Neuro-Symbolic Classifier for Depression Analysis

Ashwin Umadi, Prajakta Kini, Theodora Chaspari
Department of Computer Science
University of Colorado Boulder
{ashwin.umadi,prajakta.kini,theodora.chaspari}@colorado.edu

Abstract

Diagnosing depression has become increasingly crucial as more patients seek medical assistance for mental health concerns. The DAIC dataset is commonly used for depression classification, where an interviewer interacts with an interviewee to assess their mental state. Previous studies on this dataset have predominantly focused on analyzing speech signals, yielding moderate performance. In this work, we introduce a novel approach that leverages the linguistic content of participants' spoken language to enhance depression classification. We design a rule-based system to classify depression and evaluate its effectiveness against other baseline models built on language features. Furthermore, we incorporate the Rule-Based model into the pre-trained Mental-RoBERTa model and observe a 1.81% improvement in performance due to the integration of this symbolic framework.

1 Introduction

A rule-based system is a type of artificial intelligence (AI) system that applies predefined rules to data to derive conclusions or make decisions. It is a computer system in which domain-specific knowledge is represented in the form of rules and general-purpose reasoning is used to solve problems in the domain (Wikipedia contributors, 2024). There are different kinds of rule-based systems, 1) production-based rules and 2) Logic-Based rules. The first one is built using if-else rules, for example - If a patient has a fever and soar throat, then consider the patient has a throat infection. In the second one, rules are written in the form of clauses, for example - If a student is enrolled in CSCI 7000 class **AND** the student gets more than 95% in all the assignments, then consider the student a Pass. Comparing the two methods, product-based rules are like automated systems that apply direct actions on specific conditions. In contrast, logic-based rules often enable more complex inference and are a part of strong reasoning systems.

Medical applications frequently involve diagnosing diseases from images or various signal formats, such as EEG, ECG, or speech signals. With recent advancements in machine learning, researchers globally are tackling these challenges using cutting-edge transformers, generative models, and large language models (LLMs). For instance, (Yang et al., 2024) employs advanced adversarial techniques to classify whether a participant is depressed through sophisticated neural networks. This introduces significant challenges when working within low-computation frameworks. For example, implementing real-time fatigue detection on a smartwatch limits the ability to use models that require substantial computational resources on such a compact platform. Moreover, even if computational limitations are addressed, these complex models introduce additional latency in generating outputs.

Implementing rule-based systems in these scenarios can effectively address the challenges outlined above. For example, (Pota et al., 2017) applies fuzzy logic rules for medical classification, outperforming complex neural network models in cancer classification, Haberman's survival, and diabetes detection. This success inspires our approach to using a rule-based system for the depression classification of participants in the DAIC-WOZ dataset. In this paper, we present a novel rule-based technique for detecting depression.

We approach the task of classification of depression in a manner similar to sentiment classification, where certain strongly indicative words play a key role in the detection of depression. These words are often adjectives, so we obtain their contextual embeddings within sentences using the MentalBERT model (Ji et al., 2021) and compare these embeddings to those of the same adjectives from our SentenceBank. Our SentenceBank consists of 7500 sentences sourced from Reddit posts, all classified as cases of depression. To determine if a participant is depressed, we apply two specific rules.

- Is the adjective word used by the participant indicative of depression?
- Is the interview indicative of depression?

To identify whether a word indicates depression (Rule 1), we calculate the cosine similarity between the two embeddings. To determine whether the general interview suggests depression, we assess the number of adjectives flagged as indicative of depression and then make a final decision(Rule 2).

Furthermore, we develop an advanced neuro-symbolic model, MentalRoBERTa++, which combines the interpretability of a rule-based system with the robust capabilities of the pre-trained transformer model, MentalRoBERTa. MentalRoBERTa is an extension of the RoBERTa model, fine-tuned specifically for mental health applications in medical contexts. To seamlessly integrate the rule-based system with the MentalRoBERTa model, we introduce a novel loss function incorporating a penalty factor, λ. This factor ensures an optimal balance between the contributions of the rule-based system and the pre-trained MentalRoBERTa model.

Experimental results on the DAIC-WOZ dataset demonstrate that Mental RoBERTa++ achieves higher balanced accuracy compared to the standalone Mental RoBERTa or rule-based systems. This improvement highlights the benefits of combining symbolic reasoning with neural representations for the task of depression detection. A comprehensive graphical analysis is presented in the Experiments and Results section. Our newly proposed Mental-RoBERTa++ model outperforms all other models, achieving a balanced accuracy of 72.41%, which is 1.81% higher than the best-performing baseline model.

2 Related Work

Below are some related work in the domain that we are trying to solve our challenge.

(Vacareanu et al., 2024) focuses on relation classification by combining rule-based methods with deep learning techniques to leverage the explainability of rules and the generalization power of neural networks. This approach uses syntactic rules derived from dependency parse trees and a semantic rule matching (SRM) component trained on a vast dataset of rule-sentence pairs. The proposed method uses a two-step approach. First, it attempts to match predefined rules with the input text in a strict, binary fashion. If no match is found, the system utilizes a neural semantic rule matching (SRM) component that semantically aligns rules with text. The approach outlined in this paper shares a relevant structure with our method in depression classification, as both emphasize balancing interpretability and generalization. Like their method, which combines strict rule matching with a semantic rule matching component to handle nuanced relations, our approach uses rule-based checks to detect depression by identifying strongly indicative words, enhanced by MentalBERT's contextual embeddings. This hybrid setup is relevant for us because their semantic alignment strategies for rule-sentence pairs offer insights that could improve our classification accuracy, especially in recognizing depression expressed subtly across different contexts.

The work of (Hu et al., 2016) introduces a method to enhance the interpretability of deep neural networks by integrating structured logic rules through knowledge distillation. By training a "student" network to mimic the predictions of a "teacher" network embedded with logic-based constraints, this approach effectively transfers structured rule knowledge into the network's parameters, making predictions possible without rule evaluation at test time. This method has shown success in tasks like sentiment analysis, which is relevant to our work in depression classification, as both rely on identifying specific, strongly indicative words within the text. While their focus is on embedding rules into a neural network for prediction, our approach maintains an independent rule-based system, using cosine similarity between contextual embeddings from MentalBERT to evaluate key adjectives as indicators of depression.

3 Methodology

3.1 The Neural Part:

Embeddings of Interviewee: We utilize MentalBERT as our neural model to extract embeddings for the given context. Specifically, our approach follows these steps.

Identifying Adjectives: We utilize Python's Spacy module to identify all adjectives in the conversation between the interviewer and interviewee. Although we also tested the NLTK module, but Spacy provides more accurate POS classification for the sentences.

Embeddings of Adjectives:

After identifying the adjective words in a sentence, we pass the sentence—along with preceding and following conversational context—to MentalBERT (Ji et al., 2021) to obtain embeddings for the adjective. We selected MentalBERT because it is well-regarded for understanding depressive language. The window size for the previous and subsequent conversations provided to MentalBERT is set as a hyperparameter, with a value of 10 conversations each. MentalBERT has a maximum token limit of 512, truncating any tokens that exceed this limit. Ultimately, MentalBERT provides the contextual embeddings of the targeted adjective.

Embeddings from SentenceBank:

SentenceBank:

We have a dataset downloaded from Kaggle consisting of 7500 sentences, all classified as depression-related. We refer to this dataset as our SentenceBank.

Finding Adjectives in SentenceBank: We use the same Spacy module to identify all adjective words within our SentenceBank.

Embeddings of Adjectives: After identifying adjectives in each of the 7500 sentences, we obtain their contextual embeddings using MentalBERT, as previously discussed. For adjectives that appear multiple times in the SentenceBank, we calculate the average of all embeddings generated for that adjective and use this as the final embedding value from our SentenceBank.

$$E(w) = \frac{1}{n} \sum_{i=1}^{n} E(w_i)$$

where $E(w_i)$ is the embedding from MentalBERT for adjective w_i.

Integration of two embeddings

Once we obtain the embeddings from Sections 3.1 and 3.2, we measure their similarity using cosine similarity. A higher score indicates a greater similarity between the two words, taking their respective contexts into account. Figure 1 illustrates the complete process of the proposed methodology up to the calculation of cosine similarity.

3.2 The Symbolic Part:

Rule 1: After calculating the similarity score between the two embeddings of the same adjective, we establish a rule to determine if the word indicates a depressed class. Specifically, we check whether the cosine similarity score exceeds a certain threshold (t1) to classify the word as indicative of depression. The threshold value is determined based on experimental results, and we are still working to identify the optimal value.

Rule 2: Once each adjective is classified as either depressed or not depressed, we calculate the ratio of adjectives classified as depressed to the total adjectives in the conversation. If this ratio exceeds a specific threshold (t2), we classify the interviewee as depressed. We are still determining the optimal value for this threshold.

All the findings have been discussed in the Experimentation section.

3.3 Mental-RoBERTa++

Mental RoBERTa++ combines the neural capabilities of Mental RoBERTa with the symbolic reasoning of the adjective-based rule system to create a neuro-symbolic framework. This integration leverages the contextual understanding of Mental RoBERTa and the rule-based system's explicit reasoning, enabling robust performance and interoperability.

3.3.1 Chunking Process

To process the interview transcripts effectively, we split the question-answer pairs into chunks, each containing up to 300 words. We allowed up to 80-word overlap between chunks to ensure smooth transitions. Additionally, we leave extra space for RoBERTa's subword tokenization using Byte-Pair Encoding during chunk preparation. We also ensure that each question-answer pair remains intact and is not split across chunks.

3.3.2 Fine-Tuning Mental RoBERTa++ using Combined Loss

Each chunk of text is passed through Mental RoBERTa, where the pooled [CLS] token embeddings are extracted for classification. Mental RoBERTa, a model pre-trained on mental health-related datasets, is then fine-tuned on our specific dataset for depression classification. The fine-tuning process involves optimizing the neural model using a combined loss function that integrates both the neural model's predictions and the symbolic reasoning based on predefined rules.

The neuro-symbolic framework introduces a hybrid loss function that integrates the predictions from both the neural and symbolic components.

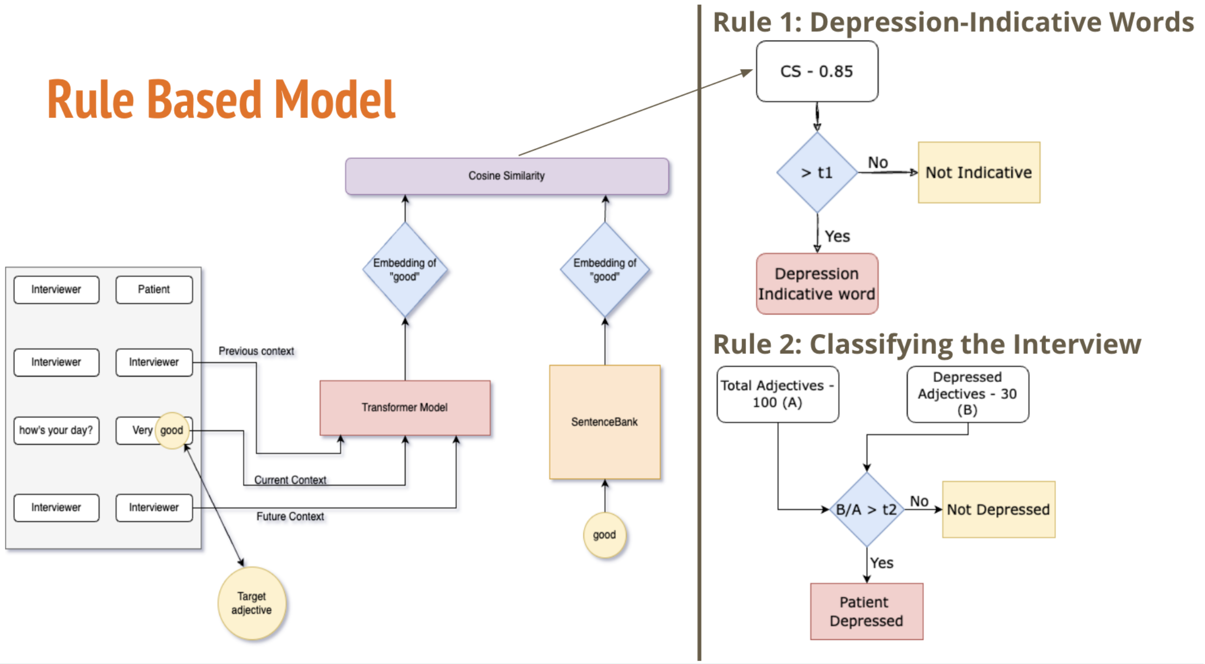

Figure 1: Proposed methodology to get the cosine similarity of the adjective present in interview process and SentenceBank. In the next stage, this cosine similarity is used to implement the two rules to predict if the patient is depressed.

The total loss is defined as:

$$Total\ Loss = Neural\ Loss + \lambda \cdot Rule\ Penalty$$

where:

- **Neural Loss:** The cross-entropy loss between the neural model's predictions and the ground-truth labels.
- **Rule Penalty:** A penalty term for disagreements between the neural and rule-based predictions, scaled by λ.
- **λ:** A weighting factor that controls the influence of the rule-based penalty. Higher values of **λ** force greater agreement with the rules.

3.3.3 Integration Process

1. The neural model generates predictions for each chunk based on its contextual understanding.
2. The symbolic model evaluates adjectives using cosine similarity and applies the two predefined rules (Section 3.2).
3. The hybrid loss function encourages the neural model to align with symbolic reasoning and leverage its contextual embeddings, balancing robustness and interpretability.

Hyperparameters such as thresholds (t1 and t2) and λ are optimized based on the validation dataset to maximize balanced accuracy.

4 Experiments & Results

4.1 Dataset:

All experiments are conducted on the DAIC-WOZ dataset, consisting of 189 participants, 48 of whom are classified as depressed. The conversations between the interviewer and participant are used without any pre-processing, and results are reported using the Balanced Accuracy metric. Since this is the first study to classify depression based on participants' spoken language, we also generate baseline figures to evaluate the performance of our proposed rule-based method.

4.2 Baseline models:

We conduct baseline experiments using XGBoost, FNN, BERT, RoBERTa, Mental-RoBERTa model, and the Llama 3.2-1B model with in-context learning(passing the entire question-answer pair as input in the prompt). For XGBoost and FNN, input sentences are tokenized with BERT encodings before being fed into the training models.

For RoBERTa and Mental RoBERTa, we employ the chunking process explained in Section 3.3.1

and then fine-tune these pre-trained models on our dataset specifically for depression classification.

The Balanced Accuracy performance of each model is presented in Table 1, with Balanced Accuracy selected as the primary metric in line with (Yang et al., 2024). The best-performing model for us anyway is the Mental-RoBERTa++ model giving an accuracy of 72.41%.

Model	Balanced Accuracy (%)
XG-Boost	50.50
FNN	49.98
BERT	56.53
RoBERTa	69.50
Mental-RoBERTa	70.60
Llama 3.2	55.67
Rule Based Model	55.50
Mental-RoBERTa++	**72.41**

Table 1: Balanced Accuracy of Depression Classification across several models including traditional Machine Learning models, Deep Learning models and Neuro symbolic models

4.3 Discussion:

4.3.1 Rule Based System

The first set of results we examine pertains to the simple rule-based systems that operate based on the two rules outlined earlier. The words used here while getting the embeddings are only adjectives present in the dataset. Table 2 summarizes seven distinct combinations of values for thresholds t_1 and t_2, with values ranging from a minimum of 0.15 to a maximum of 0.85. These combinations were systematically evaluated to identify the configuration yielding the best performance. Notably, the optimal performance was achieved when t_1 was set to 0.75 and t_2 was set to 0.25, resulting in a Balanced Accuracy of 55.5%. This finding highlights the sensitivity of the system to specific threshold values and underscores the importance of selecting appropriate parameters for optimal outcomes.

Table 3 evaluates the rule-based system using POS tags other than adjectives, such as verbs, pronouns, and nouns. These POS tags were assessed using the same symbolic framework, applying cosine similarity to contextual embeddings. The balanced accuracy ranged from 45.19% to 52.93%, with verbs consistently outperforming pronouns and nouns across all thresholds.

Despite capturing specific linguistic patterns, these POS features were less indicative of depression compared to adjectives. This finding underscores the unique importance of adjectives in sentiment-driven tasks such as depression detection. These results suggest that while other POS categories may contribute marginally to rule-based reasoning, adjectives remain the most reliable indicators for symbolic reasoning in this context.

Combination	t1	t2	BAcc (%)
0	0.65	0.15	51.1
1	0.7	0.2	52.5
2	0.7	0.3	52.5
3	0.75	0.15	46.25
4	0.75	0.25	**55.5**
5	0.8	0.75	50.0
6	0.85	0.75	50.0

Table 2: Combinations of pairs t_1 and t_2 with their Balanced Accuracy. Rules t_1 and t_2 use only adjective words and its embeddings.

t1	t2	Verb	Pron	Noun
0.75	0.15	52.83	47.32	45.19
0.75	0.05	52.83	47.32	45.19
0.75	0.95	52.83	47.32	45.19
0.85	0.90	51.69	52.93	48.61
0.85	0.95	51.69	52.93	48.61
0.85	0.98	51.69	52.93	48.61

Table 3: Balanced accuracy of POS Rule-Based Models using Mental RoBERTa's contextualized embeddings.

4.3.2 Mental RoBERTa++

Mental RoBERTa++ combines the strengths of rule-based reasoning and pretrained transformer model to enhance performance. As shown in Table 4, the neuro-symbolic framework consistently outperforms the standalone Mental RoBERTa model. This hybrid framework incorporates rule-based penalties using depression-indicative adjectives, enabling it to refine predictions and achieve superior balanced accuracy (BAcc).

Table 4 demonstrates the significant improvements achieved by Mental RoBERTa++, a neuro-symbolic framework that integrates rule-based reasoning with pretrained transformer model. The standalone Mental RoBERTa model served as the baseline, achieving a balanced accuracy (BAcc) of 70.60%. Incorporating symbolic reasoning, with thresholds t_1 = 0.75 and t_2 = 0.50 and setting λ = 0.1, resulted in a slight improvement to 70.80%.

Model	t1	t2	λ	BAcc(%)
Mental RoBERTa	-	-	-	70.60
Mental RoBERTa++	0.75	0.50	0.1	70.80
Mental RoBERTa++	0.75	0.25	0.1	71.63
Mental RoBERTa++	0.75	0.25	0.3	**72.41**

Table 4: Comparison of Neural Mental RoBERTa model with Neuro-Symbolic Models. It uses ADJ Rule-based model for imposing rule penalty.

Fine-tuning these thresholds to $t_1 = 0.75$, $t_2 = 0.25$, while keeping $\lambda = 0.1$, further boosted the BAcc to 71.63%. Finally, increasing the influence of the rule-based system with $\lambda = 0.3$ achieved the highest performance, with a BAcc of 72.41%.

These results highlight the effectiveness of Mental RoBERTa++ in combining the strengths of symbolic reasoning and neural modeling. The weight λ in the hybrid loss function plays an important role in balancing the contributions of these two components. Increasing λ encourages the model to align more closely with the rule-based system, enhancing interpretability and boosting performance. However, excessively high values of λ can overly constrain the neural model, reducing its ability to generalize. By striking the right balance, the neuro-symbolic framework of Mental RoBERTa++ delivers superior performance while maintaining interpretability, setting a new standard for depression classification tasks.

5 Limitations

In our symbolic design, we rely on the words from the SentenceBank, a collection of 7500 sentences derived from Reddit posts, all of which are associated with depression-related content. We conducted a comprehensive analysis to determine the number of nouns, verbs, and adjectives from the SentenceBank that appear in patient transcripts. Table 5 provides a detailed summary of this analysis.

The column labeled **SentenceBank** lists the total counts of nouns, verbs, and adjectives identified in the Reddit posts. The **Transcripts** column captures the corresponding counts of these parts of speech (POS) in the interview transcripts between the interviewer and the patient. The **Overlap Count** column indicates how many nouns, verbs, and adjectives found in the transcripts are also present in the SentenceBank. Finally, the **Overlap %** column calculates the percentage of POS in the transcripts that overlap with those in the SentenceBank. A higher overlap percentage indicates greater reliability of the results.

POS	SentenceBank	Transcripts	Overlap Count	Overlap %
Nouns	8037	3982	2022	50.79
Verbs	5480	2501	1653	66.06
Adjective	3143	1572	1070	68.06

Table 5: Exploratory Data Analysis of Nouns, Adjectives and Verbs in SentenceBank and Transcript datasets.

From the analysis presented in Table 5, it is evident that adjective words, which were utilized in our model design, show an overlap percentage of only 68.08%. This implies that 31.92% of adjectives in the transcripts did not contribute to predicting whether a patient is depressed. In future work, we aim to incorporate additional datasets to increase this overlap percentage, ideally approaching 100%.

6 Conclusion and Future Scope:

We utilize a range of models, including traditional machine learning models, neural networks, deep learning models, and neuro-symbolic approaches, to classify whether a patient is depressed based on transcripts from the DAIC-WOZ dataset. Among these, our proposed neuro-symbolic model, Mental RoBERTa++, achieves state-of-the-art performance with a balanced accuracy of 72.41%. This demonstrates the effectiveness of combining neural networks with symbolic rules, resulting in more interpretable predictions. Specifically, we incorporate two simple rules that rely on the presence of adjectives in the transcripts as indicators of depression. These rules are integrated with the pre-trained transformer model, Mental RoBERTa, by modifying its loss function to include a penalty term. This term optimally balances the contributions of the transformer and the rule-based system.

The DAIC-WOZ dataset is predominantly male, with over 60% of the transcripts belonging to male participants, and more than 50% representing White and African American racial groups. Moving forward, we aim to implement bias mitigation techniques to explore whether these models' performance can be further enhanced.

Acknowledgments

This work is being guided by Prof. Theodora Chaspari in the Computer Science department of University of Colorado Boulder since June 2024.

References

Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. 2016. Harnessing deep neural networks with logic rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2410–2420, Berlin, Germany. Association for Computational Linguistics.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. Mentalbert: Publicly available pretrained language models for mental healthcare. *Preprint*, arXiv:2110.15621.

Marco Pota, Massimo Esposito, and Giuseppe De Pietro. 2017. Designing rule-based fuzzy systems for classification in medicine. *Knowledge-Based Systems*, 124:105–132.

Robert Vacareanu, Fahmida Alam, Md Asiful Islam, Haris Riaz, and Mihai Surdeanu. 2024. Best of both worlds: A pliable and generalizable neuro-symbolic approach for relation classification. *Preprint*, arXiv:2403.03305.

Wikipedia contributors. 2024. Rule-based system. [Accessed: 2024-02-28].

M. Yang, A. A. El-Attar, and T. Chaspari. 2024. Deconstructing demographic bias in speech-based machine learning models for digital health. *Frontiers in Digital Health*, 6:1351637.