Integrating Textual and Emotional Dynamics for Accurate Detection of Mental Health Disorders in Social Media

Jiajun Zou[1]✉(zoujj18@tsinghua.org.cn), Zhixiao Qi[2](qizx24@mails.tsinghua.edu.cn), Jinshuai Yang[1](yjs20@mails.tsinghua.edu.cn), Zhechen Wei[1](weizc22@mails.tsinghua.edu.cn), Congqi Wang[1](wangcq22@mails.tsinghua.edu.cn), Shizhong Yang[3] (ysza02008@btch.edu.cn), Minghu Jiang[2](jiang.mh@mail.tsinghua.edu.cn), Yongfeng Huang[1,4]✉((yfhuang@mail.tsinghua.edu.cn)

[1]Department of Electronic Engineering, Tsinghua University, Beijing, China
[2]School of Humanities, Tsinghua University, Beijing, China
[3]Institute for Precision Medicine, Tsinghua University, Beijing, China
[4]Zhongguancun Laboratory, Beijing, China

Abstract

Mental health disorders impact nearly one billion people worldwide, yet stigma and insufficient awareness often prevent individuals from seeking timely professional help. The proliferation of social media platforms has introduced new opportunities for detection of mental health conditions, enabling the analysis of user-generated content to identify whether one has a mental disorder. Traditional approaches to this task have largely relied on content-based models, such as n-grams or language embeddings, which are prone to domain-specific biases and often fail to account for the emotional dynamics inherent in mental health expressions. In this work, we propose a novel framework for detecting mental disorders through the analysis of Reddit conversations, integrating both temporal textual data and emotional cues. Our model addresses the limitations of prior methods by explicitly capturing the evolving relationship between textual content and emotional expression over time. Experimental results demonstrate a significant improvement in detection accuracy compared to existing approaches, while ablation studies highlight the critical role of temporal emotional information in enhancing performance. These findings suggest that a more nuanced, emotion-aware approach offers substantial promise for advancing computational mental health diagnostics.

Keywords: mental illness detection; emotion fusion; emotional states; social media

Introduction

Emotion has long been a focal point of human inquiry, explored extensively in philosophy, literature, and the arts. Over time, the understanding of emotions has evolved from abstract speculation to empirical study, becoming a central area of psychological research. The nuanced fluctuations of emotion—particularly those subtle, imperceptible changes over time—offer valuable insights into an individual's psychological well-being, often serving as indicators of mental health conditions such as depression and anxiety. Emotional dysregulation, for example, is widely recognized as a core feature of depression (Rottenberg, 2017).

In recent years, social media has emerged as a novel platform for emotional expression, providing a vast, dynamic social environment where individuals document their psychological states. Textual data generated on these platforms capture the emotional trajectories and psychological shifts of users, offering a unique lens into mental health that is unavailable in traditional laboratory settings. According to the Social Compensation Theory (SCT), individuals with mental health challenges often turn to social media for support, particularly when navigating personal difficulties (O'Day & Heimberg, 2021). This new mode of communication and emotional exchange presents rich opportunities for studying mental health outside conventional research paradigms, positioning social media as a powerful tool in mental health research (Zhang, Yang, Ji, & Ananiadou, 2023; Roy et al., 2020; Kelley & Gillan, 2022; Guo, Sun, & Vosoughi, 2021).

The detection of mental health disorders through social media content has become an increasingly critical application of Natural Language Processing (NLP). The World Health Organization (WHO) estimates that one in eight individuals globally is affected by mental disorders, equating to approximately one billion people[1]. This number has been further exacerbated by the COVID-19 pandemic (Ettman et al., 2022). Despite the widespread prevalence of mental illness, societal stigma, ignorance surrounding mental health assessments, and a lack of accessible care often result in underdiagnosis and inadequate treatment (Zhang et al., 2023). As a result, detection and intervention through social media analysis have emerged as promising avenues.

Current approaches to mental health detection rely heavily on sentiment analysis, machine learning, and deep learning techniques (Zhang et al., 2023; Kabir et al., 2022; Cha, Kim, & Park, 2022; Roy et al., 2020). While these methods have shown promise, they tend to focus primarily on textual content, often overlooking the emotional undercurrents that accompany written expressions. This limitation is particularly notable in cases where the emotional state of an individual evolves over time. Although some studies have attempted to integrate textual and emotional features, these efforts typically treat them as isolated components, without fully exploring the dynamic and temporal relationship between emotion and language. As a result, existing models often fail to capture the complexity of how emotional states fluctuate and interact with linguistic expression, ultimately reducing the accuracy and robustness of mental health detection.

In this paper, we introduce a novel approach that integrates both historical textual data and emotional information to enhance the accuracy of mental health disorder detection.

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

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

By fusing these two elements, our method aims to provide a more comprehensive understanding of the interplay between language and emotion, ultimately improving detection performance.

Related Works

Significant advances have been made in the identification of mental disorders through the analysis of social media content, leveraging both machine learning-based methods (Gamon, Choudhury, Counts, & Horvitz, 2013; Guntuku, Giorgi, & Ungar, 2018; Benamara, Moriceau, Mothe, Ramiandrisoa, & He, 2018) and deep learning-based methods (Abdullah & Negied, 2024; Vandana, Marriwala, & Chaudhary, 2023; Uban, Chulvi, & Rosso, 2022; Yadav, Shinde, & Shedge, 2023). For instance, a random forest model has been effectively utilized to detect suicidal ideation risks within social media data (Roy et al., 2020). More recently, deep learning-based methods have gained traction due to their ability to capture complex patterns in large-scale, unstructured textual data. In a similar vein, Kabir et al. (Kabir et al., 2022) employed BERT (Devlin, Chang, Lee, & Toutanova, 2019) and other advanced models to assess the severity of depression presented in X (formerly Twitter) posts, contributing a dataset that captures a spectrum of depressive symptoms. Cha et al. (Cha et al., 2022) developed a depression-specific dictionary to identify users exhibiting depressive tendencies on X. Conversely, Kim (Kim, Lee, Park, & Han, 2020) applied deep learning algorithms to classify a wide range of mental health disorders, including depression, anxiety, bipolar disorder, borderline personality disorder, schizophrenia, and autism, using Reddit data as a source for diagnostic classification.

While these approaches have predominantly focused on textual content analysis, a growing body of research suggests that emotional expression plays a crucial role in enhancing detection accuracy. For instance, Guo et al. (Guo et al., 2021) introduced a novel framework that incorporates emotional features alongside textual content by leveraging distant supervision to build a mental disorder dataset from Reddit. Their findings emphasized that while content-based models often suffer from domain-specific biases and limited generalizability across different contexts, emotional features—particularly those reflecting emotional transitions—demonstrate a more robust ability to generalize across various platforms and user behaviors. This work underscores the importance of considering both text and emotion as complementary signals for improving the detection of mental health conditions. Further research has corroborated the importance of integrating textual and emotional data for more robust mental health detection (Zhang et al., 2023).

Recent studies have shown that such integrated models offer improved performance in identifying a broader range of mental health conditions, highlighting the limitations of models that focus solely on textual features. Moreover, emotional dynamics are now recognized as critical in capturing the subtleties of mental health conditions, as they reflect the psychological fluctuations that often precede clinical diagnoses. These advancements underscore the potential of combining sentiment analysis with emotion-aware models to enhance the accuracy and generalizability of mental health diagnostics in online settings.

In summary, the evolving body of work in social media-based mental health detection has begun to emphasize the need for a more nuanced approach that incorporates both linguistic and emotional cues. The integration of these features promises to overcome the limitations of traditional methods and pave the way for more accurate, scalable, and contextually adaptive systems for detection and intervention in mental health care. And in this paper, we focused on textual and emotional information in the temporal dimension.

Methodology

In this section, we will introduce the proposed method.

Problem Definition

Given a user U, the historical sequence of comment texts is denoted as $D = [D_{T_1}, D_{T_2}, ..., D_{T_M}]$, where D_{T_i} represents the comment posted at time instance i, and the total number of comment texts is M. Each D_{T_i} comprises K sentences, with each sentence annotated with an emotional category e, thus the sequence of emotional labels corresponding to D_{T_i} is $E = [e_1, e_2, ..., e_K]$. For each user U, it is necessary to detect whether he or she is a patient with a mental disorder based on the comment texts and the associated emotional labels.

Model Structure

The model proposed in this study is shown in Figure 1. This model integrates historical text information with emotional features.

Initially, this study concatenates each historical comment text D_{T_i} of every user U with the sequence of emotional labels $E_{T_i} = [e_1, e_2, ..., e_K]$, where $e_i \in E_o = [e_1^o, e_2^o, ..., e_m^o]$. To eliminate redundancy, this study deduplicates E_{T_i}, resulting in $E_p^i = [E_{p_0}, E_{p_1}, ..., E_{p_J}]$, and simultaneously obtains each emotion and its corresponding count: $\{(E_{p_0} : N_0), (E_{p_1} : N_1), ..., (E_{p_J} : N_J)\}$. The final concatenation outcome is: $D_{T_i}^e = E_p^i + D_{T_i}$. Subsequently, BERT is utilized to encode the concatenated text, yielding:

$$H_{T_i}^e = \text{BERT}(D_{T_i}^e) = \text{BERT}(E_p^i + D_{T_i}), \tag{1}$$

at this point, each user's comment text is infused with emotional information.

Subsequently, this study incorporates the count of each emotion from D_{T_i} into $H_{T_i}^e$, resulting in the following equation:

$$H_{T_i}^{en} = [H_{T_i}^e, N_0, N_1, ..., N_J], \tag{2}$$

where $H_{T_i}^{en}$ denotes the representation after incorporating the counts of emotional labels, reflecting the proportion of

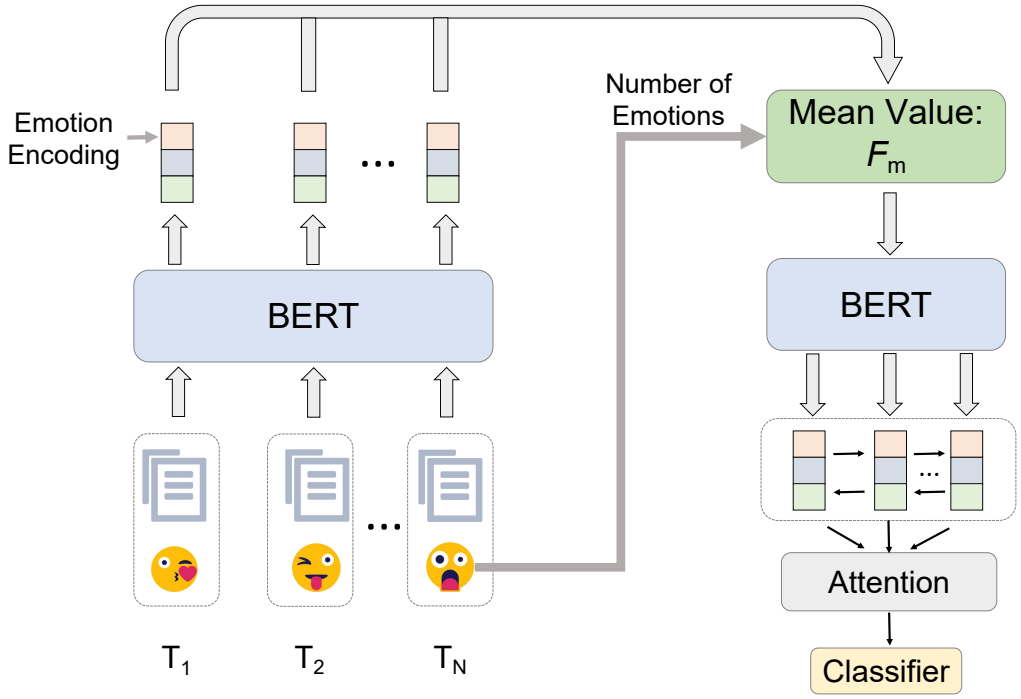

Figure 1: The overall architecture of the proposed model.

emotions in each post within the model input. Next, the text averages the emotional representations with all comment texts for each user:

$$H_U = \frac{\sum_i^M H_{T_i}^{en}}{M}, \tag{3}$$

where H_U represents the final feature for each user, utilized as input for model training.

Building upon the aforementioned foundation, this study further encodes the user features with BERT to integrate the emotional proportion information within the global context. Subsequently, a Bidirectional Long Short-Term Memory (BiLSTM) (Hochreiter & Schmidhuber, 1997) network is employed to model the contextual features of the input information. Thereafter, an attention mechanism (Vaswani et al., 2017) is utilized to focus on salient features. The process is as follows:

$$H_U^e = \text{BERT}(H_U), \tag{4}$$

$$H_U^{LSTM} = \text{BiLSTM}(H_U), \tag{5}$$

$$H_U^o = \text{Attention}(H_U^{LSTM}, H_U^{LSTM}), \tag{6}$$

ultimately, H_U^o is fed into a classifier to obtain the probability that the user has a mental disorder:

$$p_{U_i}^o = \text{softmax}(WH_U^o + \mathbf{b}), \tag{7}$$

where W and $\mathbf{b}$ are parameters trained in the model.

This study employs binary cross-entropy loss as the optimization objective, where y_i denotes the true label indicating whether the user has the disorder:

$$\mathcal{L} = -\frac{1}{N}\sum_i \left[y_i \cdot \log p_{U_i}^o + (1-y_i)\cdot \log(1-p_{U_i}^o)\right]. \tag{8}$$

Experimental Setup

For ease of comparison, this study utilizes the dataset from Guo et al. (Guo et al., 2021), which was derived from Reddit user texts and categorizes users into four groups: bipolar disorder, depression, anxiety, and a control group, with 1997 users in each category. The number of texts per group is as follows: 686,359 for bipolar disorder, 914,082 for depression, 686,369 for anxiety, and 2,516,696 for the control group. During detection, the control group is mixed with one of the mental disorder datasets (Guo et al., 2021), thereby forming a binary classification task.

In the experimental process, consistent with the work of Guo et al., the BERT-large model was employed with a text length of 140, a batch size of 16, a maximum of 1000 training epochs, a learning rate of 0.0005, and the use of k-fold cross-validation (with k set to 5), along with the AdamW optimizer. The evaluation metrics for the method include accuracy, precision, recall, and the $F1$ score.

To validate the effectiveness of this approach, the comparative methods include: Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Term Frequency-Inverse Document Frequency (tf-idf), BERT-large model, and the Emotion Representation (ER) method (Guo et al., 2021). This study directly adopts the results reported in the paper (Guo et al., 2021).

Table 1: Comparison of detection performance of different models for bipolar disorder

	Accuracy	$F1$	Precision	Recall
SVM	79.1	79.0	79.1	79.5
LogReg	79.5	79.4	79.5	80.0
RF	85.1	85.0	85.1	85.5
tf-idf	83.2	83.2	83.2	83.2
BERT	86.6	86.6	86.6	86.6
ER	86.6	86.5	86.9	86.6
Ours	**88.9**	**88.9**	**88.9**	**89.0**

Table 2: Comparison of detection performance of different models for depression disorder

	Accuracy	$F1$	Precision	Recall
SVM	77.1	77.1	77.1	77.5
LogReg	77.4	77.3	77.4	78.0
RF	80.7	80.6	80.7	81.7
tf-idf	77.4	77.4	77.4	77.4
BERT	82.5	82.5	82.6	82.5
ER	83.2	83.1	83.9	83.2
Ours	**84.0**	**83.9**	**84.0**	**84.7**

Table 3: Comparison of detection performance of different models for anxiety disorder

	Accuracy	$F1$	Precision	Recall
SVM	78.2	78.1	78.2	78.9
LogReg	79.1	79.1	79.1	79.5
RF	83.9	83.8	83.9	84.4
tf-idf	80.6	80.6	80.7	80.6
BERT	84.1	84.1	84.1	84.1
ER	85.3	85.2	85.7	85.3
Ours	**86.6**	**86.6**	**86.8**	**86.8**

Experimental Results

Model Performance

Tables 1, 2, and 3 present the detection performance of various methods for bipolar disorder, depression, and anxiety, respectively. The results demonstrate a clear advantage of the proposed method in detecting individuals with these mental health conditions.

For bipolar disorder detection, our approach outperforms the next best model by notable margins of 2.3%, 2.4%, 2.0%, and 2.4% in accuracy, $F1$ score, precision, and recall, respectively. In the case of depression, the proposed method shows improvements of 0.8%, 0.8%, 0.1%, and 1.5% across these same metrics. Similarly, for anxiety detection, the method yields improvements of 1.3%, 1.4%, 1.1%, and 1.5% over the next best model.

These results underscore the effectiveness of integrating historical textual data with emotional information in enhancing the detection of mental health disorders. The observed improvements across all three conditions suggest that the incorporation of emotional cues significantly contributes to the model's overall performance, providing a more comprehensive approach to identifying users with mental health challenges.

Ablation Study

Additionally, we conducted ablation experiments to assess the contribution of temporal emotional information in the detection of mental disorders. In these experiments, the emotional content extracted from the comment texts was removed, effectively excluding the explicit role of emotion in the model. The results of this ablation study are presented in Figure 2, which illustrates the impact on the model's performance across the three mental disorders, with performance metrics focused on the $F1$ score and accuracy.

From Figure 2, it is evident that the removal of emotional information leads to a noticeable decline in detection performance across all three disorders. This reduction underscores the critical role of emotional cues in improving the model's accuracy and $F1$ score. The temporal variations in emotional expression exhibited by individuals with mental disorders play a pivotal role in distinguishing between different conditions, offering valuable insights into the underlying psychological dynamics.

Collectively, these ablation results further validate the effectiveness of the proposed approach in mental health detection. They emphasize the substantial benefit of integrating historical text data with emotional features, highlighting that the combination of both elements is essential for accurate and nuanced mental disorder detection.

Discussion

The experimental results on public datasets demonstrate that the proposed method significantly outperforms existing approaches in detecting mental disorders. Ablation experiments further validate the pivotal role of emotional information, particularly the temporal variations in historical emotional states, in enhancing detection performance. These findings highlight the importance of incorporating emotional dynamics into models designed for mental health detection, offering a pathway toward improving both performance and interpretability.

Despite these advancements, the approach described in this study employs a simplistic averaging mechanism to integrate historical information, which may inadequately capture the complex, non-linear relationships between historical context and current textual and emotional expressions. The influence of past emotional states on present behavior is inherently non-linear and may exhibit varying degrees of significance over time. Addressing this limitation requires more sophisticated techniques, such as utilizing closed-form

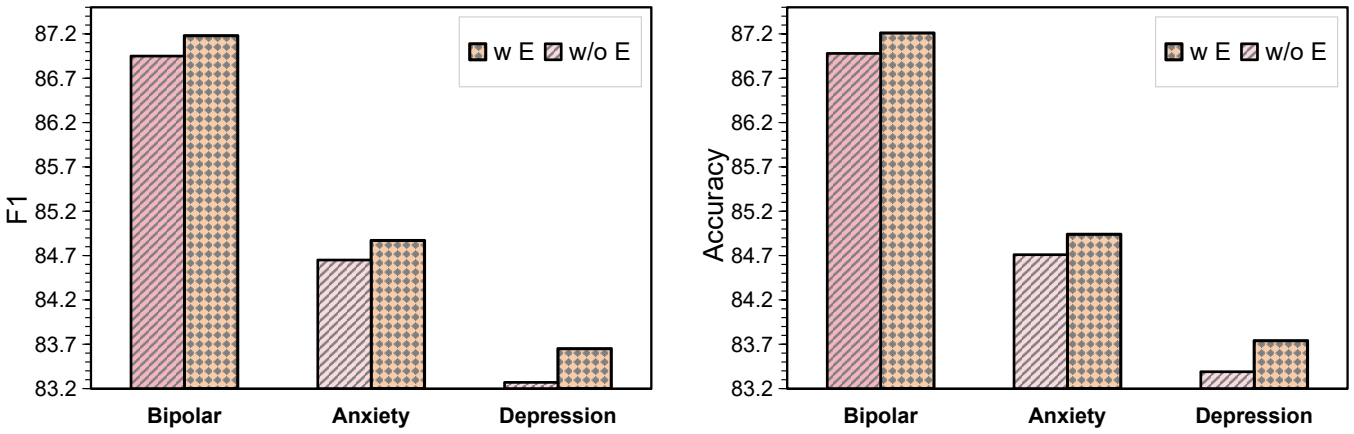

Figure 2: The results of the ablation study.

continuous-depth neural networks (Hasani et al., 2022). These advanced approaches could enable a more precise characterization of the interplay between historical and current information.

Furthermore, this study adopts a binary classification framework for detecting mental disorders, consistent with prior research (Guo et al., 2021). While effective for comparative evaluation, this simplification does not reflect the complexity of real-world scenarios. Social media platforms feature a diverse population of users, many of whom experience overlapping symptoms of multiple mental health conditions. To better align with clinical realities, future research should explore multi-class or multi-label classification tasks. Such approaches would allow the detection of multiple co-occurring disorders in a single user and could provide a more granular understanding of the spectrum of mental health challenges.

In addition to these technical considerations, there remains the challenge of ethical application. While leveraging social media data for mental health detection has great potential, it also raises concerns about privacy, consent, and the risk of stigmatization. Future work must ensure that these technologies are deployed in a way that prioritizes user rights, safeguards sensitive information, and promotes responsible use of findings.

Furthermore, the proposed framework demonstrates measurable improvements in classification accuracy; however, it retains an inherently opaque nature that limits clinical interpretability. In mental health prediction contexts where clinical decision-making requires transparent diagnostic rationale, the development of explainable artificial intelligence components becomes imperative. Future implementations could benefit from integrating Shapley Additive Explanations (SHAP) (Rodríguez-Pérez & Bajorath, 2019) values to quantify feature importance distributions, coupled with attention heatmap visualizations that elucidate temporal dependencies between lexical patterns and affective states. Such interpretability mechanisms would provide clinicians with an explicable rationale for model predictions.

By addressing these challenges, the field can move closer to creating robust, interpretable, and ethically sound tools for mental health detection, ultimately improving diagnosis and intervention for individuals in need.

Conclusion

This study introduces a novel method for detecting mental disorders by integrating historical textual and emotional information, with a focus on the temporal dynamics of emotional expression. By leveraging the interplay between linguistic content and emotional trajectories, the proposed approach significantly improves detection accuracy, as demonstrated through experimental validation on public datasets. This work underscores the transformative potential of incorporating emotional dynamics into computational frameworks for mental health research, offering both a practical tool for detection and a conceptual advance in the study of emotion and mental health. Future research should further refine this approach, exploring its applicability across diverse populations and extending its use to support clinical decision making.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 82090053 and 62036001. It was also supported by Tsinghua University Initiative Scientific Research Program of Precision Medicine (2022ZLA007). We also express our high respect for the CogSci 2025 organizers and our sincere gratitude to the reviewers of the papers.

References

Abdullah, M., & Negied, N. (2024, 01). Detection and prediction of future mental disorder from social media data using machine learning, ensemble learning, and large language models. *IEEE Access*, *PP*, 1-1. doi: 10.1109/ACCESS.2024.3406469

Benamara, F., Moriceau, V., Mothe, J., Ramiandrisoa, F., & He, Z. (2018). Automatic detection of depressive users in social media. In *Conférence en recherche d'infomations et applications.*

Cha, J., Kim, S., & Park, E. (2022). A lexicon-based approach to examine depression detection in social media: the case of twitter and university community. *Humanities & Social Sciences Communications*, *9*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Ettman, C., Cohen, G., Abdalla, S., Trinquart, L., Castrucci, B., Bork, R., ... Galea, S. (2022, 03). Assets, stressors, and symptoms of persistent depression over the first year of the covid-19 pandemic. *Science Advances*, *8*, eabm9737.

Gamon, M., Choudhury, M., Counts, S., & Horvitz, E. (2013, 07). Predicting depression via social media..

Guntuku, S. C., Giorgi, S., & Ungar, L. (2018, June). Current and future psychological health prediction using language and socio-demographics of children for the CLPysch 2018 shared task. In K. Loveys, K. Niederhoffer, E. Prud'hommeaux, R. Resnik, & P. Resnik (Eds.), *Proceedings of the fifth workshop on computational linguistics and clinical psychology: From keyboard to clinic* (pp. 98–106). New Orleans, LA: Association for Computational Linguistics.

Guo, X., Sun, Y., & Vosoughi, S. (2021). Emotion-based modeling of mental disorders on social media. In *Ieee/wic/acm international conference on web intelligence and intelligent agent technology* (pp. 8–16).

Hasani, R., Lechner, M., Amini, A., Liebenwein, L., Ray, A., Tschaikowski, M., ... Rus, D. (2022). Closed-form continuous-time neural networks. *Nature Machine Intelligence*, *4*(11), 992–1003.

Hochreiter, S., & Schmidhuber, J. (1997, 11). Long short-term memory. *Neural Computation*, *9*, 1735-1780.

Kabir, M., Ahmed, T., Hasan, M. B., Laskar, M. T. R., Joarder, T. K., Mahmud, H., & Hasan, K. (2022). Deptweet: A typology for social media texts to detect depression severities. *Comput. Hum. Behav.*, *139*, 107503.

Kelley, S. W., & Gillan, C. M. (2022). Using language in social media posts to study the network dynamics of depression longitudinally. *Nature Communications*, *13*.

Kim, J., Lee, J., Park, E., & Han, J. (2020). A deep learning model for detecting mental illness from user content on social media. *Scientific Reports*, *10*.

O'Day, E. B., & Heimberg, R. G. (2021). Social media use, social anxiety, and loneliness: A systematic review. *Computers in Human Behavior Reports*, *3*, 100070.

Rodríguez-Pérez, R., & Bajorath, J. (2019). Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values. *Journal of Medicinal Chemistry*, *63*(16), 8761–8777.

Rottenberg, J. (2017). Emotions in depression: What do we really know? *Annual review of clinical psychology*, *13*, 241-263.

Roy, A., Nikolitch, K., McGinn, R., Jinah, S., Klement, W., & Kaminsky, Z. A. (2020). A machine learning approach predicts future risk to suicidal ideation from social media data. *NPJ Digital Medicine*, *3*.

Uban, A.-S., Chulvi, B., & Rosso, P. (2022). Explainability of depression detection on social media: From deep learning models to psychological interpretations and multimodality. In F. Crestani, D. E. Losada, & J. Parapar (Eds.), *Early detection of mental health disorders by social media monitoring: The first five years of the erisk project* (pp. 289–320). Cham: Springer International Publishing.

Vandana, Marriwala, N., & Chaudhary, D. (2023). A hybrid model for depression detection using deep learning. *Measurement: Sensors*, *25*, 100587.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st international conference on neural information processing systems* (p. 6000–6010). Red Hook, NY, USA: Curran Associates Inc.

Yadav, P., Shinde, S., & Shedge, R. (2023). Mental health disorder detection using machine learning and deep learning tehniques. In *2023 3rd asian conference on innovation in technology (asiancon)* (p. 1-6).

Zhang, T., Yang, K., Ji, S., & Ananiadou, S. (2023). Emotion fusion for mental illness detection from social media: A survey. *Information Fusion*, *92*, 231-246.