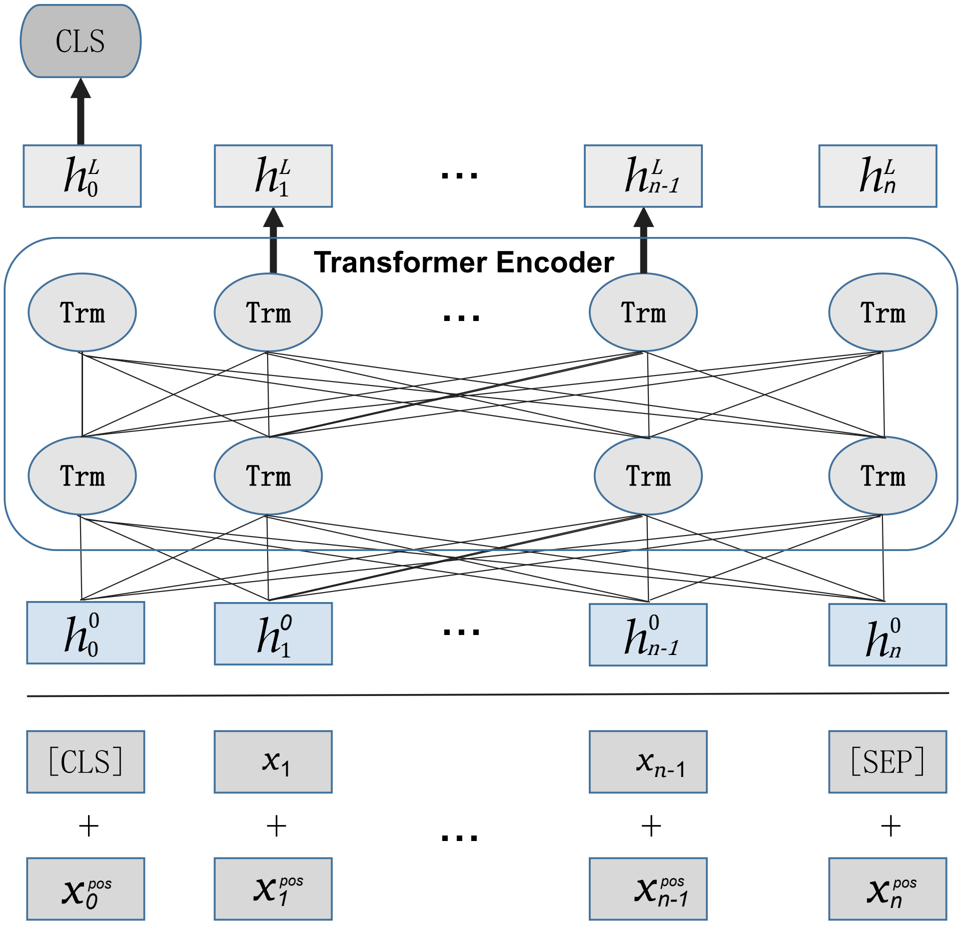

Fig 6. Structure of BERT model

equipped with a simple classifier afterward, as a classification model.

In the field of depression detection, Yadav et al. [107] pioneered a new BERT-based multi-task learning framework called FiLaMTL (Figurative Language Enabled Multi-Task Learning Framework). This framework accurately identifies depression symptoms by incorporating an auxiliary task of detecting figurative language usage. The research results indicate that BERT has a strong feature extraction capability. However, BERT trained on general corpora may not adapt well to specific domains. Moreover, the experimental results demonstrate the effectiveness of introducing figurative language detection for identifying depression symptoms. Domain-specific pretraining, as explored by Wang et al. [108] using BERT on depression datasets, outperforms all proposed Transformers-based models in depression detection and severity classification tasks. To address issues such as the massive size of the classical BERT model making it challenging to deploy in practical applications, Zeberga et al. [109] proposed a new framework that applies knowledge distillation techniques to transfer knowledge from a large pre-trained network (BERT) to a smaller one (Distilled_BERT). Compared to BERT, Distilled_BERT not only further enhances detection performance but also has a relatively smaller model size. In the application of an improved BERT structure, Khan et al. [110] used the DeBERTa (Decoding-Enhanced BERT with Disentangled Attention) model to differentiate depression from other diseases. DeBERTa introduces a disentangled attention mechanism and an enhanced mask decoder, enabling simultaneous consideration of content, relative position, and absolute position information of vocabulary, thus fully learning the content and dependencies of words. In comparisons with multiple advanced models, this model performs best in distinguishing depression from other diseases.

In summary, researchers exploring depression detection using deep learning models have achieved good results by addressing issues such as balancing data categories, feature extraction methods, and incorporating multidimensional features. Overall, compared to traditional machine learning, deep learning has stronger stability and generalization capabilities due to its ability to automatically extract features, achieving more outstanding detection performance. However, deep learning models have relatively large parameter sizes and often require the support of large-scale data; their performance on small datasets may not match that of traditional machine learning. In deep learning methods, attention mechanisms and BERT pretraining models deserve attention. Attention mechanisms enhance model performance and provide interpretability for model predictions, holding potential for clinical applications. BERT-like models, although having strong feature extraction capabilities and achieving considerable performance by extracting key information representing depression in the text, possess complex structures and large model parameters, making them less suitable for retraining. Additionally, using generic pre-trained BERT models may lead to performance losses, especially in medical domains with specific characteristics like depression.

The summary of deep learning algorithms in depression detection is shown in Table 4,5,6.

Tab 4. Summary of deep learning algorithms for depression detection-1

Algorithm Structure	**Year**	**Data Source**	**Advantage**	**Limitation**
CNN	2018	ERiskD-2017	Adds user-level language metadata, the model performance is improved; User-friendly; Not suitable for casual documents like social media	Convergence strategy obsolescence
CNN+SMOTE	2020	Reddit	Alleviating the problem of unbalanced data categories	Time series information that does not consider the vertical behavior pattern of users
MGL-CNN	2020	RSDD	Add a gated unit to effectively capture key emotional information in your posts	Unable to trace connections between distant words
RNN-LSTM	2022	Kaggle	Better performance in processing text sequence information than CNN	Data set limitations do not adequately reflect the course of depression
BiLSTM	2020	Orabi and CLPsych	Better understanding of contextual semantics	Limited data set, not counted
X-A-BiLSTM	2018	RSDD	Alleviating data set imbalances in BiLSTM	The information loss is serious when the text is long
CNN+BiGRU+Att	2020	ERiskD 2018	It can fully capture the fine grained emotions of depressed patients	Post information is not used enough
DepressionNet	2021	MDDL	Condense information by refining summaries	The structure of the model is complex and does not consider the URL content of the post

Tab 5. Summary of deep learning algorithms for depression detection-2

Algorithm Structure	**Year**	**Data Source**	**Advantage**	**Limitation**
MAM	2021	ERiskD 2018	Innovatively proposed metaphor-based method	TThe RNNR module in the model is not advanced enough: it fails to explain metaphor and essence
BiLSTM+Att	2022	Twitter(Arabic) 2022	Effective learning of important hidden features of depression	The data set is small and the model generalization is weak
EAN	2022	Reddit	It can capture both contextual semantic information and affective semantic information	Emotion analysis is not sophisticated enough
FAN	2018	RSDD	Strong interpretability of simulated expert diagnosis process	Not incorporating more posts from users results in poor model performance
MDHAN	2022	RSDD	Considering many comprehensive features, it is interpretable	User posts are not analyzed in relation to subject matter and emotion
FiLaMTL	2020	CLPsych 2015	The detection of figurative usage is introduced to improve the robustness and reliability of the model	The model structure is complex. Without the introduction of memetic modes, the model's understanding of metaphor is not deep enough

Tab 6. Summary of deep learning algorithms for depression detection-3

Algorithm Structure	**Year**	**Data Source**	**Advantage**	**Limitation**
BERT_IDP	2020	China Weibo	It has a strong feature extraction ability and divides the degree of depression	Medical knowledge of depression is not included without consideration of user-level context
BERT+Knowledge distillation	2022	Reddit+Twitter	Applying knowledge distillation technology to compress model parameters	User-level context is not considered

2.6 Neuro-Symbolic AI

Neuro-Symbolic AI is an interdisciplinary approach that combines elements of symbolic reasoning and neural network-based approaches. It aims to integrate the strengths of both symbolic AI, which excels in logic and reasoning, and neural networks, which are powerful in learning from data. This combination seeks to create more robust and versatile AI systems.

2.6.1 Symbolic Reasoning

Symbolic reasoning plays a crucial role in the field of artificial intelligence, relying on operations based on symbols and logical rules to simulate human reasoning processes. By utilizing formal symbolic representations, systems can perform various complex logical operations, making the handling of knowledge and problem-solving more flexible. A classic example is expert systems[111], which employ rule-based reasoning using a knowledge base to generate meaningful outputs based on input symbolic information.

Over the past few decades, symbolic reasoning has made significant strides in areas such as knowledge representation, problem-solving, and planning. For instance, classical reasoning engines like CLIPS and Prolog[112] have found widespread application in professional domains.

3 TAM Network for Early Detection of Depression

3.1 Introduction

Early risk prediction on the Internet (eRisk) has been a long-running Lab at CLEF [113, 114, 115, 116, 117, 118], which aims at exploring the early detection technologies to predict potential risks in the Internet users' health and safety. In this year, the Early Detection of Depression task at the CLEF eRisk 2022 Lab [118] focuses on predicting the depression risk in users based on their social media postings. A user is depression-positive if an explicit mention of being diagnosed with depression was made by the user [113, 114]. By observing the posts of a user from the very beginning, a detection system needs to raise the risk *decision* as early as possible if the user is depression-positive and to estimate a risk-ranking *score* indicating the level of depression.

The early studies of language usage in depression patients [119, 120, 121, 122, 123] suggest that depression and language usage are internally correlated, while the recent psychological studies of depression [124, 125] indicate that depression is indeed a complex emotional state and highly associates with several negative emotions [126, 127], such as sad and anxiety. These findings have inspired recent studies to explore linguistic features [128, 129, 130, 131, 11, 132], emotions, and sentiments [133, 134, 135] in user posts for detecting depression and several related mental disorders, such as suicide ideation [136].

We extend these studies by exploring the history of user affective states, based on the connection between depression and the long-term negative affects reflected in one's posts, that is, the difficulty of removing negative feelings from one's working memory [127, 137]. We consider affective state as the embedding of user emotion in a post, which is retrieved by a pre-trained DistilBERT-Emotion model. A Time-Aware Affective Memories (TAM) network is proposed to maintain the memory of an Internet user's affective state, which gets update with the user's latest affective state and the time interval $\Delta\tau_t$ between the user's latest (τ_t) and last (τ_{t-1}) postings. This affective memory is fed together with the semantic information of the latest post to a Transformer Decoder, and TAM uses the decoded information to predict a user's depression risk.

To encourage early detection of the depression risk, we propose a latency penalty that penalizes the latency of the *first-positive* predictions for the depression-positive users. Our initial experiment suggests that latency penalty is effective for reducing the Early Risk Detection Error (ERDE) score for the Early Detection of Depression task.

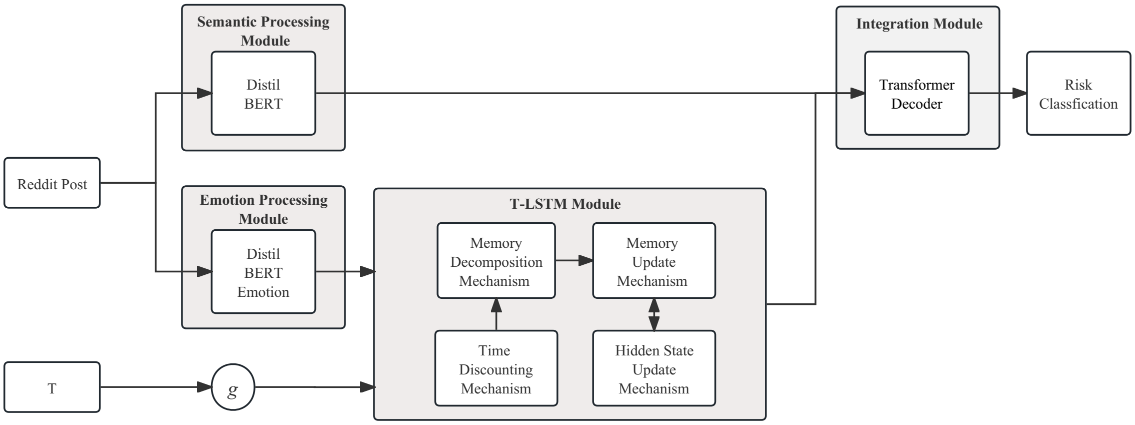
Semantic Processing Module
Distil BERT
Integration Module
Transformer Decoder
Risk Classfication
Reddit Post
Emotion Processing Module
Distil BERT Emotion
T-LSTM Module
Memory Decomposition Mechanism
Memory Update Mechanism
Time Discounting Mechanism
Hidden State Update Mechanism
T
g

Fig 7. Architectural design of the Time-Aware Affective Memories (TAM) network, showcasing its four core components: the Emotion Processing Module, the Time-Aware Long Short-Term Memory (T-LSTM) Module, the Semantic Processing Module, and the Integration Module.

3.2 Time-Aware Affective Memories Network

To explore the history of user affective states for Early Detection of Depression, we propose a Time-Aware Affective Memories (TAM) network as shown in Fig. 7. TAM is composed of an affective processing module and a semantic processing module, which are indicated in the light green and the light blue squares, respectively.

First, the affective processing module expects the latest post $x_t^{(i)}$ from user i at step t and the time interval $\Delta\tau_t^{(i)}$ between the user's latest and last postings as input. We concatenate the title and body of a post into $x_t^{(i)}$, with the user-sensitive and task-insensitive information replaced with special tokens[1]. The time interval $\Delta u_t^{(i)}$ is given by

$$\Delta\tau_t^{(i)} = \tau_t^{(i)} - \tau_{t-1}^{(i)}, \tag{1}$$

where $\tau_t^{(i)}$ and $\tau_{t-1}^{(i)}$ are the time logs of the user's latest and last postings.

Second, the user's emotion in post $x_t^{(i)}$ is mapped into an affective state $A_t^{(i)}$ based on a pre-trained DistilBERT Emotion classification model DistilBERT$_E$ [2]. The mapping is

[1] User-sensitive Email addresses and phone numbers are replaced with ⟨EMAIL⟩ and ⟨PHONE⟩, and task-insensitive numbers and currency symbols are replaced with ⟨NUMBER⟩ and ⟨CUR⟩, respectively with clean-text.

[2] `https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion`

given by

$$A_t^{(i)} = \varphi\left(\text{DistilBERT}_E(x_t^{(i)})\right), \tag{2}$$

where the affective state $A_t^{(i)} \in \mathbb{R}^{d_{\text{BERT}}}$ corresponds to a φ-pooled activation of the pre-classification layer in DistilBERT_E with input $x_t^{(i)}$, $\varphi(\cdot)$ corresponds to either a mean-pooling or a CLS-pooling among the first dimension of a tensor, and d_{BERT} is the DistilBERT model dimension. DistilBERT_E is pre-trained on an English Twitter Emotion dataset [138], which classifies user postings into *joy*, *love*, *surprise* , *sadness*, *anger*, and *fear*. The pre-trained DistilBERT is slightly inferior to that of BERT in emotion classification but is over two times faster in processing speed.

Third, the affective states $A^{(i)}$ of user i is remembered by a Time-Aware LSTM (T-LSTM) network [139]. T-LSTM takes the affective state $A_t^{(i)} \in \mathbb{R}^{d_{\text{BERT}}}$ for the current post $x_t^{(i)}$ as the first input and discounts its internal affective memory in $C \in \mathbb{R}^{d_{\text{MEM}}}$ with the time interval $\Delta\tau_t^{(i)} \in \mathbb{R}_{>0}$ as the second input. In the following description we omit the user index i for abbreviation. Given the internal memory $C_{t-1} \in \mathbb{R}^{d_{\text{MEM}}}$ and the hidden state $h_{t-1} \in \mathbb{R}^{d_{\text{MEM}}}$ at the last step $t-1$ as well as inputs $A_t^{(i)}$ and $\Delta\tau_t^{(i)}$ at the latest step t, T-LSTM updates its internal memory and hidden state by

$$C_{t-1}^S = \tanh(W_d C_{t-1} + b_d) \qquad \text{(Short-term memory)}$$

$$\hat{C}_{t-1}^S = C_{t-1}^S g_*(\Delta\tau_t) \qquad \text{(Discounted short-term memory)}$$

$$C_{t-1}^L = C_{t-1} - C_{t-1}^S \qquad \text{(Long-term memory)}$$

$$C_{t-1}^A = C_{t-1}^L + \hat{C}_{t-1}^S \qquad \text{(Adjusted previous memory)}$$

$$f_t = \sigma(W_f A_t + U_f h_{t-1} + b_f) \qquad \text{(Forget gate)}$$

$$i_t = \sigma(W_i A_t + U_i h_{t-1} + b_i) \qquad \text{(Input gate)}$$

$$\begin{aligned}
o_t &= \sigma(W_o A_t + U_o h_{t-1} + b_o) && \text{(Output gate)} \\
\tilde{C}_t &= \tanh(W_c A_t + U_c h_{t-1} + b_c) && \text{(Candidate current memory)} \\
C_t &= f_t C^*_{t-1} + i_t \tilde{C}_t && \text{(Current memory)} \\
h_t &= o_t \tanh(C_t), && \text{(Current hidden state)}
\end{aligned}$$

where $W_d \in \mathbb{R}^{d_{\text{MEM}} \times d_{\text{MEM}}}$ and $b_d \in \mathbb{R}^{d_{\text{MEM}}}$ are parameters for decomposing the memory. $W_f, W_i, W_o, W_c \in \mathbb{R}^{d_{\text{BERT}} \times d_{\text{MEM}}}$, $U_* \in \mathbb{R}^{d_{\text{MEM}} \times d_{\text{MEM}}}$, and $b_f, b_i, b_o, b_c \in \mathbb{R}^{d_{\text{MEM}}}$ are parameters for calculating the forget, input, output gates and the candidate current memory, respectively. g_* is a set of discount functions that monotonically decrease with the time interval $\Delta\tau_t$. We employ two discount functions g_{slog} and g_{flex} for the detection of depression task, in which g_{slog} is reciprocal to the logarithm of interval seconds

$$g_{\text{slog}}(\Delta\tau) = 1/\log(\Delta\tau + \varepsilon), \tag{3}$$

with a hyper-parameter ε of 1.0, and the g_{flex} is a flexible power function of the interval seconds inspired by [140]

$$g_{\text{flex}}(\Delta\tau) = \frac{q_1}{a\Delta\tau} + \frac{q_2}{1 + (\Delta\tau/b)^c}, \tag{4}$$

with trainable parameters $q_1, q_2, a, b, c \in \mathbb{R}$.

Last, a linear layer is employed to map the T-LSTM hidden state $h_t^{(i)} \in \mathbb{R}^{\mathbb{R}^{d_{\text{MEM}}}}$ to an affective memory $M_t^{(i)} \in \mathbb{R}^{d_{\text{BERT}}}$ by

$$M_t^{(i)} = W_M h_t^{(i)} + b_M, \tag{5}$$

with parameters $W_M \in \mathbb{R}^{d_{\text{MEM}} \times d_{\text{BERT}}}$ and $b_M \in \mathbb{R}^{d_{\text{BERT}}}$. To enrich the memorization of a user's affective states for TAM, we concatenate the most recent l_{MEM} affective memories by

$$\hat{M}_t^{(i)} = \text{Concat}(M_{t-l_{\text{MEM}}+1}^{(i)}, \ldots, M_t^{(i)}), \tag{6}$$

where $\hat{M}_t^{(i)} \in \mathbb{R}^{l_{\text{MEM}} \times d_{\text{BERT}}}$ is the *enriched* affective memory.

The **semantic processing module** takes the latest post writing $x_t^{(i)}$ from user i at time step t as input, which is similar as the affective processing module, and encodes it to a semantic embedding $S_t^{(i)}$ with a pre-trained DistilBERT model[3] by

$$S_t^{(i)} = \text{DistilBERT}(x_t^{(i)}), \tag{7}$$

where $S_t^{(i)} \in \mathbb{R}^{l_{\text{TEXT}} \times d_{\text{BERT}}}$ corresponds to the activation of the pre-classification layer in DistilBERT, l_{TEXT} corresponds to the length of $x_t^{(i)}$, and d_{BERT} is the DistilBERT model dimension.

To integrate the *enriched* affective memory $\hat{M}_t^{(i)}$ and the semantic embedding $S_t^{(i)}$ in TAM, we employ a Transformer Decoder network as shown in Fig. 7. We denote $\hat{M}_t^{(i)}$ and $S_t^{(i)}$ as A and B for illustrating the integration mechanism as below. A Transformer Decoder is a multi-head cross-attention architecture, each head of which makes queries for elements from an input sequence A and retrieves new values from a reference input sequence B, based on the element-wise similarity between A and B. Specifically, the cross-attention $\text{MultiHead}(A,B)$ is concatenated by $\text{head}_1, \ldots \text{head}_H$ with

$$\text{MultiHead}(A,B) = \text{Concat}(\text{head}_1, \ldots \text{head}_H)W^O, \tag{8}$$

$$\text{head}_h = \text{Attention}(AW_h^Q, BW_h^K, BW_h^V), \tag{9}$$

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^\intercal}{\sqrt{d_2}})V, \tag{10}$$

where $A \in \mathbb{R}^{n \times d_1}$, $B \in \mathbb{R}^{m \times d_1}$ are sequences of n and m embeddings and d_1 is the embedding dimension. To empower the attention mechanism, A and B are first mapped from the d_1-dimensional space to query $Q_h \in \mathbb{R}^{n \times d_2}$, key $K_h \in \mathbb{R}^{m \times d_2}$, and value $V_h \in \mathbb{R}^{m \times d_2}$ in a larger d_2-dimensional space through linear projection with parameters $W_h^Q, W_h^K, W_h^V \in \mathbb{R}^{d_1 \times d_2}$, and h is the index of attention heads. Each $\text{head}_h \in \mathbb{R}^{m \times d_2}$ is then calculated by the Attention function with the corresponding query, key, and value as the input. Last,

[3] https://huggingface.co/distilbert-base-uncased

the concatenated attention head is mapped from $m \times d_2$ back to d_1 dimension with a projection parameter $W^O \in \mathbb{R}^{Hd_2 \times d_1}$.

We propose to integrate the affective memory and semantic embedding with either one Transformer Decoder by

$$H_t^{(i)} = \text{mean}(\text{MultiHead}(\hat{M}_t^{(i)}, S_t^{(i)})), \tag{11}$$

or two Transformer Decoders by

$$\begin{aligned} H_t^{(i)} = \text{mean}\left(\text{MultiHead}\left(\text{mean}(\hat{M}_t^{(i)}), S_t^{(i)}\right)\right) + \\ \text{mean}\left(\text{MultiHead}\left(\varphi(S_t^{(i)}), \hat{M}_t^{(i)}\right)\right), \end{aligned} \tag{12}$$

where $\text{mean}(\cdot)$ indicates a mean-pooling in the first dimension of a tensor while $\varphi(\cdot)$ corresponds to either a mean-pooling or a CLS-pooling. Both decoding strategies render an integration $H_t^{(i)} \in \mathbb{R}^{d_{\text{BERT}}}$.

The depression probability $p_t^{(i)}$ and its logit $\gamma_t^{(i)}$ are predicted by a Risk Classification network, based on $H_t^{(i)}$ and a φ-pooled semantic embedding $\varphi(S_t^{(i)})$. Specifically, the concatenation of $H_t^{(i)}$ and $\varphi(S_t^{(i)})$ is passed through a linear layer with layer normalization and ReLU activation, a dropout layer, and a final classification layer of the Risk Classification network. The outputs are a *score* $\hat{s}_t^{(i)}$ that indicates the level of depression

$$\begin{aligned} \hat{s}_t^{(i)} &= p_t^{(i)} - (1 - p_t^{(i)}) \\ &= 2p_t^{(i)} - 1, \end{aligned} \tag{13}$$

and a risk *decision* $\hat{y}_t^{(i)}$

$$\hat{y}_t^{(i)} = 1\{\gamma_t^{(i)} > 0\}, \tag{14}$$

where $1\{\cdot\}$ is an indicator function.

Besides the stepwise risk classification, we employ a score accumulation technique

[141] that accumulates the historical risk scores for the current *score* by

$$\tilde{s}_t^{(i)} = \sum_{t'=1}^{t} \hat{s}_{t'}^{(i)}, \tag{15}$$

and predict the risk *decision* by

$$\tilde{y}_t^{(i)} = 1\left\{\tilde{s}_t^{(i)} > \text{median}\left(\tilde{s}_{[1:t]}^{(i)}\right) + \gamma\text{MAD}\left(\tilde{s}_{[1:t]}^{(i)}\right)\right\}, \tag{16}$$

where $\tilde{s}_{[1:t]}^{(i)}$ is a list of the accumulated scores for user i up to time step t and median$(\cdot)$ renders the median value of a list. The MAD function is given by

$$\text{MAD}(\tilde{s}_t^{(i)}) = \text{median}\left(\left|\tilde{s}_{[1:t]}^{(i)} - \text{median}\left(\tilde{s}_{[1:t]}^{(i)}\right)\right|\right), \tag{17}$$

which evaluates the Median Absolute Deviation of the accumulated scores $\tilde{s}_{[1:t]}^{(i)}$.

3.3 Latency Penalty

We propose a latency penalty ψ that penalizes TAM for the latency of the *first-positive* predictions, in terms of the depression-positive users. The latency penalty for user i at time step t is given by

$$\begin{aligned}\psi\left(y^{(i)}, \gamma_t^{(i)}, \gamma_{\max(t)}^{(i)}, t; \alpha, o\right) = \\ \sigma(\gamma_t^{(i)}) \cdot y^{(i)} \cdot lc\left(t \cdot \tanh\left(\alpha \cdot \text{ReLU}\left(\gamma_t^{(i)}\right) \cdot \text{ReLU}\left(-\gamma_{\max(t)}^{(i)}\right)\right); o\right),\end{aligned} \tag{18}$$

where $y^{(i)} \in \{0,1\}$ is the ground truth label, $\gamma_t^{(i)} \in \mathbb{R}$ is the current predicted logit, $\gamma_{\max(t)}^{(i)} = \max_{t'=1}^{t-1} \gamma_{t'}^{(i)}$ is the maximum logit up to $t-1$, and $t \in \mathbb{Z}$ indicates the current time step. α and o are two hyper-parameters, respectively, which control the latency sensitivity and the time step at which the latency cost grows quickly as described below. σ is the sigmoid function. The latency cost function lc is first proposed in the ERDE metric [142], which is given by

$$lc(t;o) = 1 - \frac{1}{1+\exp^{t-o}}, \tag{19}$$

with input t denoting the *latency* step of a true-positive prediction. The latency cost $lc \in (0,1)$ monotonically grows with the *latency* step t and grows the most quickly at the step o with a latency cost of 0.5. In practice, o is usually set to 5 and 50, the latter of which is employed for training the proposed TAM network.

In Eq. 18, we obtain the *latency* of the *first-positive* prediction for user i through a series of neural activation functions of the sequence of logit predictions $\gamma^{(i)}_{[1:t]}$. Specifically, $\mathrm{ReLU}(\gamma^{(i)}_t)$ renders a positive value $\gamma^{(i)}_t$ if the logit with respect to the latest (t) posting is positive, and renders 0 otherwise. Similarly, $\mathrm{ReLU}(-\gamma^{(i)}_{\max(t)})$ renders a positive value $-\gamma^{(i)}_{\max(t)}$ if all logits up to the last ($t-1$) posting are negative, and renders 0 otherwise. We scale their product with the latency sensitivity $\alpha = 10000$ and feed the result to $\tanh(\cdot)$. The output turns to be an indicator that takes a value close to 1 if the model renders *a positive prediction* for the latest posting for user i and *all-negative predictions* before that, while takes the value of 0 otherwise. By multiplying the latest time step t with the indicator, we obtain the step of *first-positive* prediction, that is the *latency*, and feed it to the latency cost function in Eq. 19. The latency penalty ψ is finally given by the product of the depression probability $\sigma(\gamma^{(i)}_t)$, the ground-truth label $y^{(i)}$, and the latency cost lc.

We add the latency penalty in Eq. 18 to a cross-entropy loss to produce the final training target for Early Detection of Depression by

$$\begin{aligned}\ell(y,\gamma;\alpha,o) = \sum_{t=1}^{T}\sum_{i=1}^{N} \\ -\left(y^{(i)}\log\sigma(\gamma^{(i)}_t) + (1-y^{(i)})\log(1-\sigma(\gamma^{(i)}_t))\right) + \psi\left(y^{(i)},\gamma^{(i)}_t,\gamma^{(i)}_{\max(t)},t;\alpha,o\right),\end{aligned} \tag{20}$$

where N and T are the number of users and the number of time steps in the training data, respectively.

Tab 7. Number of positive and negative users in the training data of Early Detection of Depression at the CLEF 2022 Lab.

	Positive	Negative
2017	135	752
2018	79	741
Total	214	1493

3.4 Experiment

The training data of Early Detection of Depression at the CLEF eRisk 2022 Lab [118] consists of the training and test data of CLEF eRisk 2017 Lab and the test data of CLEF eRisk 2018 Lab. The details can be found in Table 7.

The test data of Early Detection of Depression at the CLEF eRisk 2022 Lab [118] consists of 1400 users. The posts of these users are accessible in an interactive manner during the test phase, that is, the server only replies one post per-user at step t after receiving the depression predictions for all users at step $t-1$. Posts at step 0 from all users are accessible at the very beginning.

We submit five groups of risk *decisions* and risk *scores* for 2000 steps in this interactive manner, which takes around 16.5 hours. Among all participants in Early Detection of Depression, our system turns to be *the most efficient*.

The distinctive configurations of the submitted models are shown in Table 8. Specifically, Balance Strategy indicates the way of selecting positive and negative users from the training data, for which *Balance* indicates that as many as the positive users are randomly selected from the negative set while *All* indicates that all users are utilized. $\varphi_{\text{DistilBERT}}$ and $\varphi_{\text{DistilBERT-Emo}}$ indicate a *Mean*-pooling or a *CLS*-pooling for $\varphi(\cdot)$ in Eq. 12 and Eq. 2, respectively. Max Memory Len corresponds to l_{MEM}, which is the length of *enriched* affective memory $\hat{M}$. Discount Function indicates the utilization of either g_{slog} or g_{flex} for discounting the short-term memory $\hat{C^S}$. Decoder Num specifies the number of Transformer Decoders in the TAM network for integrating the affective memory $\hat{M}$ and the semantic embedding S. Score Accumulation indicates predicting the depression scores

Tab 8. Distinctive configurations of the submitted models.

Configuration	TUA1#0	TUA1#1	TUA1#2	TUA1#3	TUA1#4
Balance Strategy	Balance	All	Balance	All	Balance
$\varphi_{\text{DistilBERT}}$	Mean	CLS	Mean	CLS	N/A
$\varphi_{\text{DistilBERT}-\text{Emo}}$	CLS	Mean	CLS	Mean	N/A
Max Memory Len	30	1	30	1	Full
Discount Function	g_{slog}	g_{flex}	g_{slog}	g_{flex}	N/A
Decoder Num	1	2	1	2	N/A
Score Accumulation	False	False	True	True	True

and risk decisions by either accumulating the historical risk scores or not. TUA1#0 to TUA1#3 corresponds to the TAM-based models with distinctive configurations, while TUA1#4 is a SS3-based model [141]. Configurations which are not applicable to the model are denoted as *N/A*. To avoid making reckless risk *decisions*, we halt the positive predictions by producing all-zero decisions in the first two time steps for all models.

Table 9 shows the decision-based evaluation results. First, we find that Score Accumulation in the TAM-based models obtains similar decision-based evaluation scores, which is possibly because that the TAM network has already maintained a long-term memory of the affective states through T-LSTM as well as an *enriched* affective memory. Next, TUA1#0 and TUA1#2 achieve better Precision, F1, ERDE_{50} and $\text{F}_{\text{latency}}$ scores than TUA1#1 and TUA1#3, which indicates a long affective memory and a balanced training data could be helpful for improving the decision predictions in TAM. Our results also suggest the importance of exploring language usage patterns for predicting the depression decisions. Last, it is reasonalbe to speculate that halting positive predictions for the first two time steps could be an important factor that reduces the latency-sensitive metric scores, such as ERDE_{5}, ERDE_{50}, $\text{latency}_{\text{TP}}$, and $\text{F}_{\text{latency}}$, in our result.

Table 10 shows the ranking-based evaluation results. First, the ranking-based decisions of TUA1#0 and TUA1#2 render the state-of-the-art results in P@10 and NDCG@10 based on only 1 user post. The result suggests that the TAM network with a long affective memory could effectively recognize the users' depression risk at a very early state. It also implies that taking the decision-halting strategy off from TAM might render better

Tab 9. Decision-based evaluation for the Early Detection of Depression task. Results obtained by our models and the best performing models on each metric are included.

Model	P	R	F_1	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	speed	$F_{latency}$
TUA1#0	0.155	0.806	0.260	0.055	0.037	3.0	0.922	0.258
TUA1#1	0.129	0.816	0.223	0.053	0.041	3.0	0.992	0.221
TUA1#2	0.155	0.806	0.260	0.055	0.037	3.0	0.992	0.258
TUA1#3	0.129	0.816	0.223	0.053	0.041	3.0	0.992	0.221
TUA1#4	0.159	0.959	0.272	0.052	0.036	3.0	0.992	0.270
CYUT#2	0.106	0.867	0.189	0.056	0.047	**1.0**	**1.000**	0.189
LauSAn#0	0.137	0.827	0.235	0.041	0.038	**1.0**	**1.000**	0.235
LauSAn#4	0.201	0.724	0.315	**0.039**	0.033	**1.0**	**1.000**	0.315
BLUE#2	0.106	**1.000**	0.192	0.074	0.048	4.0	0.988	0.190
NLPGroup-IISERB#0	0.682	0.745	**0.712**	0.055	0.032	9.0	0.969	**0.690**
Sunday-Rocker2#0	0.091	**1.000**	0.167	0.080	0.053	4.0	0.988	0.165
Sunday-Rocker2#4	0.108	**1.000**	0.195	0.082	0.047	6.0	0.981	0.191
SCIR2#3	0.316	0.847	0.460	0.079	**0.026**	44.0	0.834	0.383
E8-IJS#0	**0.684**	0.133	0.222	0.061	0.061	**1.0**	**1.000**	0.144

decision-based evaluation results. Next, TUA1#1 obtains better results than TUA1#3, which indicates that Score Accumulation might not be necessary for the ranking-based prediction in TAM. TUA1#0 and TUA1#2 generally obtain better P@10, NDCT@10, NDCG@100 scores for 1 post, 100 posts, 500 posts, and 1000 posts, which suggests that long affective memory and balanced data are also helpful in improving the ranking-based predictions for TAM. Last, the TAM-based models significantly outperform the SS3-based model in terms of the ranking-based metrics.

3.5 Conclusion

In this section, we propose a Time-Aware Affective Memories (TAM) network with a latency-penalized cross-entropy loss for Early Detection of Depression at the CLEF eRisk 2022 Lab. Both decision- and ranking-based evaluation results indicate that affective state is an important indicator of depression and that a long affective memory is crutial for TAM to explore the users' affective states. Our initial experiment suggests that adding a latency penalty to the cross-entropy loss is effective for training early detection models. Among all participants, our system turns to be the most efficient and achieves

Tab 10. Ranking-based evaluation for the Early Detection of Depression task. Results obtained by our models and the best performing models on each metric are included.

Model	1 post			100 posts			500 posts			1000 posts		
	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100
TUA1#0	**0.80**	**0.88**	**0.44**	**0.60**	**0.72**	**0.52**	**0.60**	**0.67**	**0.52**	**0.70**	**0.80**	**0.57**
TUA1#1	0.70	0.77	**0.44**	0.50	0.54	0.39	0.50	0.56	0.42	0.50	0.65	0.43
TUA1#2	**0.80**	**0.88**	**0.44**	**0.60**	**0.72**	**0.52**	**0.60**	**0.67**	**0.52**	**0.70**	**0.80**	**0.57**
TUA1#3	0.60	0.69	0.43	0.50	0.54	0.39	0.50	0.56	0.42	0.50	0.65	0.43
TUA1#4	0.50	0.37	0.35	0.00	0.00	0.36	0.00	0.00	0.36	0.20	0.12	0.31

two state-of-the-art results in terms of the ranking-based evaluation. Our results also suggest that language usage patterns, such as n-grams, could be an important feature for depression detection. Integrating language usage patterns into the TAM network could be a promising work in the future.

4 TAM-SenticNet: A Neuro-Symbolic AI Approach for Early Depression Detection via Social Media Analysis

4.1 Introduction

In light of recent advancements in artificial intelligence, Neuro-Symbolic AI has emerged as a pioneering approach. We presents TAM-SenticNet, a specialized Neuro-Symbolic AI framework, explicitly designed for early depression detection. The framework integrates neural networks, proficient in sentiment analysis, with symbolic reasoning, known for its ability to perform nuanced logical inference. This amalgamation enables TAM-SenticNet to provide a more comprehensive and accurate interpretation of emotional indicators present in user-generated content on social media platforms.

The primary aim of this study is to rigorously assess the efficacy of TAM-SenticNet in the realm of early depression detection through the scrutiny of social media data.

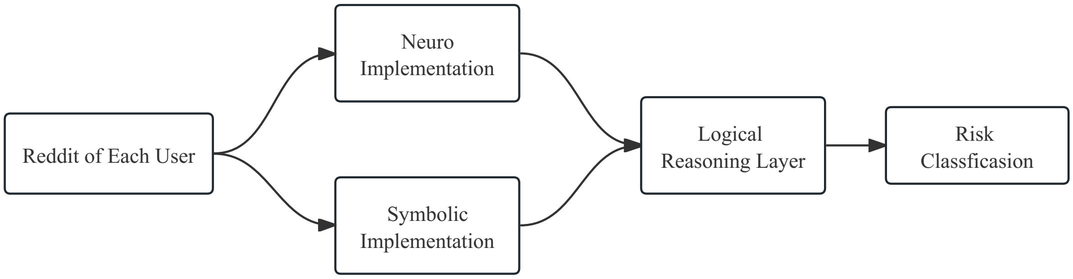

Fig 8. Schematic overview of the TAM-SenticNet framework, illustrating the seamless integration of Neural and Symbolic Implementation modules for early depression detection.

The paper will outline the architecture of this Neuro-Symbolic AI framework, clarify its specific applications in sentiment analysis, and empirically substantiate its performance by conducting a comparative evaluation against existing models. This will highlight its unique advantages and potential contributions to the field of mental health research.

4.2 Neuro-Symbolic AI for Early Depression Detection

To investigate the temporal dynamics of user affective states for the early identification of depression, we present TAM-SenticNet, a specialized Neuro-Symbolic AI framework, as depicted in Fig. 8. This framework comprises two principal modules: the Neural Implementation and the Symbolic Implementation.

4.2.1 Neural Implementation

Drawing inspiration from Time-Aware Affective Memories (TAM) network [14], the neural component of our framework employs a Time-Aware Long Short-Term Memory (T-LSTM) model. This model,It has been specifically explained in the previous chapter, and there is not much to say here.

4.2.2 Symbolic Implementation

To mitigate the interpretability challenges commonly associated with neural networks, we integrate SenticNet [143] as our symbolic reasoning engine. SenticNet func-

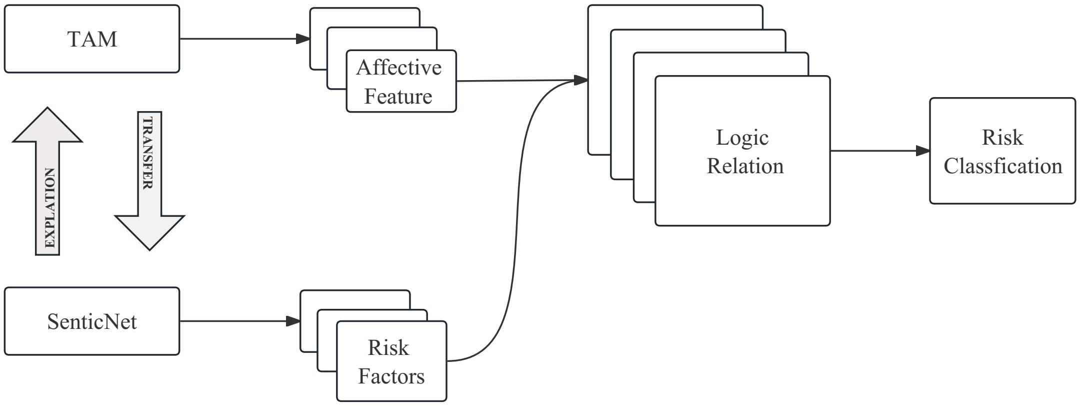

Fig 9. Symbolic Implementation utilizing SenticNet as the reasoning engine, emphasizing how symbolic logic and semantic relations contribute to enhanced model interpretability.

tions as a sentiment knowledge repository, utilizing symbolic logic and semantic relations to furnish a structured emotional understanding, thereby augmenting the framework's interpretability. The architecture and functionality of this symbolic reasoning engine are visually depicted in Fig. 9, emphasizing its role in enhancing model interpretability.

Our framework synergistically fuses the outputs from the neural network with symbolic knowledge to deliver higher-order explanations and predictions. The Risk Factors identified by SenticNet are graphically represented in Fig. 10, which also delineates the logical interconnections among these factors. For example, an individual manifesting TakeSleepingPills, Irritability, and AttemptSuicide is highly likely to be at elevated risk for depression.

In our framework, we exploit both the neural network capabilities of TAM and the symbolic reasoning of SenticNet to evaluate a user's depression risk based on linguistic features. Figure 11 elucidates the logical nexus between user language patterns and depression risk factors, thereby highlighting the synergistic interplay between SenticNet and TAM.

This logical diagram serves to illuminate the collaborative strength between SenticNet and TAM. Through this integrated methodology, we aspire to identify early indicators of depression with enhanced accuracy and efficiency, thereby enabling more timely pro-

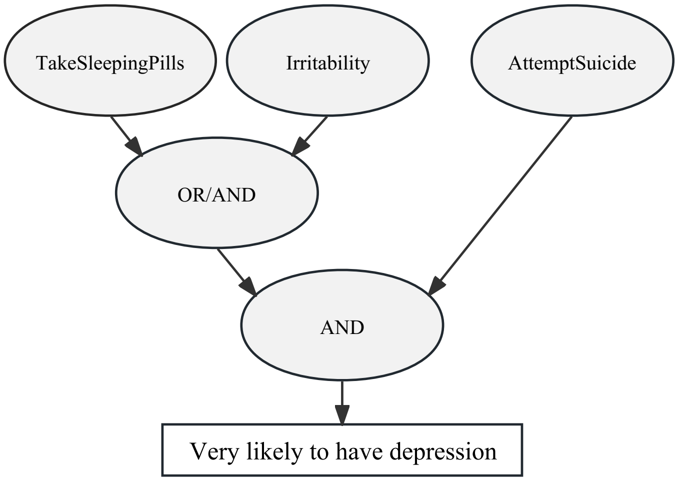

Fig 10. Visual representation of Risk Factors as delineated by SenticNet, illustrating the logical interconnections among TakeSleepingPills, Irritability, and AttemptSuicide.

fessional interventions and improving mental health outcomes.

5 Experiment and Results

5.1 Experimental Data

The dataset employed in this research for early depression identification is sourced from the CLEF eRisk 2022 Lab. This dataset amalgamates both the training and testing sets from the CLEF eRisk 2017 Lab, in addition to the testing set from the CLEF eRisk 2018 Lab. A comprehensive overview of the dataset's attributes is presented in Table 11.

Tab 11. Positive and negative samples in CLEF eRisk 2017 and 2018 Labs.

Year	Positive Samples	Negative Samples
2017	135	752
2018	79	741
Aggregate	214	1493

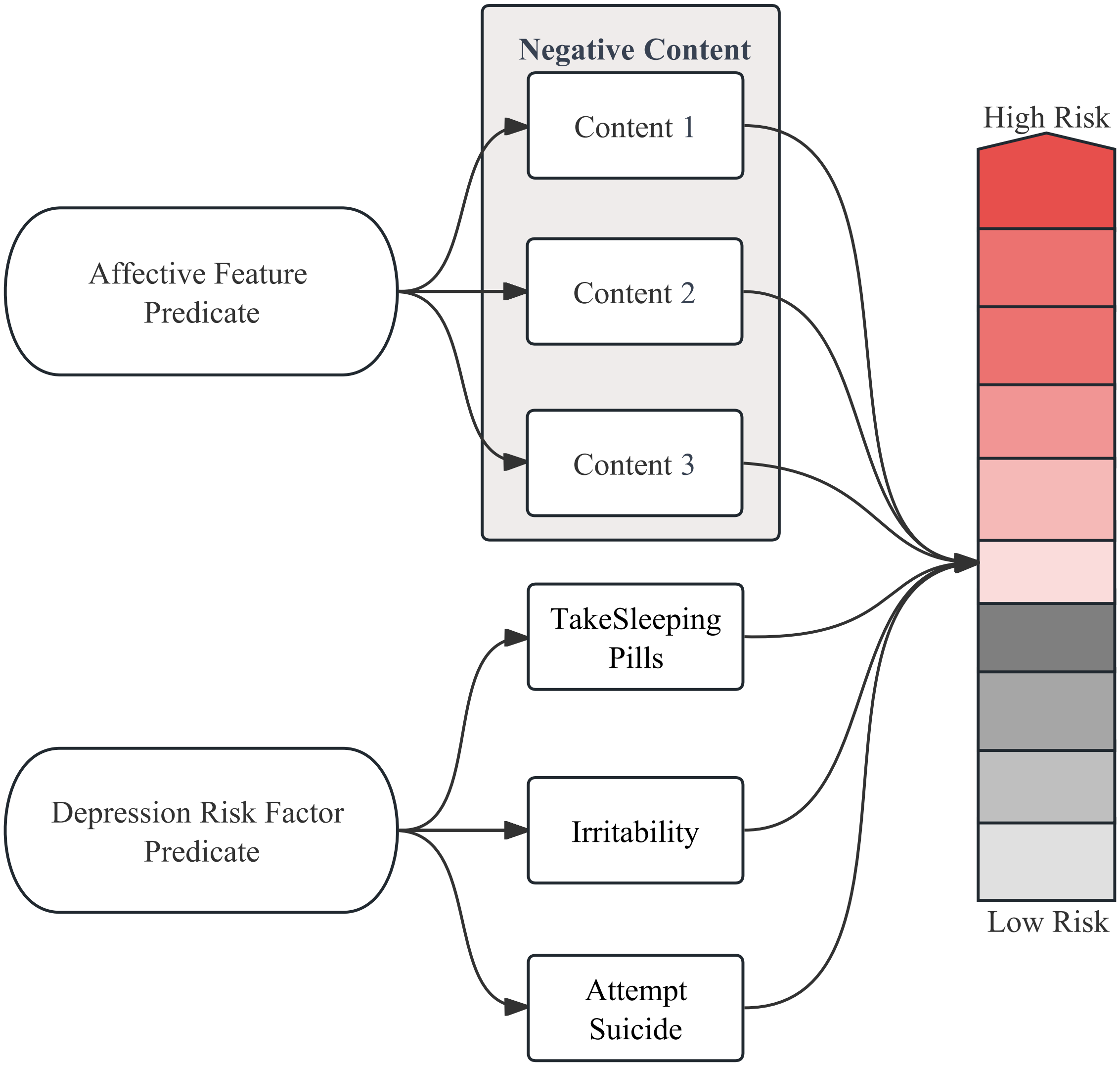

Fig 11. Logical Relationship diagram illustrating the collaborative efficacy between SenticNet and TAM in assessing depression risk based on user language and emotional states.

5.2 Evaluation Metrics

5.2.1 Classification Metrics: Precision, Recall, and F_1 Score

In our early depression identification framework, we utilize essential classification metrics for a balanced evaluation: Precision, Recall, and the F_1 Score. Precision, defined as $\text{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$, measures the model's accuracy in identifying true depression instances, with TP and FP denoting True Positives and False Positives, respectively. Recall, calculated as $\text{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$, assesses the model's ability to capture genuine depression cases, where FN represents False Negatives. The F_1 Score, given by $\mathrm{F}_1 = \frac{2\times(\text{Precision}\times\text{Recall})}{\text{Precision}+\text{Recall}}$, provides a harmonic mean of Precision and Recall, offering a comprehensive metric that considers both types of classification errors.

5.2.2 Latency Metrics: ERDE, Latency$_{TP}$, and $\mathrm{F}_{\text{latency}}$

Timeliness constitutes another pivotal dimension in the domain of early depression identification. To address this, we deploy a suite of latency metrics—ERDE, Latency$_{TP}$, and $\mathrm{F}_{latency}$—to assess the model's efficacy and efficiency in real-time decision-making.

The ERDE metric amalgamates both the efficacy and timeliness of a decision by accounting for the relative costs of false negatives and false positives. It is mathematically formulated as

$$\mathrm{ERDE}_o = \frac{1}{N}\sum_{i=1}^{N}\left(c_{\mathrm{FN}} \times \mathrm{FN}_i \times \phi(o,k_i) + c_{\mathrm{FP}} \times \mathrm{FP}_i\right), \tag{21}$$

where N denotes the total number of instances, and c_{FN} and c_{FP} represent the costs of false negatives and false positives, respectively.

Latency$_{TP}$ quantifies the median quantity of textual items required to accurately discern a true positive instance. It is mathematically expressed as

$$\text{Latency}_{TP} = \text{median}\{k_u : u \in \mathrm{U}, d_u = g_u = 1\}, \tag{22}$$

where U signifies the set of users and k_u indicates the number of textual items for user u

[115].

$\mathrm{F}_{latency}$ integrates the F_1 Score with a latency penalty term, striving to harmonize the model's accuracy and timeliness in the context of early depression detection. It is mathematically defined as

$$\mathrm{F}_{latency}(\mathrm{U},\mathrm{sys}) = \mathrm{F}_1(\mathrm{U},\mathrm{sys}) \times \left(1 - \mathrm{median}_{u\in\mathrm{U}\wedge\mathrm{ref}(u)=+}\mathrm{P}_{latency}(u,\mathrm{sys})\right), \tag{23}$$

where U denotes the set of users, sys refers to the depression detection system under evaluation and $\mathrm{P}_{latency}$ represents the proportion of true positives that are identified within the latency period [144].

5.3 Experimental Results and Discussion

5.3.1 Baseline Models

We compare TAM-SenticNet's performance with various baseline models, each characterized by their core approach – neural, symbolic, or a combination of both. These models, documented in the eRisk survey paper [145], include:

Neural Models: **CYUT (CY)**, **BLUE (BL)**, and **SCIR2 (SC)** employ advanced deep learning architectures for user-level classification. **LauSAn (LS)** and **NITK-NLP2 (NK)** utilize dynamic neural network analysis.

Symbolic Models: **BioInfo_UAVR (BU)** and **NLPGroup-IISERB (NI)** use classical machine learning with feature engineering. **E8-IJS (E8)** focuses on Logistic Regression models with varied input representations.

Hybrid Models: **TUA1 (T1)** and **UNSL (UN)** integrate neural techniques with feature-centric symbolic approaches. **RELAI (RL)** combines pre-trained word vectors with feature sets for automatic questionnaire population. **UNED-MED (UM)** incorporates tf-idf and sentiment analysis with a Deep Learning classifier. **Sunday-Rocker2 (SR)** adopts a multifaceted approach with tf-idf, linguistic features, and machine learning algorithms.

Tab 12. Comparative metrics of our **TAM-SenticNet (TS)** and other models. Model types are indicated as follows: † for Neural Models, ‡ for Symbolic Models, and § for Hybrid Models.

Model	Prc	Rcl	F_1	E_5	E_{50}	L_{TP}	F_l
CY†	0.142	0.918	0.245	0.082	0.041	8.0	0.239
BL†	0.106	**1.000**	0.192	0.074	0.048	4.0	0.190
SC†	0.274	0.847	0.460	0.045	0.031	3.0	0.411
LS†	0.201	0.724	0.315	0.039	**0.025**	**1.0**	0.315
NK†	0.149	0.724	0.248	0.049	0.039	2.0	0.247
BU‡	0.378	0.857	0.525	0.069	0.031	16.0	0.494
NI‡	0.653	0.500	0.566	0.067	0.046	26.0	0.511
E8‡	0.242	0.959	0.387	0.068	0.036	20.5	0.357
T1§	0.159	0.959	0.271	0.052	0.036	3.0	0.270
UN§	0.144	0.929	0.249	0.055	0.035	3.0	0.247
RL§	0.085	0.847	0.155	0.114	0.092	51.0	0.125
UM§	0.084	0.163	0.111	0.079	0.078	251.0	0.028
SR§	0.108	**1.000**	0.195	0.082	0.047	6.0	0.191
TS	**0.665**	0.881	**0.758**	**0.035**	**0.025**	**1.0**	**0.675**

5.3.2 Results and Discussion

In our study, TAM-SenticNet (TS) exhibits high Precision (0.665) and F_1 (0.758), demonstrating its effectiveness in accurately and comprehensively identifying depression cases as detailed in Table 12. This performance is particularly notable when compared with neural models like CY, BL, and SC, which primarily focus on deep learning techniques. Unlike these models, TAM-SenticNet integrates symbolic reasoning, enabling more nuanced data interpretation.

Symbolic models like BU, NI, and E8 rely heavily on feature engineering. TAM-SenticNet surpasses these models in Precision and F_1, highlighting the advantage of combining symbolic reasoning with neural network robustness. This combination allows TAM-SenticNet to capture complex patterns in data without solely relying on feature engineering.

Hybrid models such as T1, UN, RL, UM, and SR attempt to balance neural and symbolic approaches. However, TAM-SenticNet's Neuro-Symbolic AI approach further optimizes this balance, as evidenced by its superior F_1 score (0.758), which is indicative of a well-rounded performance in both Precision and Recall aspects.

In latency metrics, TAM-SenticNet excels with the lowest ERDE_5 (0.035) and ERDE_{50} (0.025) scores, demonstrating its efficiency in early risk detection—a crucial aspect of mental health applications. The $\mathrm{Latency}_{TP}$ score (1.0) and the highest $\mathrm{F}_{latency}$ score (0.675) further establish TAM-SenticNet as a model that proficiently balances timely and accurate detection. These metrics are especially important in early depression detection, where prompt intervention can significantly alter outcomes.

These findings strongly advocate for the integration of neural networks and symbolic reasoning in TAM-SenticNet, marking a significant stride in mental health informatics. Its capability to accurately, promptly, and efficiently detect early signs of depression, leveraging the strengths of both neural and symbolic methodologies, positions it as a highly promising tool in this critical field of research. The comprehensive evaluation against various model types underscores TAM-SenticNet's potential to set a new benchmark in early depression detection.

5.3.3 Case Study: Symbolic Reasoning in Action

To further illuminate the capabilities of TAM-SenticNet in real-world applications, we present a case study that exemplifies how symbolic reasoning functions within the model. Figure 12 delineates this by monitoring the fluctuating risk levels of depression for a Reddit user based on their posts.

Initially, the user posted, "When I was 24 I began taking sleeping pills." The symbolic reasoning component of TAM-SenticNet ascertained the risk factor "TakeSleepingPills," leading to a medium risk level of depression. Subsequently, the user articulated, "I never considered myself an addict but now I realize how angry and irritable I am when I don’t have any sleeping pills." Here, the symbolic risk factor "Irritability" was identified, escalating the risk level to high. Finally, the user divulged, "I’ve been through this before (with a suicide attempt) taking sleeping pills and gashing my wrist wide open." The model discerned the symbolic risk factor "AttemptSuicide," culminating in a very

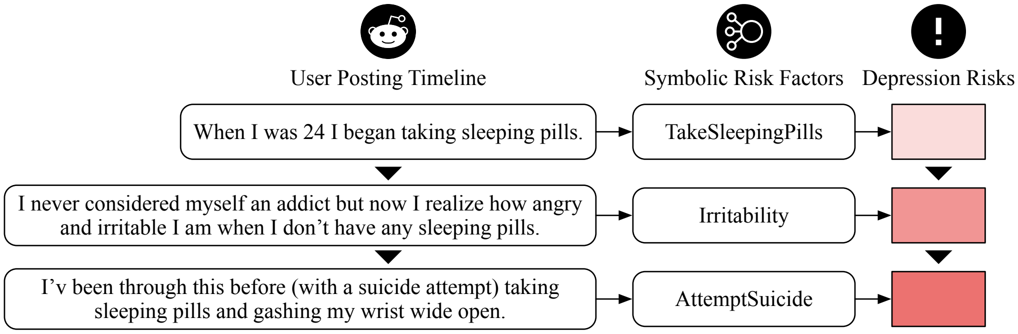
User Posting Timeline
Symbolic Risk Factors
Depression Risks
When I was 24 I began taking sleeping pills.
TakeSleepingPills
I never considered myself an addict but now I realize how angry and irritable I am when I don't have any sleeping pills.
Irritability
I'v been through this before (with a suicide attempt) taking sleeping pills and gashing my wrist wide open.
AttemptSuicide

Fig 12. TAM-SenticNet's symbolic reasoning applied to a Reddit user's posts. The model discerns risk factors—"TakeSleepingPills," "Irritability," and "AttemptSuicide"—to evaluate evolving depression risk levels from medium to very high.

high risk of depression. These risk assessments are derived from the logical relations delineated in the symbolic implementation of TAM-SenticNet, highlighting the model's capacity for nuanced and precise estimations.

For additional examples demonstrating TAM-SenticNet's application in a wider range of scenarios, please refer to the case studies provided in the appendix. These cases further exemplify the model's versatility and effectiveness in analyzing various emotional and linguistic patterns indicative of depression risk.

6 Conclusion and Future Work

6.1 Conclusion

This study introduces TAM-SenticNet, a groundbreaking Neuro-Symbolic Artificial Intelligence framework, as an efficacious instrument for the early detection of depression. The model excels in seamlessly amalgamating neural networks for feature extraction and sentiment analysis with symbolic reasoning for intricate logical inference. Empirical assessments corroborate the model's superlative performance across an expansive array of evaluation metrics, encompassing not only conventional metrics like Precision, Recall, and F_1 score but also specialized metrics such as $ERDE_5$, $ERDE_{50}$, $Latency_{TP}$, and $F_{latency}$. These findings cumulatively position TAM-SenticNet as a persuasive contender for implementation in mental health informatics. In future work, our research agenda endeavors to further refine the symbiotic relationship between neural networks and symbolic reasoning to construct more robust and interpretable models, an essential requirement in healthcare applications. Additionally, we aim to extend the model's versatility by enhancing the integration of data and domain-specific knowledge across diverse fields.

6.2 Future Work

Our future work will concentrate on augmenting TAM-SenticNet with medical expert knowledge from internationally recognized depression diagnosis manuals such as DSM-5-TR, ICD-11, and PHQ-9. This enhancement is envisioned to bolster the framework's clinical relevance and trustworthiness by aligning its analytical capabilities with established diagnostic criteria and best practices in mental health care. The integration of this expert knowledge is anticipated to not only refine the model's interpretability and accuracy but also to ensure that its output aligns with clinical insights, thus bridging the gap between AI-driven analysis and real-world clinical applications in depression detection and early intervention. Through these efforts, we aim to create a Neuro-Symbolic AI system that is not only technologically advanced but also deeply rooted in medical expertise, making it a valuable asset in the field of mental health.

Acknowledgement

I would like to express my heartfelt gratitude to Prof. Matsumoto, Prof. Kang, and the examination committee, especially Prof. Shishibori, Prof. Fuketa, and Prof. Nagata, for their invaluable guidance and support throughout the completion of my thesis. The lecturer's profound professional knowledge, rigorous academic attitude, excellent work style, tireless noble teacher style, strict self-discipline, tolerant noble behavior, simple, difficult to approach and approachable personality charm have had a profound impact on me. I not only set ambitious learning goals and mastered research methods, but also made me understand a lot of truths about dealing with people. From the selection of the topic to the completion of the thesis, every step is completed under the careful guidance of the tutor, and a lot of energy has been paid. Here, I would like to express my high respect and heartfelt thanks to the teacher! In the process of writing the paper, I met many problems, and under the teacher's patient guidance, the problems were solved. So here, I say to the teacher again, thank you!

Time flows swiftly, and in the blink of an eye, graduation is upon us. I want to convey my sincere gratitude to all the teachers at Tokushima University. Thank you for your dedicated efforts over the past four years, for imparting life principles, and for your unwavering commitment.

In the past four years, what I have gained is not only richer knowledge, but more importantly, the way of thinking, expression ability and broad vision cultivated in reading and practice. I have been fortunate to have met many mentors and helpful friends over the past four years. Whether it is study, life or work, they have given me selfless help and warm care, so that I can spend four years of academic life in a warm environment. It is difficult to measure gratitude in words. I would like to express my great respect in the simplest words.

A Symbolic Reasoning Examples

Table 13 presents a series of case studies extracted from the testing set of the CLEF eRisk 2022 Lab dataset. These studies demonstrate the TAM-SenticNet framework's application in analyzing real-world social media content, particularly Reddit posts, to assess depression risk.

In the table, each entry includes the original Reddit post, with crucial negative expressions highlighted in bold. The "Reasoning Steps" column delineates TAM-SenticNet's logical process for determining depression risk from these posts. This column vividly illustrates the framework's ability to intricately analyze emotional and linguistic patterns. The risk level predicted by TAM-SenticNet is denoted in parentheses, on a scale from 1 to 10, where higher values indicate increased risk severity. This compilation of case studies underscores TAM-SenticNet's practical utility and effectiveness in detecting early signs of depression through the detailed examination of social media content.

Tab 13. Case Studies of Symbolic Reasoning in TAM-SenticNet

Reddit Post	**Reasoning Steps**
It’s 1:11am on the West Coast. I have no alcohol to calm me down.I always **feel unloved**. Every single day I **get closer to suicide**. “**I want to die**” is my first thought of the morning or after I nap. I have **no hope or will to live**... I have mental breakdowns every single day.	“feel unloved” → SocialIsolation; “I want to die” → AttemptSuicide; “get closer to suicide” → AttemptSuicide; “no hope or will to live” → ProlongedSadness; SocialIsolation ∧ AttemptSuicide ∧ ProlongedSadness → Depressed (6)
I love them more than the moon and the stars and would give my life for them in a heartbeat, but it has caused me tremendous heartache more than anything. I raised them as best as I could and I gave them my all. I feel like parenting just **beat the living shit out of me**. I have nothing left to give anybody I wish I had more to give, I’m emotionally and mentally bankrupt from it.	“beat the living shit out of me” → ProlongedSadness ∧ AnxietyAndWorry; ProlongedSadness ∧ AnxietyAndWorry → Depressed (4)
Hey, I’ ve been on **Sertraline** for nearly two years I had to change because the fluoxetine(Prozac) was causing worse anxiety.	“Sertraline” → SubstanceAbuse; SubstanceAbuse → Depressed (1)
I felt like I was putting in all the effort and that it was really me driving the relationship. I always knew he loved me, but I still **felt very lonely** in the relationship as it currently stood.	“felt very lonely” → AnxietyAndWorry; AnxietyAndWorry → Depressed (2)
I have **insomnia and anxiety** as well and may even have **hypomania**. A few weeks ago I was optimistic... And then I crashed and have been in what feels like a **never-ending depressive** episode since June.	“insomnia and anxiety” → SuddenBehavioralChange; “never-ending depressive” → ProlongedSadness; “hypomania” → ErraticBehavior; SuddenBehavioralChange ∧ ProlongedSadness ∧ ErraticBehavior → Depressed (5)
My mother has a bad knee which has prevented her from doing some activities, and I have **always worried** that the same thing would happen to me since it sometimes feels like my knee caps are "out of place"	“always worried” → AnxietyAndWorry ∧ ProlongedSadness; AnxietyAndWorry ∧ ProlongedSadness → Depressed (3)

References

[1] Ronald C Kessler, Wai Tat Chiu, Olga Demler, and Ellen E Walters. Prevalence, severity, and comorbidity of 12-month dsm-iv disorders in the national comorbidity survey replication. *Archives of general psychiatry*, 62(6):617–627, 2005.

[2] Evelyn Bromet, Laura Helena Andrade, Irving Hwang, Nancy A Sampson, Jordi Alonso, Giovanni De Girolamo, Ron De Graaf, Koen Demyttenaere, Chiyi Hu, Noboru Iwata, et al. Cross-national epidemiology of dsm-iv major depressive episode. *BMC medicine*, 9(1):1–16, 2011.

[3] Ronald C Kessler, Maria Petukhova, Nancy A Sampson, Alan M Zaslavsky, and Hans-Ullrich Wittchen. Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the united states. *International journal of methods in psychiatric research*, 21(3):169–184, 2012.

[4] Fiona C Bull, Salih S Al-Ansari, Stuart Biddle, Katja Borodulin, Matthew P Buman, Greet Cardon, Catherine Carty, Jean-Philippe Chaput, Sebastien Chastin, Roger Chou, et al. World health organization 2020 guidelines on physical activity and sedentary behaviour. *British Journal of Sports Medicine*, 54(24):1451–1462, 2020.

[5] Hassan Zaidi, Mohamed Bader-El-Den, and James McNicholas. Using the national early warning score (news/news 2) in different intensive care units (icus) to predict the discharge location of patients. *BMC Public Health*, 19(1):1–9, 2019.

[6] Yubo Hou, Dan Xiong, Tonglin Jiang, Lily Song, and Qi Wang. Social media addiction: Its impact, mediation, and intervention. *Cyberpsychology: Journal of psychosocial research on cyberspace*, 13(1), 2019.

[7] Daria J Kuss and Mark D Griffiths. Social networking sites and addiction: Ten lessons learned. *International journal of environmental research and public health*, 14(3):311, 2017.

[8] E Donnelly and DJ Kuss. Depression among users of social networking sites (snss): The role of sns addiction and increased usage. *Journal of Addiction and Preventive Medicine*, 1(2):107, 2016.

[9] Lin Lin, Xuri Chen, Ying Shen, and Lin Zhang. Towards automatic depression detection: A bilstm/1d cnn-based model. *Applied Sciences*, 10(23):8701, 2020.

[10] Betul Ay, Ozal Yildirim, Muhammed Talo, Ulas Baran Baloglu, Galip Aydin, Subha D Puthankattil, and U Rajendra Acharya. Automated depression detection using deep representation and sequence learning with eeg signals. *Journal of medical systems*, 43:1–12, 2019.

[11] Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893, 2019.

[12] Fidel Cacheda, Diego Fernandez, Francisco J Novoa, and Victor Carneiro. Early detection of depression: social network analysis and random forest techniques. *Journal of medical Internet research*, 21(6):e12554, 2019.

[13] Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*, 32(3):588–601, 2018.

[14] Xin Kang, Rongyu Dou, and Haitao Yu. Tua1 at erisk 2022: Exploring affective memories for early detection of depression. In *Proceedings of CLEF (Working Notes)*, pages 1–12, 2022.

[15] Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135, 2008.

[16] Sheng-Tun Li and Fu-Ching Tsai. A fuzzy conceptualization model for text mining with application in opinion polarity classification. *Knowledge-Based Systems*, 39:23–33, 2013.

[17] María-Teresa Martín-Valdivia, Eugenio Martínez-Cámara, Jose-M Perea-Ortega, and L Alfonso Ureña-López. Sentiment polarity detection in spanish reviews combining supervised and unsupervised approaches. *Expert Systems with Applications*, 40(10):3934–3942, 2013.

[18] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[19] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, 2012.

[20] Randolph R Cornelius. *The science of emotion: Research and tradition in the psychology of emotions.* Prentice-Hall, Inc, 1996.

[21] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the international AAAI conference on web and social media*, volume 5, pages 450–453, 2011.

[22] Robert Plutchik. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier, 1980.

[23] Chu Wang, Daling Wang, Shi Feng, and Yifei Zhang. An approach of fuzzy relation equation and fuzzy-rough set for multi-label emotion intensity analysis. In *Database Systems for Advanced Applications: DASFAA 2016 International Workshops: BDMS, BDQM, MoI, and SeCoP, Dallas, TX, USA, April 16-19, 2016, Proceedings 21*, pages 65–80. Springer, 2016.

[24] Bing Liu et al. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666, 2010.

[25] Lei Li, Yabin Wu, Yuwei Zhang, and Tianyuan Zhao. Time+ user dual attention based sentiment prediction for multiple social network texts with time series. *IEEE Access*, 7:17644–17653, 2019.

[26] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418, 2011.

[27] Stephan Raaijmakers and Wessel Kraaij. A shallow approach to subjectivity classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 2, pages 216–217, 2008.

[28] Andrés Montoyo, Patricio Martínez-Barco, and Alexandra Balahur. Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. *Decision Support Systems*, 53(4):675–679, 2012.

[29] Philip Beineke, Trevor Hastie, Christopher Manning, and Shivakumar Vaithyanathan. Exploring sentiment summarization. In *Proceedings of the AAAI spring symposium on exploring attitude and affect in text: theories and applications*, volume 39. The AAAI Press Palo Alto CA, 2004.

[30] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058*, 2004.

[31] Dingding Wang, Shenghuo Zhu, and Tao Li. Sumview: A web-based engine for summarizing product reviews and customer opinions. *Expert Systems with Applications*, 40(1):27–33, 2013.

[32] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004.

[33] G Jalaja and C Kavitha. Sentiment analysis for text extracted from twitter. *Integrated Intelligent Computing, Communication and Security*, pages 693–700, 2019.

[34] Ben He, Craig Macdonald, Jiyin He, and Iadh Ounis. An effective statistical approach to blog post opinion retrieval. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 1063–1072, 2008.

[35] Liqiang Guo and Xiaojun Wan. Exploiting syntactic and semantic relationships between terms for opinion retrieval. *Journal of the american society for information science and technology*, 63(11):2269–2282, 2012.

[36] Bing Liu. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press, 2020.

[37] Yohei Seki, Noriko Kando, and Masaki Aono. Multilingual opinion holder identification using author and authority viewpoints. *Information Processing & Management*, 45(2):189–199, 2009.

[38] Bishan Yang and Claire Cardie. Joint inference for fine-grained opinion extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1640–1649, 2013.

[39] Elena Filatova. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Lrec*, pages 392–398, 2012.

[40] Antonio Reyes, Paolo Rosso, and Davide Buscaldi. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12, 2012.

[41] Antonio Reyes and Paolo Rosso. Making objective decisions from subjective data: Detecting irony in customer reviews. *Decision support systems*, 53(4):754–760, 2012.

[42] Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World wide web*, pages 751–760, 2010.

[43] Danushka Bollegala, Tingting Mu, and John Yannis Goulermas. Cross-domain sentiment classification using sentiment sensitive embeddings. *IEEE Transactions on Knowledge and Data Engineering*, 28(2):398–410, 2015.

[44] Tareq Al-Moslmi, Nazlia Omar, Salwani Abdullah, and Mohammed Albared. Approaches to cross-domain sentiment analysis: A systematic literature review. *Ieee access*, 5:16173–16192, 2017.

[45] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14, 2017.

[46] Nan Xu and Wenji Mao. A residual merged neutral network for multimodal sentiment analysis. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, pages 6–10. IEEE, 2017.

[47] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2539–2544, 2015.

[48] Hira Ahmed, Saman Hina, and Raheela Asif. Evaluation of descriptive answers of open ended questions using nlp techniques. In *2021 4th International Conference on Computing & Information Sciences (ICCIS)*, pages 1–7. IEEE, 2021.

[49] A Haripriya, Santoshi Kumari, and C Narendra Babu. Location based real-time sentiment analysis of top trending event using hybrid approach. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1052–1057. IEEE, 2018.

[50] Arjun Mukherjee, Bing Liu, and Natalie Glance. Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st international conference on World Wide Web*, pages 191–200, 2012.

[51] David Zimbra, Ahmed Abbasi, Daniel Zeng, and Hsinchun Chen. The state-of-the-art in twitter sentiment analysis: A review and benchmark evaluation. *ACM Transactions on Management Information Systems (TMIS)*, 9(2):1–29, 2018.

[52] Shiliang Sun, Chen Luo, and Junyu Chen. A review of natural language processing techniques for opinion mining systems. *Information fusion*, 36:10–25, 2017.

[53] Munir Ahmad, Shabib Aftab, Muhammad Salman Bashir, and Noureen Hameed. Sentiment analysis using svm: a systematic literature review. *International Journal of Advanced Computer Science and Applications*, 9(2), 2018.

[54] Lei Zhang, Shuai Wang, and Bing Liu. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253, 2018.

[55] Barbara Calabrese and Mario Cannataro. Sentiment analysis and affective computing: Methods and applications. In *Brain-Inspired Computing: Second International Workshop, BrainComp 2015, Cetraro, Italy, July 6-10, 2015, Revised Selected Papers 2*, pages 169–178. Springer, 2016.

[56] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.

[57] Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani, et al. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204, 2010.

[58] Hassan Saif, Miriam Fernandez, Leon Kastler, and Harith Alani. Sentiment lexicon adaptation with context and semantics for the social web. *Semantic Web*, 8(5):643–665, 2017.

[59] Farhan Hassan Khan, Usman Qamar, and Saba Bashir. A semi-supervised approach to sentiment analysis using revised sentiment strength based on sentiwordnet. *Knowledge and information Systems*, 51:851–872, 2017.

[60] Chihli Hung. Word of mouth quality classification based on contextual sentiment lexicons. *Information Processing & Management*, 53(4):751–763, 2017.

[61] Shi Feng, Kaisong Song, Daling Wang, and Ge Yu. A word-emoticon mutual reinforcement ranking model for building sentiment lexicon from massive collection of microblogs. *World Wide Web*, 18:949–967, 2015.

[62] Jaspreet Singh, Gurvinder Singh, and Rajinder Singh. Optimization of sentiment analysis using machine learning classifiers. *Human-centric Computing and information Sciences*, 7:1–12, 2017.

[63] Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 151–161, 2011.

[64] Meng Joo Er, Fan Liu, Ning Wang, Yong Zhang, and Mahardhika Pratama. User-level twitter sentiment analysis with a hybrid approach. In *Advances in Neural Networks–ISNN 2016: 13th International Symposium on Neural Networks, ISNN 2016, St. Petersburg, Russia, July 6-8, 2016, Proceedings 13*, pages 426–433. Springer, 2016.

[65] Youngseok Choi and Habin Lee. Data properties and the performance of sentiment classification for electronic commerce applications. *Information Systems Frontiers*, 19:993–1012, 2017.

[66] Ammar Hassan, Ahmed Abbasi, and Daniel Zeng. Twitter sentiment analysis: A bootstrap ensemble framework. In *2013 international conference on social computing*, pages 357–364. IEEE, 2013.

[67] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[68] Arman Cohan, Sydney Young, Andrew Yates, and Nazli Goharian. Triaging content severity in online mental health forums. *Journal of the Association for Information Science and Technology*, 2017.

[69] Altug Akay, Andrei Dragomir, and Bjorn Erik Erlandsson. Network-based modeling and intelligent data mining of social media for improving care. *IEEE Journal of Biomedical Health Informatics*, 19(1):210, 2015.

[70] Andrew Yates, Arman Cohan, and Nazli Goharian. Depression and self-harm risk assessment in online forums. 2017.

[71] David E. Losada, Fabio Crestani, and Javier Parapar. Overview of erisk: Early risk prediction on the internet. *Springer, Cham*, 2018.

[72] JavierParapar, FabioCrestani, and Davide. Losada. Overview of erisk 2019 early risk prediction on the internet. 2019.

[73] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Workshop on Computational Linguistics Clinical Psychology: from Linguistic Signal to Clinical Reality*, 2015.

[74] Tiancheng Shen, Jia Jia, Guangyao Shen, Fuli Feng, and Wendy Hall. Cross-domain depression detection via harvesting social media. In *Twenty-Seventh International Joint Conference on Artificial Intelligence IJCAI-18*, 2018.

[75] Thin Nguyen, Svetha Venkatesh, and Dinh Phung. Textual cues for online depression in community and personal settings. In *International Conference on Advanced Data Mining and Applications*, 2016.

[76] Iram Fatima, Hamid Mukhtar, Hafiz Farooq Ahmad, and Kashif Rajpoot. Analysis of user-generated content from online social communities to characterise and predict depression degree. *SAGE Publications*, (5), 2018.

[77] Songqiao Han, Hailiang Huang, and Yuqing Tang. Knowledge of words: An interpretable approach for personality recognition from social media. *Knowledge-Based Systems*, 194(2):105550, 2020.

[78] Víctor M Prieto, Sérgio Matos, Manuel Alvarez, Fidel Cacheda, and José Luís Oliveira. Twitter: A good place to detect health conditions. *PLoS ONE*, 9(1):e86191, 2014.

[79] Wesley Ramos Dos Santos, Amanda Maria Martins Funabashi, and Ivandre Paraboni. Searching brazilian twitter for signs of mental health issues. In *International Conference on Language Resources and Evaluation*, 2020.

[80] Thin Nguyen, O'Dea, Bridianne, Larsen, Mark, Dinh Phung, Venkatesh, Svetha, Christensen, and Helen. Using linguistic and topic analysis to classify sub-groups of online depression communities.

[81] Xuetong Chen, Martin Sykora, Thomas Jackson, Suzanne Elayan, and Fehmidah Munir. Tweeting your mental health: an exploration of different classifiers and features with emotional signals in identifying mental health conditions. In *Hawaii International Conference on System Sciences*, 2018.

[82] Victor Leiva and Ana Freire. Towards suicide prevention: Early detection of depression on social media. In *International Conference on Internet Science*, 2017.

[83] Quan Hu, Ang Li, Fei Heng, Jianpeng Li, and Tingshao Zhu. Predicting depression of social media user on different observation windows. 2015.

[84] Zhichao Peng, Qinghua Hu, and Jianwu Dang. Multi-kernel svm based depression recognition using social media data. *International Journal of Machine Learning and Cybernetics*, 2017.

[85] Subhan Tariq, Nadeem Akhtar, Humaira Afzal, Shahzad Khalid, and Ghufran Ahmad. A novel co-training-based approach for the classification of mental illnesses using social media posts. *IEEE Access*, PP(99):1–1, 2019.

[86] Sergio G. Burdisso, Marcelo Errecalde, and Manuel Montes y Gómez. t-ss3: a text classifier with dynamic n-grams for early risk detection over text streams. 2019.

[87] AntoineBriand, HaydaAlmeida, and Marie-JeanMeurs. Analysis of social media posts for early detection of mental health conditions. In *Canadian Conference on Artificial Intelligence*, 2018.

[88] Fidel Cacheda, Diego Fernandez, Francisco J Novoa, and Victor Carneiro. Early detection of depression: Social network analysis and random forest techniques. *Journal of Medical Internet Research*, 21(6), 2019.

[89] Marcel Trotzek, Sven Koitka, and Christoph M. Friedrich. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge Data Engineering*, pages 1–1, 2018.

[90] Jina Kim, Jieon Lee, Eunil Park, and Jinyoung Han. A deep learning model for detecting mental illness from user content on social media. *Scientific Reports*, 10(1), 2020.

[91] Guozheng Rao, Yue Zhang, Li Zhang, Qing Cong, and Zhiyong Feng. Mgl-cnn: A hierarchical posts representations model for identifying depressed individuals in online forums. *IEEE Access*, PP(99):1–1, 2020.

[92] Amna Amanat, Muhammad Rizwan, Abdul Rehman Javed, Maha Abdelhaq, Raed Alsaqour, Sharnil Pandya, and Mueen Uddin. Kinnaird college for women researchers describe findings in electronics (deep learning for depression detection from textual data). *Electronics Newsweekly*, (Mar.29), 2022.

[93] Hussain Ahmad, Muhammad Zubair Asghar, Fahad M. Alotaibi, and Ibrahim A. Hameed. Applying deep learning technique for depression classification in social media text. *Journal of Medical Imaging and Health Informatics*, 2020.

[94] Qing Cong, Zhiyong Feng, Fang Li, Yang Xiang, and Cui Tao. X-a-bilstm: a deep learning approach for depression detection in imbalanced data. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2018.

[95] Mario Ezra Aragón, A. Pastor López-Monroy, Luis C. González, and Manuel Montes y Gómez. Attention to emotions: Detecting mental disorders in social media. 2020.

[96] Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. Depressionnet: A novel summarization boosted deep framework for depression detection on social media. 2021.

[97] Dalia Magaa. Cultural competence and metaphor in mental healthcare interactions: A linguistic perspective. *Patient Education and Counseling*, 102(12):2192–2198, 2019.

[98] Joy Llewellyn-Beardsley, Stefan Rennick-Egglestone, Felicity Callard, Paul Crawford, Marianne Farkas, Ada Hui, David Manley, Rose Mcgranahan, Kristian Pollock, and Amy Ramsay. Characteristics of mental health recovery narratives: Systematic review and narrative synthesis. *Plos One*, 14(3), 2019.

[99] Dongyu Zhang, Nan Shi, Ciyuan Peng, Abdul Aziz, and Feng Xia. Mam: A metaphor-based approach for mental illness detection. 2021.

[100] Hongyu Gong, Kshitij Gupta, Akriti Jain, and Suma Bhat. Illinimet: Illinois system for metaphor detection with contextual and linguistic information. In *Meeting of the Association for Computational Linguistics*, 2020.

[101] Rui Mao, Chenghua Lin, and Frank Guerin. End-to-end sequential metaphor identification inspired by linguistic theories. In *Meeting of the Association for Computational Linguistics*, 2019.

[102] Abdulqader M. Almars. Attention-based bi-lstm model for arabic depression classification. *计算机,材料和连续体(英文)*, 2022.

[103] Lu Ren, Hongfei Lin, Bo Xu, Shaowu Zhang, and Shichang Sun. Depression detection on reddit with an emotion-based attention network: Algorithm development and validation. *JMIR Medical Informatics*, 9(7):e28754, 2021.

[104] Hoyun Song, Jinseon You, Jin Woo Chung, and Jong C. Park. Feature attention network: Interpretable depression detection from social media. In *Pacific Asia Conference on Language, Information, and Computation*, 2018.

[105] Ana Sabina Uban, Berta Chulvi, and Paolo Rosso. On the explainability of automatic predictions of mental disorders from social media data. 2021.

[106] Hamad Zogan, Imran Razzak, Xianzhi Wang, Shoaib Jameel, and Guandong Xu. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web*, 25(1):281–304, 2022.

[107] Shweta Yadav, Jainish Chauhan, Joy Prakash Sain, Krishnaprasad Thirunarayan, and Jeremiah Schumm. Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework. 2020.

[108] Xiaofeng Wang, Shuai Chen, Tao Li, Wanting Li, Yejie Zhou, Jie Zheng, Qingcai Chen, Jun Yan, and Buzhou Tang. Depression risk prediction for chinese microblogs via deep-learning methods: Content analysis. *JMIR Medical Informatics*, 8(7).

[109] Muhammad Ali, Jamal Hussain Shah, Muhammad Attique Khan, Majed Alhaisoni, Usman Tariq, Tallha Akram, Ye Jin Kim, and Byoungchol Chang. Brain tumor detection and classification using pso and convolutional neural network. *计算机、材料和连续体(英文)*, (012):000, 2022.

[110] Pervaiz Iqbal Khan, Imran Razzak, Andreas Dengel, and Sheraz Ahmed. A novel approach to train diverse types of language models for health mention classification of tweets. *arXiv e-prints*, 2022.

[111] Paul Harmon and David King. Expert systems. *International Review of Law Computers Technology*, 51(4):142–144, 1985.

[112] A Colmerauer. Prolog and infinite trees. *Logic Programming*, 1982.

[113] David E Losada, Fabio Crestani, and Javier Parapar. Clef 2017 erisk overview: Early risk prediction on the internet: Experimental foundations. In *CLEF (Working Notes)*, 2017.

[114] David E Losada, Fabio Crestani, and Javier Parapar. Overview of erisk: early risk prediction on the internet. In *International conference of the cross-language evaluation forum for european languages*, pages 343–361. Springer, 2018.

[115] David E Losada, Fabio Crestani, and Javier Parapar. Overview of erisk 2019 early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 340–357. Springer, 2019.

[116] David E Losada, Fabio Crestani, and Javier Parapar. Overview of erisk at clef 2020: Early risk prediction on the internet (extended overview). *CLEF (Working Notes)*, 2020.

[117] Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. Overview of erisk at clef 2021: Early risk prediction on the internet (extended overview). *CLEF (Working Notes)*, 2021.

[118] Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. Evaluation report of erisk 2022: Early risk prediction on the internet. *CLEF (Working Notes)*, 2022.

[119] Minsu Park, David McDonald, and Meeyoung Cha. Perception differences between the depressed and non-depressed users in twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, pages 476–485, 2013.

[120] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th annual ACM web science conference*, pages 47–56, 2013.

[121] Javier Parapar, David E Losada, and Alvaro Barreiro. A learning-based approach for the identification of sexual predators in chat logs. In *CLEF (Online working notes/labs/workshop)*, volume 1178, 2012.

[122] Megan A Moreno, Lauren A Jelenchick, Katie G Egan, Elizabeth Cox, Henry Young, Kerry E Gannon, and Tara Becker. Feeling bad on facebook: Depression disclosures by college students on a social networking site. *Depression and anxiety*, 28(6):447–455, 2011.

[123] Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133, 2004.

[124] Sidney J Blatt. *Experiences of depression: Theoretical, clinical, and research perspectives.* American Psychological Association, 2004.

[125] M.D. Aaron T. Beck and Ph.D. Brad A. Alford. *Depression: Causes and Treatment*. University of Pennsylvania Press, 2014.

[126] Jonathan Rottenberg. Mood and emotion in major depression. *Current Directions in Psychological Science*, 14(3):167–170, 2005.

[127] Jutta Joormann and Colin H Stanton. Examining emotion regulation in depression: A review and future directions. *Behaviour research and therapy*, 86:35–49, 2016.

[128] Maxim Stankevich, Vadim Isakov, Dmitry Devyatkin, and Ivan V Smirnov. Feature engineering for depression detection in social media. In *ICPRAM*, pages 426–431, 2018.

[129] Tiancheng Shen, Jia Jia, Guangyao Shen, Fuli Feng, Xiangnan He, Huanbo Luan, Jie Tang, Thanassis Tiropanis, Tat Seng Chua, and Wendy Hall. Cross-domain depression detection via harvesting social media. International Joint Conferences on Artificial Intelligence, 2018.

[130] Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. Recognizing depression from twitter activity. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 3187–3196, 2015.

[131] Andrew Yates, Arman Cohan, and Nazli Goharian. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*, 2017.

[132] Alex Rinaldi, Jean E Fox Tree, and Snigdha Chaturvedi. Predicting depression in screening interviews from latent categorization of interview prompts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7–18, 2020.

[133] Shaoxiong Ji, Celina Ping Yu, Sai-fu Fung, Shirui Pan, and Guodong Long. Supervised learning for suicidal ideation detection in online user content. *Complexity*, 2018, 2018.

[134] Shaoxiong Ji, Xue Li, Zi Huang, and Erik Cambria. Suicidal ideation and mental disorder detection with attentive relation networks. *Neural Computing and Applications*, pages 1–11, 2021.

[135] Luna Ansari, Shaoxiong Ji, Qian Chen, and Erik Cambria. Ensemble hybrid learning methods for automated depression detection. *IEEE Transactions on Computational Social Systems*, 2022.

[136] Chenghao Yang, Yudong Zhang, and Smaranda Muresan. Weakly-supervised methods for suicide risk assessment: Role of related domains. *arXiv preprint arXiv:2106.02792*, 2021.

[137] Jutta Joormann and Ian H Gotlib. Updating the contents of working memory in depression: interference from irrelevant negative material. *Journal of abnormal psychology*, 117(1):182, 2008.

[138] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[139] Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM*

SIGKDD international conference on knowledge discovery and data mining, pages 65–74, 2017.

[140] Dongyu Zhang, Jidapa Thadajarassiri, Cansu Sen, and Elke Rundensteiner. Time-aware transformer-based network for clinical notes series prediction. In *Machine Learning for Healthcare Conference*, pages 566–588. PMLR, 2020.

[141] Sergio G Burdisso, Marcelo Errecalde, and Manuel Montes-y Gómez. A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications*, 133:182–197, 2019.

[142] David E Losada and Fabio Crestani. A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 28–39. Springer, 2016.

[143] Erik Cambria, Qian Liu, Sergio Decherchi, Frank Xing, and Kenneth Kwok. Senticnet 7: A commonsense-based neurosymbolic ai framework for explainable sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3829–3839, 2022.

[144] Farig Sadeque, Dongfang Xu, and Steven Bethard. Measuring the latency of depression detection in social media. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 495–503, New York, NY, USA, 2018. Association for Computing Machinery.

[145] Patricia Martın-Rodilla, David E Losada, and Fabio Crestani. Overview of erisk 2022: Early risk prediction on the internet. In *Proceedings of the 13th International Conference of the CLEF Association*, pages 233–262. Springer Nature, 2022.