Improvement of Depression Type Classification by Patient Dialogue Embedding

Waranrach Viriyavit
Faculty of Informatics
Burapha University
Chonburi, Thailand
waranrach.vi@go.buu.ac.th

Pipat Khambun
Faculty of Informatics
Burapha University
Chonburi, Thailand
64160206@go.buu.ac.th

Ponlawat Chophuk
Faculty of Informatics
Burapha University
Chonburi, Thailand
ponlawat.ch@go.buu.ac.th

Supawadee Srikamdee
Faculty of Informatics
Burapha University
Chonburi, Thailand
srikamdee@buu.ac.th

Thatsanee Charoenporn
Asia AI Institute (AAII)
Faculty of Data Science
Musashino University
Tokyo, Japan
thatsanee@ds.musashino-u.ac.jp

Virach Sornlertlamvanich
Asia AI Institute (AAII)
Faculty of Data Science
Musashino University
Tokyo, Japan
virach@gmail.com

Abstract— ***Depression is a significant global health concern, and early detection is important for effective treatment. This study investigates the potential of using dialogue context to enhance the accuracy of depression detection. By analyzing dialogues based on the nine symptoms of the Patient Health Questionnaire-9 (PHQ-9), we developed a classifier that categorizes dialogues according to their relevance to specific PHQ-9 symptoms. Our results demonstrate that our LSTM-based dialogue approach achieved an average classification accuracy of 84.37%, significantly outperforming the baseline single-sentence analysis (80.00%). These findings suggest that analyzing dialogues can better capture contextual information often missed in single-sentence analysis, leading to more accurate classification of depressive symptoms aligned with the PHQ-9.***

Keywords- text classification, depression detection, Patient Health Questionnaire-9, natural language processing, deep learning, long-short term memory

I. INTRODUCTION

Depression is a widespread mental health disorder affecting millions globally, with significant emotional, social, and economic consequences. The World Health Organization estimates that over 264 million people are affected, emphasizing the need for effective detection and intervention [1]. The early detection of depressive symptoms is of critical importance in a clinical setting, as prompt intervention has been shown to correlate with improved patient outcomes and enhanced quality of life.

Existing approaches to depression detection have predominantly relied on keyword extraction and single-sentence analysis [2], which inherently limit the depth of psychological insight. This approach stands in contrast to how psychiatrists work in clinical settings, where they analyze complete conversations rather than isolated statements to gain deep insights into patients' mental states. Drawing inspiration from this clinical practice, our research proposes a paradigm shift towards comprehensive dialogue-based analysis, leveraging the rich contextual information embedded in interactive conversations.

The hypothesis underlying this study is that depressive tendencies can be more accurately identified through the analysis of language phenomena in dialogues as well as the paragraphs rather than isolated sentences or keywords. This approach is supported by research indicating that dialogue analysis reveals emotional complexities often missed in isolated text, making dialogues a rich source for mental health assessments [3].

Dialogue, as a collaborative and dynamic form of communication, discloses complicated patterns of language use reflecting underlying emotional states. These patterns align closely with the nine diagnostic criteria in PHQ-9, as specific linguistic phenomena associated with depression manifest differently across various symptom categories., For instance, increased use of first-person pronouns [3] indicates excessive self-focus and feelings of worthlessness (PHQ-9 criterion 6), while negative sentiment words [4] and expressions of hopelessness [5] frequently signal thoughts about death or self-harm (PHQ-9 criterion 9). These linguistic markers become more evident when analyzed across entire conversations, as they often appear as recurring patterns rather than isolated instances. The use of specific linguistic structural features can be more reliably identified and interpreted within the context of a conversation or a paragraph, allowing for a more nuanced mapping between linguistic patterns and PHQ-9 criteria. Our approach aims to develop models for detecting depression in interactive conversations as well as in the context of paragraphs, comparing their effectiveness with models based on isolated sentences.

This study proposes depression types classification using dialogue analysis based on the PHQ-9 assessment. The study analyzed dialogues from 31 movies, providing naturally occurring conversational patterns across diverse situations. To examine linguistic patterns systematically, we used three data formats for comparison: Individual Quotes representing single-sentence analysis, Focus Sentences containing explicitly depression-related content, and Full Dialogues encompassing complete conversational exchanges. This multi-format approach allows us to compare the effectiveness of isolated

979-8-3315-2040-3/25/$31.00 ©2024 IEEE

sentence analysis against more context-rich dialogue analysis. The Long Short-Term Memory (LSTM) network is employed to categorize depressive symptoms according to the nine diagnostic criteria in the PHQ-9 questionnaire. This approach enables a comprehensive analysis of how different dialogue formats can contribute to the accurate identification and classification of depression indicators.

This paper presents a comprehensive study organized into five main sections. First, we detail our dataset collection and preparation process, providing insights into the data characteristics and preprocessing steps. The methodology section describes the methods used for depression type classification. The results section presents the classification results. A discussion is given in the discussion section and followed by a conclusion section.

II. DATASET

A. Dataset Collection

The goal of this study is to analyze authentic conversations between doctors and patients with depression symptoms. However, due to medical privacy constraints and ethical considerations, we adapted our approach to examine dialogues from 31 movies that portray characters dealing with depression. Our dataset includes diverse film genres, providing a comprehensive range of contexts where depressive symptoms manifest.

The analyzed dialogues were extracted from the following films:

Winnie the Pooh
- Cake
- Fight Club
- Silver Linings Playbook
- Scarlet Witch
- The Avenger End Game
- SpongeBob
- Garden State
- A Beautiful Mind
- The Perks of Being a Wallflower
- The Fault in Our Stars
- Joker
- Eternal Sunshine
- Lost in Translation
- After sun
- Blue Valentine
- The Royal Tenenbaums
- Inside Out
- Girl, Interrupted
- Requiem for a Dream
- Revolutionary Road
- The Virgin Suicides
- Sunshine
- Manchester by the Sea
- Her
- Taxi Driver
- Bright Star
- Melancholia
- The Pursuit of Happiness
- Frozen
- Fault in star

B. Dataset Labeling

The classification of movie dialogues was conducted using the Patient Health Questionnaire-9 (PHQ-9) framework, a standardized clinical tool widely used for assessing depression severity [6]. The PHQ-9 consists of nine questions that evaluate depressive symptoms experienced over a two-week period, with scoring ranging from 0 (not at all) to 3 (nearly every day), The nine PHQ-9 categories used for classification are:

Q1: Little interest or pleasure in doing things

Q2: Feeling down, depressed, or hopeless

Q3: Trouble falling or staying asleep, or sleeping too much

Q4: Poor appetite or overeating

Q5: Feeling bad about yourself or that you are a failure or have let yourself or your family down

Q6: Trouble concentrating on things, such as reading the newspaper or watching television

Q7: Moving or speaking so slowly that other people could have noticed, or being unusually fidgety and restless

Q8: Thoughts that you would be better off dead, or of hurting yourself

Q9: Thoughts of self-harm or that you would be better off dead

The annotation process employed three annotators with backgrounds in clinical psychology and computational linguistics, each trained specifically in applying the PHQ-9 framework for text classification. The annotators worked with detailed guidelines that included example dialogues and specific linguistic markers for each category. Final classification was determined through a majority voting system, requiring at least two annotators to agree on a label. In cases where all three annotators disagreed, they followed a structured resolution process involving individual rationale presentation and collaborative review of the PHQ-9 criteria. Dialogues that remained ambiguous after this review process were excluded from the dataset to maintain data quality.

The annotation process resulted in 151 labeled dialogues distributed across the PHQ-9 categories, as shown in Table I. The distribution of data across PHQ-9 categories shows significant imbalances in our dataset. Category 2 (feeling down, depressed, or hopeless) being the most prevalent at 36 instances. Category 6 (trouble concentrating) follows closely with 35 instances, while Category 4 (poor appetite or overeating) shows the lowest frequency among the populated categories with only 5 instances. Notably, Category 8, which pertains to physical manifestations of depression, contained no entries due to its non-verbal nature. These imbalances led us to implement data augmentation techniques in the subsequent phase to ensure more robust model training.

TABLE I. LIST OF DATA FOR EACH QUESTIONAIRE

Question Number	Amount of data
1	25
2	36
3	15
4	5
5	14
6	35
7	12
8	0
9	11

C. Focus Sentence Selection

The Focus Sentence dataset was created to complement the complete Dialogue dataset, reflecting how depression symptoms can be identified at different levels of conversation. While the Dialogue dataset captures the full context of emotional expressions, the Focus Sentence dataset consists of individual sentences that explicitly convey depressive symptoms according to PHQ-9 criteria. The annotation team selected these sentences based on their ability to indicate specific symptoms independently, without requiring surrounding context. This systematic approach was applied consistently across both original and augmented datasets, enabling analysis of how contextual information influences the accuracy of depression detection.

D. Augmentation Dataset

To address the significant data imbalance across PHQ-9 categories, we employed data augmentation techniques analogous to medical case study development, where variations of authentic cases are created while preserving core symptomatic features. Using ChatGPT, we generated five contextual variations for each original dialogue, maintaining the fundamental emotional and symptomatic characteristics while introducing diverse situational elements. A team of annotators rigorously reviewed all generated samples, selecting only those that maintained fidelity to the original PHQ-9 classification criteria. Table II illustrates how the augmentation process preserves the essential emotional content while introducing natural variations in expression.

This process expanded our dataset from 151 to 960 samples, achieving more balanced representation across categories as shown in Table III.

TABLE II. EAXAMPLES OF EXCERPT BEFORE AND AFTER AUGMENTATION

Pre-Augmentation Dialogue Set	Post-Augmentation Dialogue Set
Kanga: Good morning Roo: Good morning Eeyore: If it's actually a good morning, which I doubt.	Kanga: Good morning Roo: Good morning Eeyore: If it is indeed a good morning, which I find hard to believe.
	Kanga: Good morning Roo: Good morning Eeyore: Assuming it's a good morning, though I'm not convinced.

TABLE III. EAXAMPLES OF EXCERPT BEFORE AND AFTER AUGMENTATION

Question Number	Raw Data			Augmented data		
	FD	***FS***	***IS***	***FD***	***FS***	***IS***
1	25	25	63	120	120	309
2	36	36	82	120	120	288
3	15	15	36	120	120	307
4	5	5	13	120	120	308
5	14	14	34	120	120	280
6	35	35	85	120	120	296
7	12	12	25	120	120	253
8	0	0	0	0	0	0
9	11	11	23	120	120	277

III. METHODOLOGY

A. Data Preparation and Preprocessing

In the data preparation phase, we established clear criteria for dialogue selection based on PHQ-9 diagnostic guidelines. Conversations were evaluated against specific indicators of depressive symptoms, such as expressions of persistent sadness, loss of interest, or sleep disturbances. Dialogues that did not contain these clinical indicators were excluded to maintain dataset relevance. The selected conversations were limited to 2-4 sentences, balancing the need for sufficient context while managing computational complexity. The preprocessing pipeline was designed to optimize the text for model training while preserving essential emotional and symptomatic content. First, conversations were structured in JSON format to facilitate systematic data handling. The text underwent sequential transformations: converting to lowercase for standardization, tokenization to create processable word units, and removal of stop words (common words like 'the', is', at') that carry minimal diagnostic value. Lemmatization was then applied to reduce word variations to their base forms (e.g., 'feeling', 'felt' to 'feel'), enhancing pattern recognition while maintaining semantic meaning. Finally, an embedding layer translated these processed texts into numerical values, creating a mathematical representation that captures the linguistic features relevant to depression classification.

B. Classification Model

The Long Short-Term Memory (LSTM) model was selected for its proven effectiveness in processing sequential text data and capturing long-range dependencies essential for understanding emotional context in conversations. Unlike simpler neural networks, LSTM's ability to maintain and update its internal memory state makes it particularly suited for detecting subtle patterns of depressive symptoms that may span across multiple sentences. LSTMs are effective at learning sequential dependencies and addressing the vanishing gradient problem, allowing them to maintain long-range dependencies. This is significant for natural language processing, as understanding context and relationships across sequences enables the detection of patterns indicative of mental health conditions like depression.

The model architecture was carefully designed to optimize performance for our specific classification task. The embedding layer, with input_dim=10000

and output_dim=128, transforms words into dense vector representations, capturing semantic relationships between terms commonly associated with depressive symptoms. These dimensions were chosen based on our vocabulary size and the need to balance computational efficiency with representational capacity. The subsequent LSTM layers, with 64 and 32 units respectively, create a gradually refined representation of the text, while dropout rates of 0.5 help prevent overfitting by randomly deactivating half of the neurons during training.

To address the challenge of dataset imbalance across PHQ-9 categories, we implemented two complementary strategies. First, class weights were calculated using the formula:

$$Wi = \text{Total Samples} / (\text{Number of Classes} \times \text{Frequency of Class i}) \quad (1)$$

This approach ensures that the model pays proportionally more attention to underrepresented symptom categories during training. Additionally, we employed Stratified K-Fold Cross-Validation (k=5) to maintain consistent class distribution across training and validation sets, enabling more reliable model evaluation across all symptom categories.

Fig. 1 illustrates our complete experimental framework, highlighting four distinct datasets generated through our methodology: Raw Dataset, Focus Sentence Raw Dataset, Augmented Dataset, and Focus Sentence Augmented Dataset (shown in red box). Each dataset undergoes Stratified K-Fold Cross-Validation (k=5), ensuring consistent class distribution in both training and validation phases. This comprehensive validation approach, combined with our LSTM architecture, enables robust evaluation of how different data formats and augmentation strategies affect depression symptom detection accuracy.

Table IV from below summarizes the parameters associated with the LSTM Layers, which are crucial in processing sequential data and enhancing model learning.

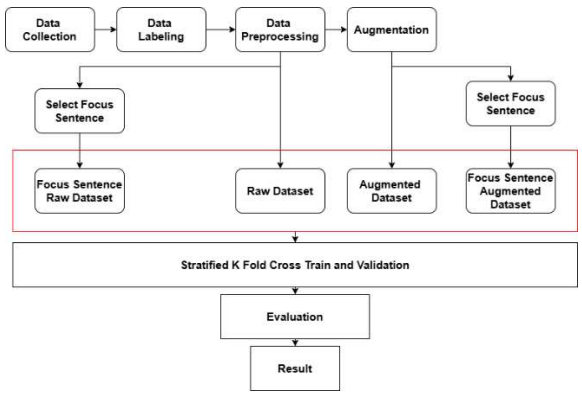

Fig. 1. Diagram of depression type classification.

TABLE IV. LONG SHORT TERM MEMORY PARAMETER

Layer	Parameter
Embedding Layer	input_dim=10000
	output_dim=128
	input_length=input_length
LSTM (First Layer)	units=64
	return_sequences=True
Dropout	rate=0.5
LSTM (Second Layer)	units=32

IV. RESULTS

A. *Performance of Raw Dataset*

In the raw dataset experiments, we observed significant instability in the model's performance metrics. Fig. 2 shows considerable fluctuation in Precision, Recall, and F1-score across different PHQ-9 categories, particularly in categories with limited samples such as Category 4 (poor appetite, which had only 5 instances in the raw dataset). This performance variability suggests that the imbalanced nature of the raw dataset impacts the model's ability to reliably classify depressive symptoms.

B. *Performance of Class-weights Method*

The application of class weights demonstrates an improvement over the raw dataset results, as illustrated in Figure 3. While this technique helps balance the model's attention across PHQ-9 categories, particularly benefiting categories with fewer samples, the high standard deviation indicates persistent instability in the model's predictions. This suggests that while class weights help address the immediate impact of data imbalance, they may not fully compensate for the fundamental limitation of having insufficient examples in certain categories.

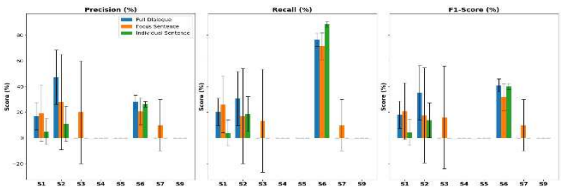

Fig. 2. Raw Dataset Performance Results.

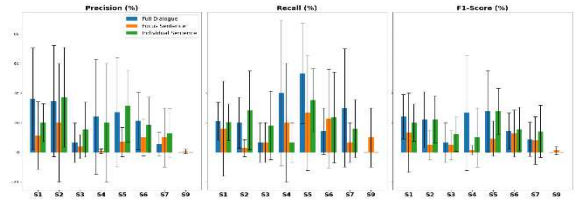

Fig. 3. Class weight Method Performance Results.

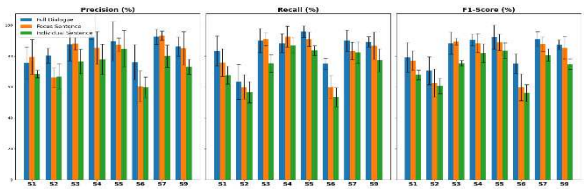

Fig. 4. Augmented Dataset Performance Results.

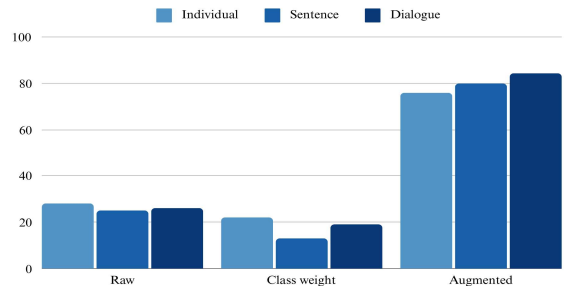

Fig. 5. Accuracy Result Across Three Dataset.

C. *Performance of Augmented Dataset*

Data augmentation demonstrates the most substantial improvement among all approaches, as shown in Fig. 4. The augmented dataset not only enhances overall model performance with improved Precision, Recall, and F1-Score, but also notably reduces standard deviation across all metrics. This increased stability suggests that the augmentation strategy successfully addresses both the data imbalance and sparsity issues that persisted in previous approaches. The reduced variability in model predictions indicates more reliable classification across all PHQ-9 categories, including those that previously suffered from limited samples. The results of the augmentation show an improvement in all cases.

D. *Comparative Performance Analysis*

The comparative analysis of the three data types reveals a clear advantage of contextual information in depression detection, as illustrated in Figure 5. The Full Dialogue approach achieved the highest accuracy at 84%, demonstrating the value of analyzing complete conversational exchanges. This performance notably surpasses the Focus Sentence approach at 80%, which relies on isolated but symptom-specific statements. The Individual Sentence method showed the lowest accuracy at 72%, confirming that analyzing sentences without their surrounding context limits the model's ability to detect subtle depressive indicators. These results emphasize how broader conversational context enhances the model's capability to identify and classify depressive symptoms more accurately.

V. DISCUSSION

The results after data augmentation demonstrate a significant enhancement in the Full Dialogue model's ability to understand depressive symptoms within their broader conversational context. This improved contextual understanding enables the model to capture subtle emotional nuances that are often missed when analyzing isolated sentences, leading to more accurate classification of depression symptoms according to PHQ-9 criteria.

A compelling example of this contextual advantage is illustrated in Table V, where the isolated statement "Honestly, I feel like ending it all" was misclassified as Case6 (trouble concentrating) when analyzed as a Focus Sentence. However, when the same statement appeared within its full conversational context - "You must be kidding. I'm serious. Honestly, I feel like ending it all" - the model correctly identified it as Case9 (thoughts of self-harm). This example demonstrates how surrounding dialogue provides crucial emotional escalation cues that help differentiate between general distress and more severe depressive symptoms.

A second example further illustrates how conversational context enhances classification accuracy. When analyzing the isolated statement 'I feel as if I've disappointed everyone' as a Focus Sentence, the model misclassified it as Case4 (poor appetite). However, when examining the complete dialogue exchange - 'I feel as if I've disappointed everyone. You haven't. You are more resilient than you realize.' - the model correctly identified it as Case6 (trouble concentrating), as shown in Table VI. This improvement stems from the model's ability to recognize how the supportive response helps characterize the nature of the speaker's negative self-assessment.

The effectiveness of our approach is further validated by quantitative improvements across multiple categories after data augmentation. For instance, the F1-Score for Case9 (thoughts of self-harm) improved significantly to 0.8744 ± 0.0318 in Full Dialogue, while Case1 (loss of interest) showed substantial enhancement from 0.6787 ± 0.0317 to 0.7915 ± 0.0965.

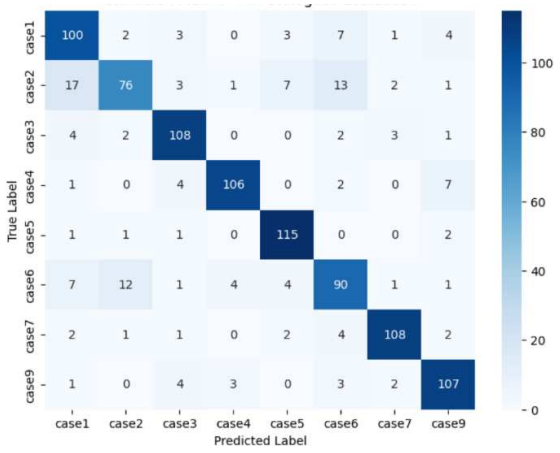

Fig. 6. Focus Sentence Confusion Matrix.

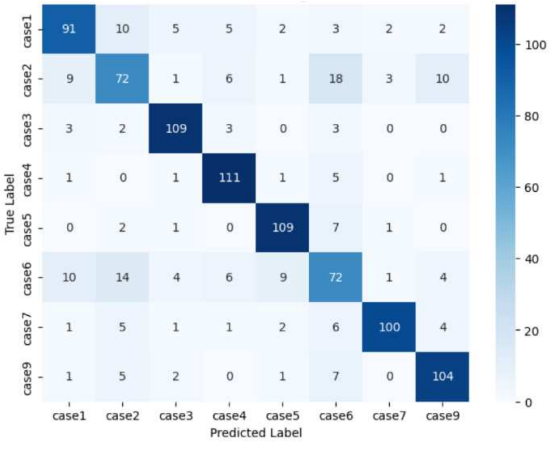

Fig. 7. Full Dialogue Confusion Matrix.

TABLE V. MISCLASSIFIED EXAMPLE BETWEEN FOCUS SENTENCE AND DIALOGUE

Data Type	Text	Results
Sentence	A: Honestly, I feel like ending it all.	Case6
Full Dialogue	A: You must be kidding. I'm serious. B: Honestly, I feel like ending it all.	Case9
Actual		Case 9

TABLE VI. MISCLASSIFIED EXAMPLE BETWEEN FOCUS SENTENCE AND DIALOGUE

Data Type	Text	Results
Sentence	A: I feel as if I've disappointed everyone.	Case4
Full Dialogue	A: I feel as if I've disappointed everyone. B: You haven't. B: You are more resilient than you realize.	Case6
Actual		Case 6

These consistent improvements across different symptom categories demonstrate how the combination of conversational context and augmented data enhances the model's capability to detect subtle variations in depressive expressions.

The augmented Full Dialogue model shows promising potential for real-world mental health applications, particularly in digital platforms where conversational context is available. Integration with chatbots could enable real-time assessment of depressive symptoms through natural user interactions, providing healthcare professionals with timely insights for intervention. For example, the model's ability to understand context in statements like 'Dad just walked out on us. Oh, he must not care about us anymore. That's really sad. I guess I should take control, right?' demonstrates its potential for detecting complex emotional patterns in actual conversations. The scalability of our augmented dataset approach suggests adaptability across various clinical and digital mental health settings.

However, important limitations exist in our current approach. The use of scripted dialogues may not fully capture the spontaneity and variability of real-world conversations, potentially limiting the model's effectiveness with informal or unstructured language. Furthermore, while data augmentation improved overall performance, some categories like Case6 showed persistent challenges (F1-Score: 0.7511 ± 0.0647), indicating that certain depressive expressions remain difficult to classify accurately.

VI. CONCLUSION

This research demonstrates the superior effectiveness of Full Dialogue analysis in detecting depressive symptoms compared to traditional single-sentence approaches. By leveraging complete conversational contexts, our model achieved an 84% accuracy rate, significantly outperforming Focus Sentence (80%) and Individual Sentence (72%) methods. The integration of data augmentation techniques further enhanced the model's ability to capture subtle emotional expressions across different PHQ-9 categories, particularly in cases where context plays a crucial role in symptom interpretation. The practical implications of these findings extend to real-world mental health applications, where early detection and intervention are critical. Our approach shows particular promise for integration with digital health platforms and chatbots, where natural conversation flows can be analyzed for depressive indicators in real-time. However, the current reliance on scripted dialogues presents limitations in capturing the full complexity of real-world conversations. Future research should focus on incorporating naturalistic data and exploring advanced transformer-based models to enhance the system's adaptability to informal language patterns and improve overall classification accuracy across all PHQ-9 categories.

ACKNOWLEDGMENT

This study is financially supported by Faculty of Informatics, Burapha University.

REFERENCES

[1] R. D. Sousa, A. R. Henriques, J. Caldas de Almeida, H. Canhão, and A. M. Rodrigues, "Unraveling Depressive Symptomatology and Risk Factors in a Changing World," *Int J Environ Res Public Health*, vol. 20, no. 16, p. 6575, Aug. 2023, doi: 10.3390/ijerph20166575.

[2] T. Zhang, K. Yang, H. Alhuzali, B. Liu, and S. Ananiadou, "PHQ-aware depressive symptoms identification with similarity contrastive learning on social media," *Information Processing & Management*, vol. 60, no. 5, p. 103417, Sep. 2023, doi: 10.1016/j.ipm.2023.103417.

[3] L. A. Cariola *et al.*, "Language use in depressed and non-depressed mothers and their adolescent offspring," *Journal of affective disorders*, vol. 366, p. 290, Aug. 2024, doi: 10.1016/j.jad.2024.08.131.

[4] R. N. Trifu, B. Nemeş, D. C. Herta, C. Bodea-Hategan, D. A. Talaş, and H. Coman, "Linguistic markers for major depressive disorder: a cross-sectional study using an automated procedure," *Front. Psychol.*, vol. 15, Mar. 2024, doi: 10.3389/fpsyg.2024.1355734.

[5] N. H. Yahya and H. Abdul Rahim, "Linguistic markers of depression: Insights from english-language tweets before and during the COVID-19 pandemic," *Language and Health*, vol. 1, no. 2, pp. 36–50, Dec. 2023, doi: 10.1016/j.laheal.2023.10.001.

[6] K. K, S. Rl, and W. Jb, "The PHQ-9: validity of a brief depression severity measure," *Journal of general internal medicine*, vol. 16, no. 9, Sep. 2001, doi: 10.1046/j.1525-1497.2001.016009606.x.