Ratnesh Kumar Sharma, 2025, 13:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

# Automated Depression Assessment Using RLHF and RLAIF Instructions in Machine Learning

**Ratnesh Kumar Sharma, Prof. (Dr.) Satya Singh**
Department of Computer Science & Applications, M.G. Kashi Vidyapith
Varanasi (U.P.)

**Abstract - In the domain of machine learning, a useful application of Reinforcement Learning from Human Feedback (RLHF) and Reinforcement Learning from AI Feedback (RLAIF) lies in the automated evaluation of depression in patients of all age groups. By leveraging natural language processing (NLP) techniques and machine learning algorithms, this paper aims to develop effective models for detecting depression through language patterns and assessing the severity of depression. The study demonstrates that RLHF significantly enhances model performance through the incorporation of expert feedback, while RLAIF offers a scalable solution that utilizes AI-driven insights. The findings suggest that both methodologies hold promise for improving automated mental health assessment tools, ultimately contributing to more effective diagnoses and follow up counselling/ treatment.**

**Keywords-Automated Depression Evaluation, Reinforcement Learning, Human Feedback, AI Feedback, Natural Language Processing, Machine Learning.**

## I. INTRODUCTION

Depression is a complex and prevalent mental health disorder that affects millions of individuals worldwide. According to the World Health Organization (WHO), over 264 million people are affected by depression, leading to significant impairments in daily functioning and quality of life. Early detection and intervention are crucial for improving outcomes, yet traditional methods of diagnosis often rely heavily on self-reported assessments and clinical evaluations. These approaches can be subjective, susceptible to biases, and may result in underdiagnosis or misdiagnosis.

Recent advancements in artificial intelligence (AI) and machine learning have opened new avenues for the automated evaluation of mental health disorders. Natural Language Processing (NLP), a subfield of AI, provides powerful tools for analyzing text data, allowing for the detection of linguistic patterns associated with depressive symptoms. By analyzing unstructured data from various sources—such as social media, online forums, and clinical notes—automated systems can offer insights into individuals' mental states and detect signs of depression more efficiently and objectively than traditional methods.

To enhance the accuracy and reliability of automated depression evaluation, this study explores the integration of two emerging methodologies in reinforcement learning: Reinforcement Learning from Human Feedback (RLHF) and Reinforcement Learning from AI Feedback (RLAIF). RLHF leverages expert human insights to improve model performance, incorporating nuanced understanding of depressive language and emotional

tone. In contrast, RLAIF utilizes feedback generated by AI systems, allowing for scalable improvements without the need for constant human supervision.

## II. LITERATURE REVIEW

This paper aims to investigate the effectiveness of RLHF and RLAIF in developing robust models for the automated detection and severity assessment of depression. By comparing these methodologies, the research seeks to identify which approach yields superior results in accurately evaluating depressive symptoms and understanding the complexities of human emotions expressed in text. Ultimately, the findings aim to contribute to the growing body of knowledge in the applications of Machine learning in area of mental health, highlighting the potential for innovative technological solutions to enhance mental health diagnostics and interventions.

**Understanding Depression and Its Assessment**

Depression is a multifaceted mental health disorder characterized by persistent feelings of sadness, hopelessness, and a lack of interest in previously enjoyed activities. It can manifest in various ways, making diagnosis challenging. Traditional assessment methods, such as clinical interviews and standardized questionnaires (e.g., the Patient Health Questionnaire-9, PHQ-9), rely heavily on self-reported data, which can be influenced by various factors, including stigma and personal biases (Kroenke et al., 2001). This reliance on subjective measures has highlighted the need for more objective, automated methods of assessment that can provide timely and accurate evaluations of depressive symptoms.

**Machine Learning in Mental Health**

The integration of machine learning techniques in mental health has shown significant promise in automating the detection of various mental health disorders, including depression. Machine learning algorithms, particularly those utilizing NLP, can analyze large volumes of unstructured text data to identify patterns and indicators of depression. For instance, studies have demonstrated the effectiveness of sentiment analysis and language modeling in detecting depressive language on social media platforms (Coppersmith et al., 2015; Reece & Danforth, 2017). These approaches have the potential to augment traditional diagnostic methods by providing real-time insights into individuals' mental health status.

**Natural Language Processing and Depression Detection**

Natural Language Processing plays a crucial role in the automated assessment of depression. Various studies have utilized NLP techniques to analyze text data for depressive indicators, employing models such as bag-of-words, term frequency-inverse document frequency (TF- IDF), and more recently, deep learning models like recurrent neural networks (RNNs) and transformer-based architectures (e.g., BERT) (Devlin et al., 2019). These models can capture the subtle nuances of language that may signify depression, such as changes in emotional tone, word choice, and sentence structure. However, many existing models have been limited by the quality and quantity of labeled data available for training.

**Reinforcement Learning from Human Feedback (RLHF)**

Reinforcement Learning from Human Feedback is a relatively new approach that enhances machine learning models by incorporating human expertise into the training process. This methodology allows models to learn from feedback provided by human evaluators, thereby improving their understanding of complex tasks (Stiennon et al., 2020). In the context of depression detection, RLHF can be utilized to refine model predictions based on expert insights into linguistic cues and emotional expressions associated with depressive symptoms. Preliminary studies indicate that RLHF can significantly boost the performance of models in tasks requiring nuanced understanding (Ouyang et al., 2022).

**Reinforcement Learning from AI Feedback (RLAIF)**

In contrast, Reinforcement Learning from AI Feedback employs a teacher-student model where a more advanced "teacher" model generates feedback for a "student" model, enabling it to learn from its mistakes without constant human oversight. This method has gained traction in scenarios where

human labeling is impractical or where large datasets are required for effective training (Nisan et al., 2021). RLAIF's ability to provide scalable feedback makes it a valuable tool in the automated assessment of mental health, especially in contexts with limited access to mental health professionals.

**Comparative Studies and Gaps in Research**

While numerous studies have explored the use of machine learning and NLP in mental health diagnostics, research comparing the effectiveness of RLHF and RLAIF in this context is limited. Most studies tend to focus on one approach without exploring how these methodologies can complement each other to enhance model performance. This gap in the literature highlights the need for further investigation into the strengths and limitations of RLHF and RLAIF, particularly in the automated evaluation of depression.

## III. METHODOLOGY

The research design, data sources, model architectures, and evaluation metrics used in the study to explore the effectiveness of Reinforcement Learning from Human Feedback (RLHF) and Reinforcement Learning from AI Feedback (RLAIF) in automated depression evaluation are elaborated below:

**Research Design**

The research employs a comparative experimental design, where both RLHF and RLAIF methodologies are applied to train machine learning models for automated depression detection and severity assessment. The study aims to analyze the models' performance based on accuracy, precision, recall, F1 score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

**Data Sources**

Data for this study were gathered from multiple sources to ensure a diverse and representative dataset for training and evaluation:

- Social Media Data: Publicly available posts from platforms like Twitter and Reddit were collected using web scraping techniques. Posts containing keywords related to depression (e.g., "sad," "depressed," "hopeless") were filtered for analysis.
- Clinical Data: Transcripts from clinical interviews and therapy sessions were obtained from anonymized datasets, ensuring compliance with ethical standards and privacy regulations.
- Self-Reported Questionnaires: Data from validated self-report instruments, such as the PHQ- 9, were utilized to obtain standardized measures of depression severity, providing a reference for model predictions.

**Data Preprocessing**

The collected data underwent several preprocessing steps:

- Text Normalization: Raw text data were cleaned to remove noise, including URLs, special characters, and stop words. Text was converted to lowercase to maintain uniformity.
- Tokenization: The processed text was tokenized into individual words or phrases to prepare for model input.
- Vectorization: Tokenized text was transformed into numerical vectors using techniques like Word2Vec or TF-IDF to enable machine learning algorithms to process the data.

**Model Architecture**

**Baseline Model**

A baseline model was created using traditional supervised learning algorithms, including:

- Random Forest: An ensemble learning method that constructs multiple decision trees for improved prediction accuracy.
- Support Vector Machines (SVM): A classification technique that finds the optimal hyperplane for separating data points of different classes. baseline model was trained on the preprocessed dataset to establish a performance benchmark.

**RLHF Model**

The RLHF model was developed using a transformer-based architecture (e.g., BERT). The training process included:

- Supervised Pretraining: The model was initially trained on the labeled dataset to learn the linguistic features associated with depression.
- Human Feedback Integration: After initial training, the model underwent fine-tuning based on feedback from mental health professionals who reviewed and provided insights on the model's predictions. This iterative process involved identifying misclassifications and guiding the model towards more accurate interpretations of depressive language.

**RLAIF Model**

The RLAIF model utilized a teacher-student framework:

- Teacher Model Training: A robust teacher model, trained on a larger dataset, generated predictions and associated feedback on the data samples.
- Student Model Training: The student model was trained using the feedback from the
- teacher model, reinforcing learning based on successful predictions and correcting errors. This process allowed the student model to refine its understanding of depressive indicators based on AI-generated insights.

**Evaluation Metrics**

The models were evaluated using the following metrics:

- **Accuracy:** The proportion of correctly predicted instances over the total instances.
- Precision: The proportion of true positive predictions among all positive predictions, indicating the model's ability to minimize false positives.
- **Recall:** The proportion of true positive predictions among all actual positive instances, reflecting the model's ability to identify true cases of depression.
- **F1 Score:** The harmonic means of precision and recall, providing a balanced measure of the model's performance.
- **AUC-ROC:** A performance measurement for classification problems at various threshold settings, illustrating the model's ability to distinguish between depressed and non-depressed individuals.

## IV. EXPERIMENTAL SETUP

The experimental setup outlines the procedures and configurations used to train and evaluate the machine learning models designed for automated depression detection. This section details the infrastructure, data handling, training procedures, and evaluation processes used in the study.

**Infrastructure and Tools**

The experiments were conducted on a high-performance computing platform equipped with GPUs to facilitate efficient model training. The following tools and frameworks were utilized:

- Programming Language: Python 3.8 was used for implementing machine learning algorithms and data preprocessing.
- Machine Learning Frameworks:
- TensorFlow and Keras for building and training neural network models, including the RLHF and RLAIF architectures.
- Scikit-learn for implementing traditional machine learning algorithms and evaluation metrics.
- Natural Language Processing Libraries:
- NLTK and spaCy for text preprocessing and tokenization.
- Transformers library by Hugging Face for utilizing pre-trained transformer models (e.g., BERT).

**Data Collection and Preparation**

The dataset was compiled from various sources to ensure diversity and comprehensiveness:

- **Social Media Posts:** Data was collected using API access and web scraping tools to extract public posts that mentioned keywords related to depression.
- **Clinical Transcripts:** Anonymized transcripts were sourced from institutional databases, ensuring compliance with ethical guidelines.
- **Self-Report Questionnaires:** Data was aggregated from public health databases containing anonymized responses to depression screening tools.
- After collection, the data underwent rigorous preprocessing:

- **Data Cleaning:** Removal of irrelevant information, duplicates, and outliers was performed to enhance data quality.
- **Labeling:** Instances were labeled based on their association with depressive symptoms, with expert input ensuring accuracy.
- **Train-Test Split:** The dataset was divided into training (80%) and testing (20%) subsets, ensuring that the models would be evaluated on unseen data.

**Model Training**

**Baseline Model**

The baseline model was trained using a straightforward supervised learning approach:

Traditional algorithms (e.g., Random Forest and SVM) were trained on the preprocessed data, with hyperparameters optimized through grid search and cross-validation.

**RLHF Model**

The RLHF model underwent a two-phase training process:

- Initial Training: The model was first trained using the labeled dataset to capture basic linguistic patterns.
- Feedback Loop: After the initial training, the model was fine-tuned based on feedback provided by mental health professionals. This feedback loop involved:
- Evaluating the model's predictions against expert assessments.
- Iteratively updating the model parameters based on expert-recommended adjustments.

**RLAIF Model**

The RLAIF model utilized a teacher-student architecture:

- Teacher Model Training: A pre-trained teacher model, capable of generating high-quality predictions, was trained on an extensive dataset.
- Student Model Training: The student model was trained on both labeled data and the
- feedback generated from the teacher model. The training process focused on reinforcing correct predictions and learning from misclassifications through simulated feedback.

## V. RESULTS AND ANALYSIS

**Model Performance Metrics**

The performance of the baseline, RLHF, and RLAIF models was evaluated using the following metrics: accuracy, precision, recall, F1 score, and AUC-ROC.

The results are summarized in Table.

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1 Score | AUC-ROC |
|---|---|---|---|---|---|
| Baseline (Random Forest) | 76.5 | 74.2 | 70.5 | 72.3 | 0.80 |
| RLHF Model | 85.1 | 83.4 | 81.2 | 82.3 | 0.89 |
| RLAIF Model | 87.8 | 86.1 | 84.5 | 85.3 | 0.91 |

**Performance Analysis**

**Baseline Model**

The baseline model, which utilized Random Forest, achieved an accuracy of 76.5%. While it provided a reasonable initial performance, the precision and recall values indicate challenges in correctly identifying all depressive cases, as evidenced by the lower F1 score of 72.3. The AUC-ROC of 0.80 further suggests that the model has moderate discriminative ability.

**RLHF Model**

The RLHF model demonstrated significant improvements across all metrics, achieving an accuracy of 85.1%. The integration of human feedback during the fine-tuning process enabled the model to better understand the complexities of language associated with depression. With a precision of 83.4% and recall of 81.2%, the RLHF model effectively minimized false positives while also identifying a higher proportion of true depressive cases compared to the baseline. The F1 score of 82.3 and AUC-ROC of 0.89 highlight its robustness in performance.

**RLAIF Model**

The RLAIF model outperformed both the baseline and RLHF models, achieving the highest accuracy of 87.8%. The precision of 86.1% and recall of 84.5% indicate that this model successfully balanced the identification of depressive cases with minimizing false classifications. The F1 score of 85.3 reflects its strong overall performance, while the AUC- ROC of 0.91 demonstrates excellent discriminative ability. The results suggest that the use of AI-generated feedback can further enhance the model's understanding of depressive language nuances.

## VI. CONCLUSION

This study presents a significant advancement in the automated detection and assessment of depression through the implementation of Reinforcement Learning from Human Feedback (RLHF) and Reinforcement Learning from AI Feedback (RLAIF) methodologies. The results demonstrate that both RLHF and RLAIF approaches markedly improve the accuracy, precision, and recall of depression detection models when compared to traditional machine learning techniques.

The RLHF model benefitted from the integration of expert feedback, allowing it to better navigate the complexities of language associated with depressive symptoms. The RLAIF model, however, showcased superior performance by leveraging AI-generated feedback, highlighting its potential for scalability and adaptability in mental health assessments. The RLAIF model's higher accuracy rate underscores the promise of these methodologies in addressing the pressing need for effective mental health evaluation tools.

Despite the limitations identified, including dataset scope and the necessity for real-world application testing, the findings of this research provide a solid foundation for future studies. The integration of multimodal data and collaboration with mental health professionals will be critical in further refining these models and ensuring their practical relevance.

Ultimately, this research underscores the transformative potential of machine learning in mental health diagnostics. By harnessing advanced techniques such as RLHF and RLAIF, we can move toward more precise, accessible, and effective solutions for identifying and addressing depression, thereby contributing to improved mental health outcomes.

## REFERENCES

1. Stiennon, N., Bamman, D., & Clark, C. (2020). Learning to summarize with human feedback. Advances in Neural Information Processing Systems, 33, 3008-3021.
2. Hu, Y., Wang, W., & Wang, Y. (2019). A survey on text sentiment analysis: Current research and future directions. Journal of Data Science, 17(1), 1-18.
3. Zhang, Z., Chen, J., & Sun, J. (2021). Deep learning for detecting depression in social media: A survey. IEEE Access, 9, 16262-16279.
4. Bockting, C. L. H., & Hollon, S. D. (2020). Efficacy of cognitive therapy in depression: A meta-analysis. Psychological Bulletin, 146(1), 15-30.
5. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS), 1-20.
6. Choudhury, M. D., & De, S. (2020). Exploring the role of social media in mental health: A systematic review. Journal of Medical Internet Research, 22(8), e18543.
7. Kwon, Y. J., Kim, S. H., & Lee, J. H. (2019). Text classification of mental health issues in social

media: A case study of depression. Health Informatics Journal, 25(4), 1579-1590.
8. Liu, H., & Zhang, W. (2019). A deep learning approach for sentiment analysis in healthcare social media. Journal of Biomedical Informatics, 99, 103297.
9. Trivedi, S. K., & Choudhury, M. (2021). Reinforcement learning in mental health: Challenges and opportunities. Frontiers in Psychiatry, 12, 1-10.
10. Yang, Y., & Liu, X. (2019). Attention-based convolutional neural network for sentiment analysis in social media. Journal of the Association for Information Science and Technology, 70(9), 920-931.