Self-State Evidence Extraction and Well-Being Prediction from Social Media Timelines

Suchandra Chakraborty **Sudeshna Jana** **Manjira Sinha** **Tirthankar Dasgupta**
TCS Research, Kolkata
suchandrac2001@gmail.com,
{sudeshna.jana, sinha.manjira, dasgupta.tirthankar}@tcs.com

Abstract

This study explores the application of Large Language Models (LLMs) and supervised learning to analyze social media posts from Reddit users, addressing two key objectives: first, to extract adaptive and maladaptive self-state evidence that supports psychological assessment (Task A1); and second, to predict a well-being score that reflects the user's mental state (Task A2). We propose i) a fine-tuned RoBERTa (Liu et al., 2019) model for Task A1 to identify self-state evidence spans and ii) evaluate two approaches for Task A2: a retrieval-augmented DeepSeek-7B (DeepSeek-AI et al., 2025) model and a Random Forest regression model trained on sentence embeddings. While LLM-based prompting utilizes contextual reasoning, our findings indicate that supervised learning provides more reliable numerical predictions. The RoBERTa model achieves the highest recall (0.602) for Task A1, and Random Forest regression outperforms DeepSeek-7B for Task A2 (MSE: 2.994 vs. 6.610). These results highlight the strengths and limitations of generative vs. supervised methods in mental health NLP, contributing to the development of privacy-conscious, resource-efficient approaches for psychological assessment. This work is part of the CLPsych 2025 shared task (Tseriotou et al., 2025).

1 Introduction

Mental health assessment using natural language processing (NLP) has evolved from static risk classification to longitudinal modeling of self-states and psychological well-being. The CLPsych Shared Task has progressively introduced more nuanced challenges, moving beyond binary risk assessment to capture dynamic shifts in mental health. The CLPsych 2022 Shared Task (Tsakalidis et al., 2022) was the first to introduce longitudinal modeling, focusing on detecting "Moments of Change" in a user's mood over time and exploring its connection to suicidality risk. The CLPsych 2024 Shared Task (Chim et al., 2024) expanded on this by requiring models to find textual evidence that supports suicide risk levels.

The CLPsych 2025 Shared Task (Tseriotou et al., 2025) extends this research by combining longitudinal modeling with evidence extraction, promoting models that generate human-interpretable rationales while recognizing mental states as they evolve. The shared task consists of four subtasks:

- **Task A1** (Self-State Evidence Extraction): Identifying spans of text that provide evidence for adaptive and maladaptive self-states in a given post.
- **Task A2** (Well-Being Score Prediction): Assigning a well-being score (1–10) to measure the user's psychological state.
- **Task B** (Post-Level Summarization): Generating a summary of the interaction between adaptive and maladaptive states identified in the post.
- **Task C** (Timeline-Level Summarization): Producing longitudinal summaries that capture the trajectory of a user's mental state across multiple posts.

This work focuses on Tasks A1 and A2, which require precise extraction of self-state evidence and structured estimation of well-being scores from the given Reddit post.

Two main approaches exist for these tasks: supervised learning and generative modeling. Supervised methods leverage annotated datasets for structured predictions, using transformer-based models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). Generative models, particularly Large Language Models (LLMs), offer contextual reasoning but require carefully designed prompts to ensure reliable outputs.

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025), pages 279–286
May 3, 2025 ©2025 Association for Computational Linguistics

In this work, we make the following contributions:

1. **Span-Based Evidence Extraction**: We fine-tune a RoBERTa model to extract adaptive and maladaptive self-state evidence (Task A1), achieving a high recall of 0.602. This demonstrates the effectiveness of structured supervision in identifying psychological markers.

2. **Comparative Study of Well-Being Score Prediction**: We evaluate two distinct approaches for Task A2:

   (a) A **retrieval-augmented DeepSeek-7B** model for contextualized well-being estimation.

   (b) A **Random Forest regression** model trained on sentence embeddings for structured numerical prediction.

Our results indicate that supervised learning outperforms LLM-based approaches for numerical well-being regression, while LLMs capture nuanced mental health signals but introduce high variance in predictions. This study contributes to the ongoing development of interpretable, data-driven methods for mental health NLP. The following sections outline our methodology, experiments, and findings.

2 Task Description and Dataset

This study focuses on **Task A1**(Self-State Evidence Extraction) and **Task A2** (Well-Being Score Prediction).

2.1 Task A1: Self-State Evidence Extraction

Given a Reddit post p_j, Task A1 requires identifying spans of text within the post that indicate **adaptive** or **maladaptive** self-states. We define the task as learning a function $f_{\text{A1}} : X_j \rightarrow \{S_{\text{adaptive}}, S_{\text{maladaptive}}\}$, where X_j represents the text of post p_j, and $S_{\text{adaptive}}, S_{\text{maladaptive}}$ are sets of non-overlapping spans belonging to X_j that reflect positive coping mechanisms or distress-driven thought patterns.

2.2 Task A2: Well-Being Score Prediction

Task A2 involves assigning a **well-being score** y_j to each post p_j, where scores range from **1 (severe distress) to 10 (minimal impairment)**, aligning with the Global Assessment of Functioning (GAF) scale. This is framed as a regression problem $f_{\text{A2}} : p_j \rightarrow y_j, \quad y_j \in \{1, 2, ..., 10\}$.

2.3 Dataset Overview

The dataset (Shing et al., 2018; Zirikly et al., 2019; Tsakalidis et al., 2022) consists of 30 user timeline JSON files (343 posts) in the training set and 10 user timeline JSON files (94 posts) in the test set. Each training JSON file contains a timeline ID, a list of posts, and a timeline summary. Each post includes a post index, post ID, timestamp, post text, post summary, well-being score, and evidence annotations. The evidence annotations consist of adaptive and maladaptive states, each with categories and highlighted evidence spans. Each test JSON consists of a timeline ID and a list of posts.

The evidence spans (`adaptive-state` and `maladaptive-state`) are **substrings of the given post text**. The dataset follows the **MIND framework** (Slonim, 2024), modeling mental health as a **dynamic fluctuation of self-states** over time.

3 Methodology

3.1 Task A1: Self-State Evidence Extraction

We frame Task A1 as a **token classification problem**, where each token in a Reddit post is labeled as **adaptive (1)**, **maladaptive (2)**, or **non-evidence (0)**.

3.1.1 Data Preprocessing and Augmentation:

The training data was extracted from annotated JSON files and converted into a CSV format containing Timeline ID, Post, Adaptive Evidence, and Maladaptive Evidence. Posts without any evidence spans were removed, resulting in 199 posts. To enhance robustness, we generated **50 additional posts** using the `nlpaug` (Ma, 2019) library, which provides various NLP-based augmentation methods. Specifically, we applied synonym replacement using the **SynonymAug** augmenter and explicitly configured it to use WordNet as the synonym source. We also applied random word swapping using the **RandomWordAug** augmenter, which randomly exchanges the positions of words within a sentence. This introduced lexical and structural variations while preserving the meaning of the posts, thereby enhancing the overall diversity of the dataset.

3.1.2 Tokenization and Labeling:

We used the **RoBERTa tokenizer** with `add_prefix_space=True` to preserve subword alignment. Evidence spans were mapped to token positions using a rule-based matching algorithm. Labels were assigned directly at the word level by matching evidence spans; each word was initially labeled as non-evidence (0) and then relabeled as adaptive (1) or maladaptive (2) if it was part of the corresponding evidence spans.

3.1.3 Model Architecture and Training:

We fine-tuned a **RoBERTaForTokenClassification** model with three output labels corresponding to evidence categories. The model was trained using Cross-Entropy Loss, AdamW (Loshchilov and Hutter, 2019) optimizer (learning rate = $2e^{-5}$), batch size = 16, and 3 epochs with early stopping based on validation loss to prevent overfitting. We used mixed precision training (fp16) to enhance GPU utilization and speed up training.

3.1.4 Post-Processing and Inference:

During inference, each post is initially split into sentences using a sentence tokenizer. The model generates token-level predictions for each sentence, and the predicted label that is most frequent in that sentence is used as its overall classification.

3.2 Task A2: Well-Being Score Prediction

Task A2 involves assigning each Reddit post a **well-being score** ranging from 1 (severe distress) to 10 (minimal impairment). The training data was extracted from annotated JSON files and converted into a CSV format containing Timeline ID, Post, and Well-being Score. Rows with missing well-being scores were removed.

3.2.1 LLM-Based Approach (DeepSeek-7B)

For this method, we use a retrieval-augmented prompting strategy using DeepSeek-7B, an instruction-tuned causal language model. An overview of this method is shown in Figure 1. The training data is used to generate sentence embeddings via all-MiniLM-L6-v2 (Wang et al., 2020). For each test post, the embedding is computed and compared against the training embeddings using cosine similarity to retrieve the top-k most similar examples. These retrieved examples, along with their well-being scores, are incorporated into a detailed few-shot prompt that begins with a description of the well-being scale based on GAF criteria, followed by instructions to produce a justification sentence and a predicted well-being score. The prompt is tokenized using DeepSeek-7B's tokenizer, and the model generates an output (with parameters such as max_new_tokens set to 50 and temperature to 0.1) from which the numerical score is parsed. Finally, this entire inference pipeline iterates over the test set, and the predicted scores are saved for evaluation. An example prompt used in our approach is provided in the appendix A for reference.

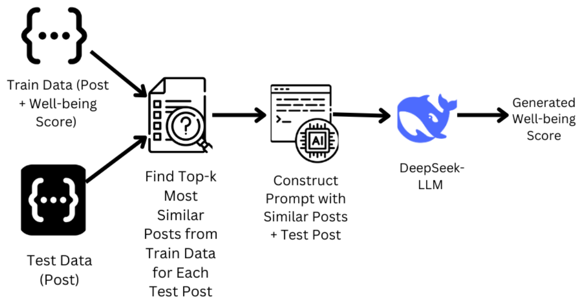
Train Data (Post + Well-being Score)
Test Data (Post)
Find Top-k Most Similar Posts from Train Data for Each Test Post
Construct Prompt with Similar Posts + Test Post
DeepSeek-LLM
Generated Well-being Score

Figure 1: LLM-based Well-being Score Prediction

3.2.2 Supervised Learning Approach (Random Forest Regression)

We also experimented with a supervised regression approach using a Random Forest model trained on sentence embeddings. Sentence representations are generated using all-MiniLM-L6-v2 (Wang et al., 2020), a compact transformer-based embedding model. The feature matrix consists of the embeddings, while the well-being scores serve as the target variable. An 80-20 train-validation split is applied, and a Random Forest Regressor with 200 estimators and a fixed random state is trained on the dataset. Predictions are made on the validation set, and post-processing ensures that outputs are rounded and clipped to integer values within the 1–10 range. For inference, embeddings are generated for the test posts and passed through the trained model. The predicted well-being scores are then stored in a CSV file alongside their corresponding Timeline_ID and Post. Validation performance is assessed using Mean Absolute Error (MAE) and accuracy.

3.2.3 Post-processing:

For DeepSeek-7B, any non-numeric outputs were filtered, and scores exceeding 1–10 were discarded. For Random Forest Regression, predictions were clipped and rounded to ensure numerical consistency.

4 Evaluation Metrics

Task A.1: Evidence of Adaptive and Maladaptive Self-States

- **Recall**: Average of maximum BERTScore for gold spans:

$$\text{Recall} = \frac{1}{|E|} \sum_{e \in E} \max_{h \in H} \text{BERTScore}(e, h)$$

- **Weighted Recall**: Adjusted for predicted span lengths:

$$w = \begin{cases} \frac{n_{\text{gold}}}{n_{\text{pred}}} & \text{if } n_{\text{pred}} > n_{\text{gold}} \\ 1 & \text{otherwise} \end{cases}$$

- **Null Handling**: Defaults to 0 if no spans are submitted.

Task A.2: Well-being Score Prediction

- **Mean Squared Error (MSE)**: Averaged over timelines, computed for:
  - Serious impairment (scores 1-4)
  - Impaired functioning (scores 5-6)
  - Minimal impairment (scores 7-10)
- **Null Handling**: Ignored if no gold score; penalized by max error if no prediction.

5 Results

In Tables 1 and 2, we present the test set results for Task A1 and Task A2. The performance of our methods is compared against baseline models.

5.1 Task A1: Self-State Evidence Extraction

Table 1 presents the results for self-state evidence extraction. Our RoBERTa-based model (MMKA RoBERTa) achieves the **second-highest performance** in the shared task, with an **overall recall of 0.602**. The model shows stronger performance in detecting maladaptive spans (0.681 recall) compared to adaptive spans (0.522 recall), suggesting that distress-related expressions were more easily identifiable by the model. The weighted recall is lower, indicating some level of over-extraction. For a detailed analysis of common misclassification patterns, refer to Appendix B.

Model	**Overall**		**Adaptive**		**Maladaptive**	
	R	W	R	W	R	W
Llama 3.1	0.358	0.337	0.306	0.293	0.382	**0.411**
w/ Window	0.496	0.262	0.365	0.252	0.627	0.272
BART	0.404	**0.382**	0.473	**0.464**	0.336	0.299
w/ Window	0.260	0.258	0.282	0.279	0.238	0.237
MRoBERTa (Ours)	**0.602**	0.343	**0.522**	0.374	**0.681**	0.313

Table 1: Results of our proposed method against baselines on Task A1. "R" and "W" denote recall and weighted recall; *w/ Window* represents the incorporation of post windows.

Model	**MSE↓**	**M-S**	**M-I**	**M-M**	**F1**
Llama 3.1	4.22	4.67	3.66	3.20	0.255
w/ Window	4.46	**1.67**	3.20	7.06	**0.274**
BERT	**2.90**	3.39	2.32	2.81	0.139
w/ Window	4.56	5.68	1.01	5.34	0.135
MMKA DS (Ours)	6.61	4.22	11.76	4.95	0.257
MMKA RF	2.99	4.25	**0.78**	**2.60**	0.197

Table 2: Results of our proposed method against baselines on Task A2. "M-S", "M-I", and "M-M" denote MSE across serious impairment, impaired, and minimal impairment. MMKA DS is our Deepseek approach, and MMKA RF is our Random Forest approach which was not a part of our initial submission.

5.2 Task A2: Well-Being Score Prediction

Table 2 presents the results for well-being score prediction. Our Random Forest Regression model (MMKA Random Forest) achieves the second lowest overall **MSE of 2.994**, outperforming both our submission model DeepSeek-7B and most baselines. However, this approach was not part of our official submission. The DeepSeek-7B model exhibited higher variance and struggled, particularly in severe distress cases, yielding an MSE of 6.610. The results indicate that while LLM-based methods (DeepSeek-7B) capture contextual information, they struggle with numerical stability, often generating inconsistent well-being scores. Additionally, while using LLM-based methods for Task A2, we have faced hallucination issues of LLMs, which is a major drawback of this method. Random Forest Regression, by contrast, provides more stable predictions but lacks interpretability compared to LLM-generated justifications.

5.2.1 Performance Comparison for Task TA2: DeepSeek-7B vs Random Forest

For Task A2 (Well-being Score Prediction), the **Random Forest model** outperformed **DeepSeek-7B**, highlighting key differences between structured machine learning and large language models

(LLMs) for numerical prediction.

Key Factors for Random Forest's Superior Performance

- **Structured Learning:** Random Forest utilizes explicit numerical features and supervised training, which helps the model predict well-being scores precisely. DeepSeek-7B relies on retrieval-augmented prompting, which lacks direct optimization for numerical regression.
- **Stability Interpretability:** Random Forest provides consistent predictions and feature importance insights, while DeepSeek-7B's black-box nature leads to variability and reduced interpretability.
- **Efficiency:** Random Forest makes deterministic predictions efficiently, whereas DeepSeek-7B is computationally expensive and sensitive to retrieval quality.

Future Improvements Enhancing DeepSeek-7B's performance could involve fine-tuning it on domain-specific data, improving retrieval mechanisms, and constraining numerical outputs. Exploring hybrid models combining structured learning with LLM-based contextual reasoning is a promising direction.

6 Conclusion

In this work, we explored approaches for self-state evidence extraction (Task A1) and well-being score prediction (Task A2) as part of the CLPsych 2025 Shared Task. Our RoBERTa-based token classification model achieved the second-best recall (0.602) for Task A1, demonstrating strong performance in detecting maladaptive self-state evidence. For Task A2, we compared a retrieval-augmented LLM (DeepSeek-7B) and a Random Forest regression model. While the DeepSeek-7B model captured contextual information, it exhibited numerical instability. Our Random Forest model outperformed all baselines (MSE = 2.994) except for BERT, but this approach was not part of the official submission.

7 Future Work

For Task A1, future work can focus on **span-level** annotation rather than **sentence-level** classification, allowing the model to distinguish adaptive and maladaptive cues within the same sentence. Future work can also explore **data augmentation using LLMs** for Task A1, which could improve self-state extraction by generating additional diverse training instances. This was not attempted due to computational constraints but presents a promising avenue for enhancing model generalization. Additionally, incorporating **stylistic features** such as sentiment shifts, discourse markers, and writing patterns could provide deeper contextual insights, improving both evidence extraction and well-being prediction. Further, hybrid models that combine the contextual reasoning of LLMs with the numerical stability of regression-based approaches could lead to more robust well-being assessments. Finally, extending models to **capture temporal trends** in user well-being may provide deeper insights into longitudinal mental health assessment.

8 Limitations

Our study has several limitations. First, initial experiments using prompting-based approaches with models such as Mistral-7B (Jiang et al., 2023) and LLaMA (Touvron et al., 2023) for Task A1 resulted in poor performance, with frequent hallucinations and unreliable evidence extraction. As a result, we opted for a RoBERTa-based token classification model, which demonstrated improved robustness. Second, data scarcity remains a significant challenge for both Task A1 and Task A2. Although we applied basic data augmentation techniques to increase the number of training instances, these methods are limited in their ability to capture the full variability of mental health expressions. More advanced data augmentation using LLMs, coupled with a BERT-based model, could potentially yield better performance. Finally, capturing the nuanced and inherently subjective aspects of self-state evidence and well-being scores proved difficult. While we initially anticipated that LLMs would excel in both tasks, they often failed to provide consistent and interpretable results. This suggests that larger models, which might better capture these subtleties, are computationally expensive and present a trade-off between performance and resource requirements.

9 Ethics

The data used in this study consists of sensitive, real user posts collected from Reddit. Although the data are publicly available, we have ensured that all processing is conducted within a secure environment, and no personally identifiable information is shared externally. We strictly adhere to ethical guidelines for data usage and privacy, ensuring that our findings are reported responsibly and without stigmatizing individuals. All analyses and results are derived solely for research purposes and to advance our understanding of mental health dynamics in social media.

References

Jenny Chim, Adam Tsakalidis, Dimitris Gkoumas, Dana Atzil-Slonim, Yaakov Ophir, Ayah Zirikly, Philip Resnik, and Maria Liakata. 2024. Overview of the CLPsych 2024 shared task: Leveraging large language models to identify evidence of suicidality risk in online posts. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 177–190, St. Julians, Malta. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, LA. Association for Computational Linguistics.

Dana Atzil Slonim. 2024. Self-other dynamics (sod): A transtheoretical coding manual.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, Jeff Leintz, and Maria Liakata. 2022. Overview of the clpsych 2022 shared task: Capturing moments of change in longitudinal user posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 184–198, Seattle, USA. Association for Computational Linguistics.

Talia Tseriotou, Jenny Chim, Ayal Klein, Aya Shamir, Guy Dvir, Iqra Ali, Cian Kennedy, Guneet Singh Kohli, Anthony Hills, Ayah Zirikly, Dana Atzil-Slonim, and Maria Liakata. 2025. Overview of the clpsych 2025 shared task: Capturing mental health dynamics from social media timelines. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.

A Example Prompt for DeepSeek-7B for Task A2.

To illustrate the retrieval-augmented prompting strategy used for well-being score prediction, we provide the following dummy example prompt.

You are an advanced language model tasked with rating the overall well-being presented in a given post on a scale from 1 (low well-being) to 10 (high well-being).
The score is based on GAF (American Psychiatric Association, 2000).
The well-being scale is given below:
1 – The person is in persistent danger of severely hurting self or has attempted a serious suicidal act with a clear expectation of death.
2 – In danger of hurting self or others (e.g., suicide attempts; frequently violent; manic excitement) or significant impairment in communication (e.g., incoherent or mute)
.
.
.
10 – No symptoms and superior functioning in a wide range of activities

Examples:
Post: "I've been feeling extremely overwhelmed with work, but I'm trying to manage it by taking breaks."
Well-being score: 7
Post: "Nothing feels enjoyable anymore, and I don't see the point in getting up most days."
Well-being score: 3

Now, read the following post and predict the well-being score. Use the above scale and examples to predict the well-being score. Before predicting the score, justify the predicted score in one full sentence.
Post: "I feel exhausted every day, but I still push through to meet my responsibilities."
Well-being score:

This example demonstrates how the model is guided to score well-being by utilising top-2 retrieved posts with their corresponding labels as examples.

B Error Analysis for Task A1 (Self-State Evidence Extraction)

In our analysis of misclassified instances, we identified several recurring patterns where the model failed to correctly classify evidence spans. Since gold labels were not available for the test data, the analysis was primarily conducted manually. Below, we categorize these errors and provide examples similar to the ones from the test dataset, which the model failed to classify. Note that the current RoBERTa based model has been fine-tuned for **sentence classification** rather than **span classification**. This means that instead of identifying specific spans within a sentence, the model assigns a label to the entire sentence. As a result, it struggles with cases where both adaptive and maladaptive evidence co-exist in a single sentence, leading to ambiguous predictions. In addition, posts with no adaptive or maladaptive evidences from the training data were excluded during the fine-tuning of RoBERTa. This likely contributed to the lower weighted recall compared to recall, as the model struggled to classify sentences as "none" and instead attempted to assign them to one of the predefined classes, even when they did not belong to either.

Mixed Sentiment

Example: "I feel really down, but I know things will get better soon."
Possible Cause: The model struggles to decide whether the sentence leans more positive or negative.

Negation Handling

Example: "I don't think I'm actually sad, just a bit tired."
Possible Cause: The presence of negation ("don't think") may confuse the model into classifying incorrectly.

Ambiguous Language

Example: "Why is everything like this?"
Possible Cause: Without context, the model might not distinguish uncertainty from definitive negative sentiment.

Strong Emotional Words

Example: "I'm completely exhausted and drained, I wish it was not like this."
Possible Cause: The model might overemphasize strong words like "exhausted" and "drained," ignoring the broader context.

The model appears to struggle with mixed sentiments, ambiguous language, emotionally charged words, and multiple ideas within a single post. It seems biased towards classifying strongly emotional statements as maladaptive, even when they contain adaptive elements. Additionally, it might not effectively handle negations or contextual shifts within a sentence, leading to inconsistent classifications. Further analysis could explore the influence of specific keywords and sentence structures in model errors.

These findings suggest that improving contextual understanding and refining the handling of ambiguity in language could enhance model performance in Task A1.