ReDepress: A Cognitive Framework for Predicting Depression Relapse from Social Media

Aakash Kumar Agarwal[1]*, **Saprativa Bhattacharjee**[1]*, **Mauli Rastogi**[2], **Jemima S. Jacob**[3], **Biplab Banerjee**[1], **Rashmi Gupta**[1] and **Pushpak Bhattacharyya**[1]

[1]Indian Institute of Technology Bombay, Mumbai, India
[2]Clinical Psychologist, Bangalore, India
[3]Mental Health Consultant, HELPED, Mumbai, India

{aakash.agarwal, bbanerjee}@iitb.ac.in, {saprativa, pb}@cse.iitb.ac.in,
{maulirastogicp, jemimajohnjacob}@gmail.com, rash_cogsci@yahoo.com

Abstract

Almost 50% depression patients face the risk of going into relapse. The risk increases to 80% after the second episode of depression. Although, depression detection from social media has attained considerable attention, depression relapse detection has remained largely unexplored due to the lack of curated datasets and the difficulty of distinguishing relapse and non-relapse users. In this work, we present *ReDepress*, the first clinically validated social media dataset focused on relapse, comprising 204 Reddit users annotated by mental health professionals. Unlike prior approaches, our framework draws on cognitive theories of depression, incorporating constructs such as attention bias, interpretation bias, memory bias and rumination into both annotation and modeling. Through statistical analyses and machine learning experiments, we demonstrate that cognitive markers significantly differentiate relapse and non-relapse groups, and that models enriched with these features achieve competitive performance, with transformer-based temporal models attaining an F1 of 0.86. Our findings validate psychological theories in real-world digital data and underscore the potential of cognitive-informed computational methods for early relapse detection, paving the way for scalable, low-cost interventions in mental healthcare.

1 Introduction

In the treatment of depression, one of the biggest challenge faced by both medical practitioners and patients alike, is the risk of getting into relapse (Burcusa and Iacono, 2007; American Psychiatric Association, 2025). Previous studies have found that, one out of every two depression patients (50%), are likely to go into relapse. After the second episode of depression, this risk drastically increases to 80%. Further, it becomes 90% for those

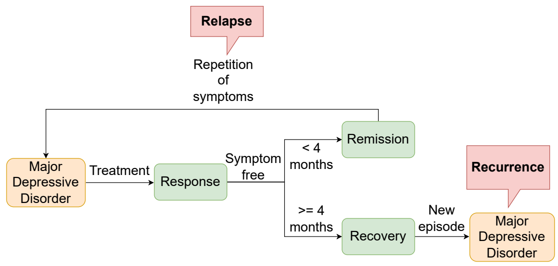

Figure 1: Visual depiction of relapse and other related terms encountered during the treatment of depression. The timelines are as per Muzammel et al. (2021). In this work we use the terms relapse and recurrence synonymously. For a more elaborate discussion please see Appendix A.

with a history of three episodes (Kupfer et al., 1996; Muzammel et al., 2021). This puts heavy burden on the patients as well as the healthcare infrastructure. Additionally, the risk of suicide among depression patients is twenty times higher than the normal populace (Harris and Barraclough, 1997). Early detection of depression relapse thus becomes crucial for devising timely preventive intervention measures.

Numerous computational studies have addressed the task of depression detection from social media posts either as a standalone disorder (Mendes and Caseli, 2024) or even in a setting with other comorbidities (Hengle et al., 2024). Moreover, quite a few datasets are also available for depression detection from social media viz., e-Risk (Losada et al., 2018) and Reddit Mental Health Dataset (Low et al., 2020).

But there has been no study on the possibility of using social media posts for detection of depression relapse. This could be attributed to (1) the lack of a high quality depression relapse dataset, and (2) the inherent difficulty of separating relapse and non-relapse users because of the high similarity

*Equal contribution.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 34664–34682
November 4-9, 2025

between the posts made by both set of users.

To address these gaps, in this work we introduce a high quality depression relapse dataset named *ReDepress*, curated from the social media website Reddit. ReDepress is meticulously curated through a multistage process involving regular expressions, large language models (LLMs), non-expert humans and finally clinical psychologists.

Studies in cognitive sciences have established the close connection between depression and cognitive processes such as rumination and information processing biases (LeMoult and Gotlib, 2019; Gupta and Kar, 2012; Gotlib and Joormann, 2010; Gupta and Kar, 2008). Moreover, the most widely used therapy for depression, Cognitive Behavior Therapy (Beck, 1967), is also based on cognitive science. Thus, in this study we chose to study the effects of four cognitive constructs– attention bias, memory bias, interpretation bias and rumination (see Figure 2). Our hypothesis is that, early identification of shifts in these biases and rumination could help predict relapse of depression. Moreover, we believe successful early detection of depression relapse based on social media posts could pave the way for a cost-effective approach to improve the quality of life for the depression patients and also ease the burden on the healthcare infrastructure.

Our contributions are:

1. **ReDepress Dataset**: We present the first depression relapse focused social media dataset consisting of 204 (83 relapse and 121 non-relapse) users curated from Reddit. (section 3)

2. **Clinical Annotation**: Our dataset is annotated by three clinical psychologists who deal with depression patients regularly ensuring that our dataset is clinically validated from the very beginning. (section 3)

3. **Cognitive Markers for Relapse Prediction**: We explore whether the four cognitive dimensions- attention bias, interpretation bias, memory bias and rumination correlate with relapse of depression, thereby validating psychological theories using real-world textual data. (section 5)

2 Related Work

2.1 Cognitive Theories of Depression

Cognitive scientists underscore that depression is not only about low mood, but also about systematic biases in attention, interpretation and memory. Gotlib and McCann (1984) found that depressed individuals show a stable attentional bias toward depression-related words. Butler and Mathews (1983) found that depressed individuals were more likely to endorse negative interpretations compared to non-depressed individuals. Lloyd and Lishman (1975) demonstrated that depressed individuals recall negative experiences faster than positive ones, linking memory bias to depression. Finally, Nolen-Hoeksema (1991) introduced rumination as a cognitive style that prolongs depressive episodes. These studies motivated our choice of picking the cognitive dimensions of attention bias, interpretation bias, memory bias and rumination for our work.

2.2 Clinical Studies on Depression Relapse

Burcusa and Iacono (2007) explores factors specifically associated with depression recurrence, differentiating them from those related to the initial onset of depression. Salvini et al. (2015) proposes a multi-relational predictive model for depression relapse in bipolar disorder patients, employing Inductive Logic Programming (ILP) techniques on relational clinical data. Lye et al. (2020) conducts a five-year longitudinal study of 201 patients with Major Depressive Disorder (MDD), examining a range of prognostic factors including clinical, personality, environmental, and genetic variables. Meanwhile, Moriarty et al. (2021, 2022) provide a systematic review of existing prognostic models for depression relapse, highlighting significant methodological limitations and the widespread lack of external validation. Together, these studies highlight the high incidence of depression relapse and stress the urgent need for robust, clinically validated prediction tools to facilitate personalized and timely interventions.

2.3 Non-Clinical Studies on Depression Relapse

Aziz et al. (2009) initiated research on modeling depression relapse through dynamic agent simulations, later extending this line of work into an integrative ambient agent framework that combined relapse dynamics with intelligent agent technology (Aziz et al., 2010) for preventing relapse in individuals with a history of unipolar depression. Building on such conceptual models, subsequent efforts shifted toward data-driven prediction. For instance, Nie et al. (2016) leveraged the STAR*D

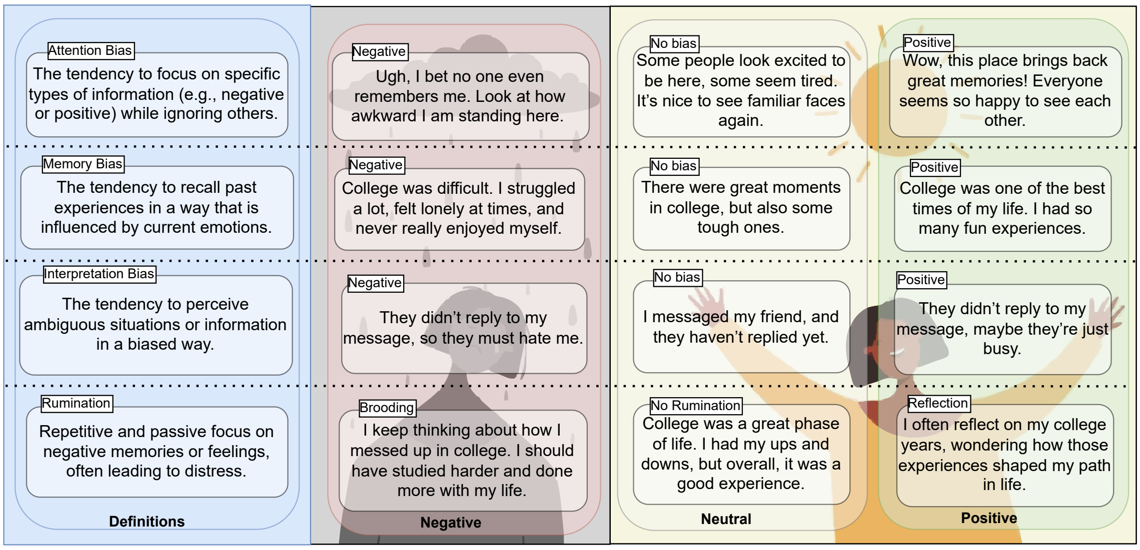

Figure 2: Definitions and examples of attention bias, memory bias, interpretation bias and rumination. Our ReDepress dataset is annotated on these cognitive dimensions (Features) at the post level.

dataset[1] and introduced a censored regression approach with truncated l_1 loss to estimate relapse risk across treatment stages. Similarly, Dwyer (2019) explored the utility of machine learning, particularly LSTMs, on anonymized electronic health records (EHRs) for relapse forecasting.

Parallelly, researchers also examined signals from everyday digital traces. Garcia et al. (2021) demonstrated that smartphone-derived lifelog data could be coupled with survival models to estimate relapse risk. Expanding further, Yin et al. (2022) proposed an intelligent mobile monitoring platform that fused acoustic, semantic, environmental, and personal features via a CNN–LSTM model. More recently, wearable technologies have also been investigated, as in the work of Matcham et al. (2024), who studied the relationship between Fitbit-derived sleep features and relapse in individuals with recurrent depression. Complementing these, Lucasius et al. (2024) introduced a multimodal approach for adolescents by integrating video and speech data, while Muzammel et al. (2021) incorporated audio-visual cues from the DAIC-WOZ dataset (Gratch et al., 2014) into deep learning systems for detecting early signs of relapse.

Despite the breadth of methodologies, ranging from agent-based simulations to EHRs, smartphones, wearables and multimodal sensing, prior research has not explored the use of social media posts as a data source for relapse detection. We argue that analyzing social media posts offers a cost-effective and unobtrusive means of continuous monitoring, with the potential to enable early identification of vulnerable patients at scale.

3 Dataset

3.1 Data Curation

Our dataset curation pipeline is depicted in Figure 3. We start with a decade of Reddit dumps[2] from 2012 to 2022 and consider only those users who have posted at least once in *r/depression* subreddit and whose number of posts (overall) lie between 100 to 500. The lower limit is to have sufficient context per user and the upper limit is to minimize the cost and effort of manual annotation. For the first level of filtering we utilize an expanded version of the patterns provided by Cohan et al. (2018) to extract the users who have self-reported being diagnosed with depression. We expanded the pattern list with the help of GPT-4o. The expanded list is provided in Appendix B. Next we use Llama 3.2 3B model on the posts of the users filtered in the previous step to further filter only those posts which has mention of any mental health issue to identify genuine depressed users and reduce false positives. We use a small model here because the number of input posts are quite high and it would incur a high computational cost if we use a bigger model. Also, we do not use GPT-4o here because of budget constraints. But smaller models are more

[1] https://medicine.yale.edu/lab/statmethods/datasets/stard/

[2] Available at https://the-eye.eu/redarcs/

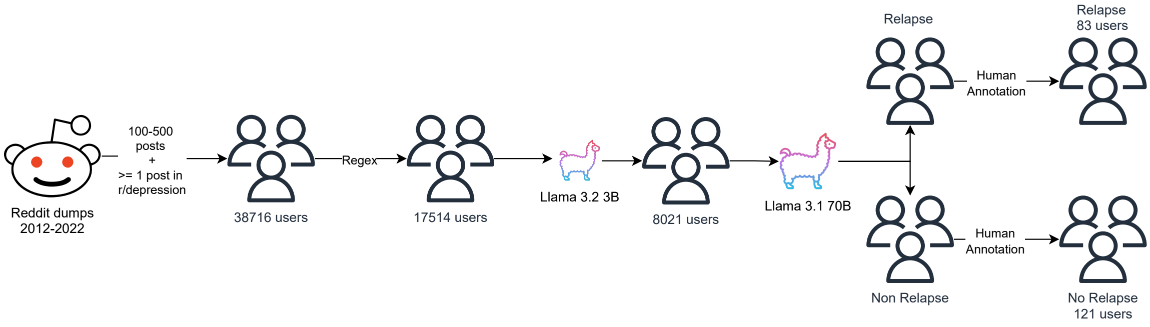

Figure 3: Dataset curation pipeline.

	#Users	Min #Posts	Max #Posts	Avg #Posts
Relapse	83	4	37	15.44
No Relapse	121	3	35	11.24

Table 1: Dataset Statistics. Min #Posts and Max #Posts refer to the minimum and maximum number of posts of any user in that class respectively. Avg #Posts refers to the average number of posts of all users in a class.

	Full Agreement	Majority Agreement	Level
Attention Bias	0.39	0.97	Post level
Memory Bias	0.18	0.98	Post level
Interpretation Bias	0.29	0.99	Post level
Rumination	0.17	0.92	Post level
Relapse/No Relapse	0.76	0.99	User level

Table 2: Inter-Annotator Agreement Metrics (Fleiss' Kappa). We follow Tsakalidis et al. (2022) in reporting the majority agreement scores.

prone to hallucinations, thus, finally we use Llama 3.1 70B model on the depression users identified by Llama 3.2 3B to further improve the accuracy. The prompts used are given in Appendix C.

3.2 Timeline Extraction

We create a dataset of Reddit user's timelines. A user's timeline is a subset of her entire posting history. Each user's posts are ordered chronologically. Depending on the answers of remission and relapse provided by Llama 3.1 70B, relapse and no relapse users are identified through manual inspection first by non-experts (first two authors). The manual inspection begins with verifying that the user has indeed self-reported being diagnosed with depression or not. Next, for creation of the timelines for a user, the post identified as remission is taken as the starting post. Then subsequent posts are checked up to a duration of one and a half year from the remission post. If a subsequent post (only those posts which had any answers given by Llama 3.1 70B as per prompt in Appendix C) mentioning either relapse or return of symptoms is found, the timeline is extracted and the user is a potential candidate belonging to the relapse class. Conversely, if no such posts are found for the specified duration, the user is most likely a no relapse case. Following this process we extract a total of 208 user timelines. We send these for annotation by clinical psychologists.

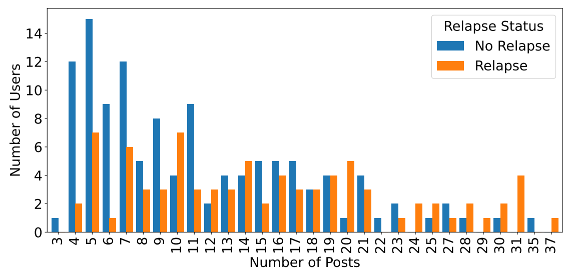

Figure 4: Distribution of user posts.

3.3 Annotation

Final annotations are carried out by three clinical psychologists based on annotation guidelines (Appendix D) created after a series of discussion rounds. The annotations are carried out at two levels– user level and post level. At the user level, the psychologists are asked to mark the remission and relapse posts of a user. Based on majority vote, i.e. if 2 out of 3 annotators mark any post of a given user as relapse, we consider that as a relapse user otherwise no relapse. Moreover, if the first couple of posts are not identified as a remission post, the user is discarded. Through this we get 83 relapse and 121 no relapse users. 4 users remained inconclusive due to lack of agreement with respect to either remission or relapse posts. The dataset statistics are presented in Table 1.

At the post level, annotators are asked to annotate each post on four dimensions– attention

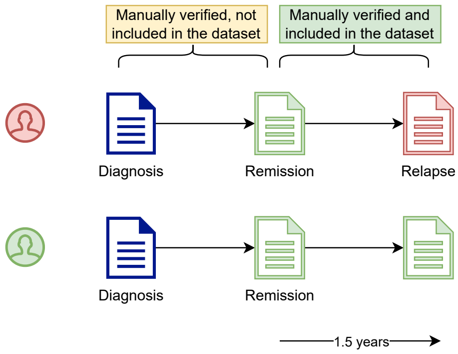

Figure 5: Timeline of a relapse vs no relapse user.

bias, interpretation bias, memory bias and rumination. The inter-annotator agreement metrics are presented in Table 2. The posts distributions are shown in Figure 4. The three annotators employed by us were all female of the age group 20-30 years. All of them are clinical psychologists working at different organizations. The annotators were compensated for their efforts as per the norms. We plan to make the dataset available upon request under a user agreement to ensure responsible usage[3].

3.4 Examples

An example timeline difference between relapse and no relapse is illustrated in Figure 5. Only the part from remission to relapse (or till one and a half years) is included in the dataset. The other part from diagnosis till remission (excluding) is only used during sample selection but not included in the final dataset. For a sample remission and relapse post from the dataset, please see Appendix E. For an elaborate discussion on the how the distribution of relapse cases differs between our dataset based on Reddit and real life, please refer Appendix F.

4 Experimental Setup

We conduct a series of experiments to evaluate the predictive power of cognitive markers for relapse detection. At the user level, post-level annotations of attention bias, interpretation bias, memory bias, and rumination are aggregated using statistical measures such as mean, median, minimum, and maximum, and multiple machine learning classifiers are trained on these aggregated features. Hyperparameters are tuned via grid search, with evaluation performed using 5-fold cross-validation on 80% of the data and the best-performing model tested on the remaining 20%.

[3]Details available at `https://github.com/saprativa/ReDepress`

Feature (Average)	Mann-Whitney p-value
Attention Bias	0.000000
Interpretation Bias	0.000002
Memory Bias	0.000212
Rumination	0.000014

Table 3: Statistical test p-values comparing Relapse vs No Relapse groups for cognitive features (Average). To determine the appropriate statistical tests, we conducted normality tests on the variables. See Appendix G for details.

To capture temporal dynamics in user timelines, we implement transformer-based encoders where each post is represented by a concatenation of its text embedding and cognitive markers, with the input being the chronologically ordered embeddings of each user. Finally, to test the feasibility of automating cognitive feature extraction, we train BERT-based classifiers for each cognitive dimension. We also benchmark large language models on this task in a zero-shot setting. Performance across all experiments are assessed with standard metrics including Accuracy, Precision, Recall, and F1-score.

5 Results and Discussion

5.1 Quantitative Analysis

Figure 6 displays the distribution of cognitive dimension scores-attention bias, interpretation bias, memory bias, and rumination-across relapse and no-relapse groups with relapse users exhibiting more negative attention, interpretation and memory biases, and also higher levels of brooding rumination. As evident from Table 3, the differences between the two groups groups are statistically significant for all cognitive features ($p < 0.01$).

In addition to static distributions, we also examined the temporal evolution of cognitive biases by aggregating multiple bias measures into combined scores and normalizing them over user timelines. Figure 7 illustrates how these combined bias scores change across time for relapse and no-relapse groups. Individuals who eventually relapsed exhibited a stronger downward trajectory, with mean values consistently below those of no-relapse group, indicating more persistent negative cognitive biases. While both groups displayed variability, the divergence between their trajectories became increasingly pronounced as time progressed, underscoring the temporal buildup of cognitive vulnerabilities in relapse cases.

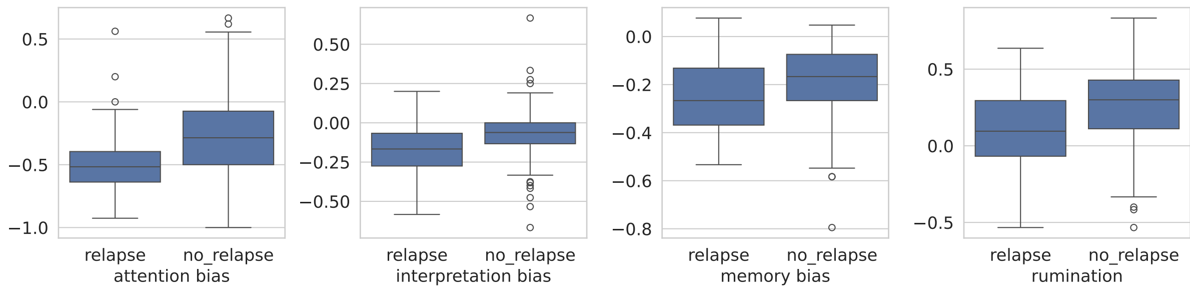

Figure 6: Distributions of average attention bias, interpretation bias, memory bias and rumination across relapse and no relapse groups. Appendix H discusses the averaging applied.

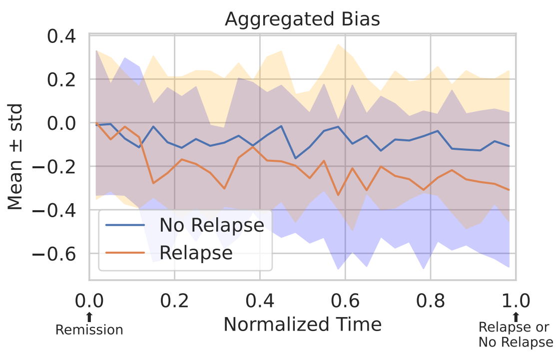

Figure 7: Temporal evolution of aggregated cognitive bias scores across normalized user timelines.

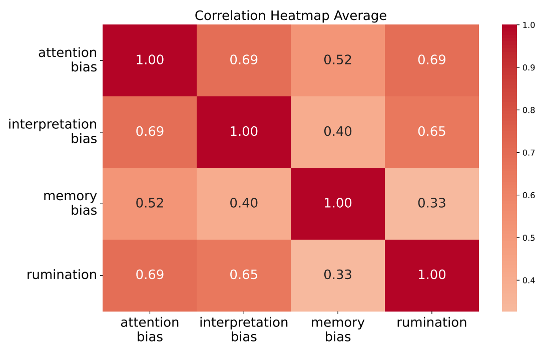

Figure 8: Cognitive dimensions correlation heatmap.

We also computed a correlation heatmap (Figure 8) based on average user-level scores across the four annotated features. The analysis revealed moderately strong positive correlations between attention bias and interpretation bias (r = 0.69), as well as between attention bias and rumination (r = 0.69), and interpretation bias and rumination (r = 0.65). Memory bias showed weaker, though still positive, associations with the other dimensions, most notably with attention bias (r = 0.52).

Importantly, none of the cognitive dimensions showed very high correlations (r > 0.80), suggesting that these measures capture related but distinct processes, rather than reflecting a single underlying factor such as sentiment alone. Overall, this correlation structure underscores the interconnected nature of cognitive vulnerabilities in depression and supports our hypothesis that a multidimensional cognitive lens is essential for understanding and predicting relapse risk.

These results are consistent with clinical studies, which demonstrate that individuals vulnerable to relapse tend to display persistent negative biases in attention, memory, and interpretation, as well as increased rumination, even after remission (Visted et al., 2018; LeMoult and Gotlib, 2019). However, as visible in the Figure 6, while group differences are significant, the distributions overlap substantially, reflecting the clinical reality that remitted individuals often retain residual cognitive vulnerabilities, though to a lesser degree than those in active relapse.

5.2 Predictive Modeling

We computed cognitive bias features for each user by aggregating their post-level ratings using statistical measures (mean, median, min and max) across the four cognitive dimensions: attention, interpretation, memory, and rumination. These aggregated features served as inputs for several machine learning models, including Random Forest, XGBoost, Logistic Regression, Gradient Boosting, SVM, KNN, and Neural Network.

Table 4 reports the performance of each model. The GradientBoosting classifier achieved the best overall performance (Accuracy: 0.80, F1-Score: 0.78), indicating its effectiveness in distinguishing between relapse and non-relapse users based on the cognitive feature set. The best performing hyperparameters are listed in Table 13 (Appendix I). Neural Network also performed competitively, with F1-score of 0.76. While these results are promising, the overall performance, with an F1-score of 0.78 at its best, is not exceptionally high. This

Model	Accuracy	Precision	Recall	F1
RandomForest	0.71	0.59	**0.94**	0.73
XGBoost	0.73	0.62	**0.94**	0.74
LogisticRegression	0.73	0.62	**0.94**	0.74
GradientBoosting	**0.80**	**0.74**	0.82	**0.78**
SVM	0.73	0.62	**0.94**	0.74
KNN	0.68	0.57	**0.94**	0.71
Neural Network	0.78	0.70	0.82	0.76

Table 4: Performance Metrics for Different Models on Test Set (rounded to 2 decimal places).

Feature	Statistic	$\mathrm{II}_{\mathrm{Ablated}}$	Δ II	Interpretation
Rumination	Mean	0.611	+0.014	Most important
Interpretation Bias	Mean	0.605	+0.008	Some importance

Table 5: Top two features contributing most to relapse prediction according to information imbalance ablation analysis. Higher positive ΔII indicates greater unique informativeness.

suggests that the models could potentially benefit from additional features, such as those present within the posts of users, to improve their predictive power and better distinguish between relapse and non-relapse users.

5.3 Interpretation and Clinical Relevance

Our findings validate cognitive models of depression, showing that users at risk of relapse exhibit more pronounced negative attention, memory, and interpretation biases, as well as elevated rumination, mirroring cognitive patterns observed in clinical populations. The predictive results suggest that cognitive features extracted from social media can meaningfully distinguish relapse from no-relapse cases, but also reinforce the subtlety of these differences: remitted individuals retain mild cognitive vulnerabilities, as reflected by the moderate but not extreme model performances and overlapping group distributions. This is in line with clinical literature indicating residual symptoms and biases often persist after remission, contributing to future relapse risk.

5.4 Ablation study

To further investigate the individual importance of each cognitive bias feature in relapse prediction, we conducted comprehensive ablation studies using both model-free and model-based approaches.

Model-Free Feature Importance: Information Imbalance Ablation We first employed a model-free, information-theoretic ablation based on the information imbalance metric[4]. This method measures the change in information transfer from features to the relapse label when a given feature is removed, independent of any specific predictive model. Table 5 shows the top two most informative features, while Table 15 in Appendix J provides the full ablation results for each type of bias and aggregation (min, max, mean, median). A larger positive ΔII indicates a higher unique contribution of that feature to the label; negative or near-zero values suggest redundancy.

Results indicate that the mean of rumination (+0.014) and the mean of interpretation bias (+0.008) are the most critical features, as their removal leads to the greatest increase in information imbalance. In contrast, features such as the median of interpretation bias (-0.025) and the maximum of rumination and attention bias (-0.020) show negative or negligible ΔII, suggesting these features are largely redundant or even introduce noise.

Feature	Mean	Min	Max	Median
Memory Bias	0.159	0.094	0.041	0.060
Attention Bias	0.078	0.041	0.111	0.060
Rumination	0.041	0.021	0.078	0.111
Interpretation Bias	0.009	0.060	0.095	0.078

Table 6: F1 drop per feature type and aggregation.

Model-Based Feature Importance: F1 Score Ablation To complement the model-free analysis, we performed a model-based ablation using the best-performing classifier (GradientBoosting classifier, F1-Score: 0.78). For each feature, we retrained the model after removing that feature and measured the drop in F1 score on the test set. The results are summarized in Table 6.

The largest F1 drop occurs when removing Memory Bias, particularly the mean aggregation (0.159), followed by Attention Bias - Max (0.111) and Rumination - Median (0.111). Removing Interpretation Bias results in only a minimal change across aggregations (maximum F1 drop 0.095), suggesting that it contributes limited unique predictive information. Overall, features related to memory and attention biases appear to have the strongest impact on the model's predictive performance.

5.5 Temporal Sequence Modeling

While aggregated features provide useful signals, they ignore the sequential nature of user timelines.

[4] https://bit.ly/3VY8IRL

Embedding	Strategy	Accuracy	Precision	Recall	F1
mentalbert	No CM	0.81	0.85	0.81	0.80
mentalbert	CM	**0.83**	**0.86**	**0.83**	**0.82**
mentalbert	CM-Emb	**0.83**	**0.86**	**0.83**	**0.82**
mentalroberta	No CM	**0.83**	0.85	**0.83**	0.80
mentalroberta	CM	0.81	0.83	0.81	**0.81**
mentalroberta	CM-Emb	0.81	**0.87**	0.81	0.79
mpnet	No CM	0.86	0.86	0.86	0.81
mpnet	CM	**0.88***	0.88	**0.88***	**0.86***
mpnet	CM-Emb	**0.88***	**0.89***	**0.88***	**0.86***

Table 7: Binary relapse prediction results using transformer embeddings under three settings: No CM = No Cognitive Markers, CM = Cognitive Markers, CM-Emb = Cognitive Markers Embedding. **Bold** highlights the best result within each embedding type, and * marks the overall best across all embeddings.

Feature	Model	Accuracy	Precision	Recall	F1
Attention bias	bert-base-uncased	0.66	0.68	0.66	0.66
	clinicalbert	0.67	0.67	0.67	0.67
	mentalbert	**0.74**	**0.73**	**0.74**	**0.73**
	mentalroberta	0.73	**0.73**	0.73	**0.73**
Interpretation bias	bert-base-uncased	0.85	0.84	0.85	0.85
	clinicalbert	0.85	0.83	0.85	0.84
	mentalbert	0.87	0.87	0.87	0.87
	mentalroberta	**0.90**	**0.90**	**0.90**	**0.89**
Memory bias	bert-base-uncased	0.80	0.78	0.80	0.79
	clinicalbert	0.82	0.80	0.82	0.81
	mentalbert	0.82	0.81	0.82	0.81
	mentalroberta	**0.83**	**0.82**	**0.83**	**0.82**
Rumination	bert-base-uncased	0.72	0.72	0.72	0.72
	clinicalbert	0.65	0.68	0.65	0.65
	mentalbert	0.75	0.74	0.75	0.75
	mentalroberta	**0.76**	**0.76**	**0.76**	**0.76**

Table 8: Results of cognitive dimension classifiers (best values in each block are highlighted in bold).

Feature	Positive	Negative	No Bias
Memory Bias	12	497	1828
Attention Bias	237	1201	899
Interpretation Bias	48	268	2021
	Reflection	**Brooding**	**No Rumination**
Rumination	1216	576	545

Table 9: Final majority label distributions for each cognitive dimension.

To model temporal dependencies, we implemented a transformer-based encoder where each Reddit post is represented by its text embedding concatenated with four cognitive markers (attention bias, interpretation bias, memory bias, rumination). The chronological sequences are encoded, and the final valid hidden state is used for classification. Unlike the original transformer (Vaswani et al., 2017) model that takes tokens as input, our model takes post embeddings as input .

We experimented with three input configurations:

- **No Cognitive Markers (No CM):** Embedding (`<post>`)
- **Cognitive Markers (CM):** Embedding (`<post>`) + `<Cognitive Markers>`
- **Cognitive Markers Embedding (CM-Emb):** Embedding (`<post> + <Cognitive Markers>`)

Table 7 summarizes results across `mentalbert`, `mentalroberta` (Ji et al., 2022), and `mpnet` (Song et al., 2020). Cognitive markers consistently enhance performance, with the largest gains observed for `mpnet`, where F1 improved from 0.81 (embeddings only) to 0.86 with markers. `mentalbert` also benefited, achieving 0.82 F1 with markers compared to 0.80 without.

Interestingly, while mental health domain-adapted models such as `mentalbert` and `mentalroberta` perform well, they are consistently outperformed by `mpnet`, which is not domain-adapted. This suggests that domain-adapted models are primarily optimized for tasks with clear diagnostic differences, such as depression detection, whereas relapse prediction on Redepress involves more subtle cognitive and behavioral cues.

Overall, these results show that temporal sequence models enriched with cognitive features better capture relapse risk, highlighting the importance of integrating clinical insights into sequential text modeling. The best hyperparameter settings for these transformer models are reported in Table 14 (Appendix I).

5.6 Cognitive Dimension Classifiers

For assessing the feasibility of automated cognitive dimension labeling, we trained dedicated classifiers for each cognitive feature using multiple transformer-based models (Table 8). We performed three-class classification for each cognitive construct individually. Specifically, attention bias, memory bias, and interpretation bias were classified into *positive*, *negative*, and *no bias* categories, while rumination was classified into *reflection*, *brooding*, and *no rumination*. The full label distributions for these are provided in Table 9.

Results indicate that transformer models can reliably identify cognitive markers from text, with the best performance observed for interpretation and memory bias classification. Notably, the mental health–domain pretrained models (MentalBERT, MentalRoBERTa) outperform general-domain BERT variant (Devlin et al., 2019) or even allied domain ClinicalBERT (Huang et al., 2019). This demonstrates that models adapted to mental health contexts are better at capturing nuanced cognitive constructs in social media posts. The relatively lower but consistent scores for attention bias and rumination suggest that these constructs are more challenging to capture, likely due to their subtler textual manifestations. Nevertheless, even for these dimensions, mental health–specific models achieve higher accuracy and F1 compared to their general-purpose counterparts.

Feature	Model	Accuracy	Precision	Recall	F1
Attention bias	gemma3_27b	0.68	0.65	0.63	0.58
	llama3.1_70b	**0.72**	0.66	**0.68**	**0.66**
	qwen2_72b	0.71	**0.70**	0.63	0.65
Interpretation bias	gemma3_27b	0.28	**0.43**	0.60	0.25
	llama3.1_70b	**0.41**	**0.43**	**0.63**	**0.33**
	qwen2_72b	0.35	**0.43**	0.58	0.30
Memory bias	gemma3_27b	0.51	0.44	**0.62**	0.38
	llama3.1_70b	0.59	0.45	**0.62**	0.43
	qwen2_72b	**0.66**	**0.48**	**0.62**	**0.47**
Rumination	gemma3_27b	0.56	**0.58**	0.58	0.54
	llama3.1_70b	**0.59**	**0.58**	**0.60**	**0.57**
	qwen2_72b	0.49	0.57	0.56	0.45

Table 10: Results of cognitive dimension classifiers for large language models (Gemma-3-27B (Team, 2025), Llama-3.1-70B (Grattafiori et al., 2024), Qwen-2-72B (Yang et al., 2024)). Best values in each block are highlighted in **bold**.

5.7 Zero-Shot Evaluation with LLMs

In addition to fine-tuned classifiers, we also evaluated large language models (LLMs) in a zero-shot setting for each cognitive dimension (Table 10). The task setup was identical to the supervised experiments: attention bias, memory bias, and interpretation bias used the categories *positive*, *negative*, and *no bias*, while rumination was classified into *reflection*, *brooding*, and *no rumination*. The exact zero-shot prompt used for these evaluations is provided in Appendix K.

Performance varied across models and dimensions. For attention bias, Llama-3.1-70B achieved the highest accuracy (0.72, F1 = 0.66), with Qwen-2-72B showing the highest precision (0.70). Interpretation bias was the most challenging (best: Llama-3.1-70B, Accuracy = 0.41, Recall = 0.63, F1 = 0.33). For memory bias, Qwen-2-72B performed best (Accuracy = 0.66, F1 = 0.47), while rumination was best captured by Llama-3.1-70B (Accuracy = 0.59, Recall = 0.60, F1 = 0.57). These results indicate that zero-shot LLMs can capture some cognitive dimensions without training, but their performance remains below domain-adapted supervised models, underscoring the need for careful annotation and domain-specific adaptation.

6 Conclusion and Future Work

This paper introduced *ReDepress*, the first clinically validated social media dataset dedicated to depression relapse, annotated by clinical psychologists and grounded in cognitive theory. Our analyses confirmed that cognitive markers correlate strongly with relapse risk and can be computationally modeled from Reddit timelines. By integrating attention, interpretation, and memory biases alongside rumination into both annotation and modeling, we demonstrated that these dimensions are statistically significant indicators of relapse and capture subtle cognitive vulnerabilities reflected in user posts. Statistical tests, machine learning classifiers, and transformer-based sequence models further underscored their predictive utility, with cognitive-enriched approaches consistently achieving performance gains. These results not only validate long-standing cognitive theories of depression but also highlight the potential of combining computational methods with clinical insights to support early relapse detection.

At the same time, our findings reveal the inherent challenges of the task: relapse and non-relapse users often share overlapping cognitive patterns, reflecting the subtle nature of residual vulnerabilities after remission. This underscores the need for more robust and context-aware approaches. Future work will focus on expanding ReDepress to larger and more diverse populations, incorporating multimodal signals such as behavioral and physiological data, and developing finer-grained temporal models that capture both short-term fluctuations and long-term cognitive shifts. We also envision personalized modeling frameworks that adapt to individual baselines and trajectories, increasing sensitivity to early relapse signs. Finally, ethical considerations around privacy, responsible deployment, and clinical translation remain central, requiring collaboration with mental health professionals to ensure safe and effective integration of such systems into real-world care.

Limitations

While our work presents a novel approach to detecting depression relapse using social media data, it is essential to acknowledge its limitations.

- **Annotation Subjectivity:** The identification of complex psychological constructs such as rumination or emotional distress relies on human annotation, which is inherently subjective. Despite our efforts to establish clear guidelines and conduct rigorous training, inter-annotator agreement remains a challenge. Variability in individual interpretations may introduce bias, particularly when labeling nuanced signals of relapse.

- **External Validity:** Our dataset is derived exclusively from Reddit, which may limit the generalizability of our findings. Reddit users are not necessarily representative of the broader population, and their self-disclosures on the platform may differ from experiences shared in offline or clinical contexts. Furthermore, the self-reporting nature of social media introduces potential biases, as users may selectively share specific aspects of their mental health. Future research should explore datasets from diverse platforms or clinical sources to validate the robustness of our findings across contexts.

- **Temporal Granularity:** The temporal resolution of social media posts does not always align with clinical definitions of depression relapse. Users may post inconsistently, leading to gaps in data that obscure important transitions in mental state. Moreover, the timing of self-disclosures may lag behind or precede actual relapse episodes. To address this limitation, future studies could integrate multimodal longitudinal data or use active user engagement methods to capture finer temporal details.

- **Ethical Considerations:** Analyzing sensitive mental health data from social media raises ethical concerns. While our data collection adheres to publicly available sources, there is an inherent risk of identifying or labeling vulnerable individuals. We prioritize user privacy by anonymizing data and following ethical research guidelines. Nevertheless, future work should consider additional safeguards, such as engaging with mental health professionals and developing frameworks to mitigate potential harm when applying these models in real-world scenarios.

Acknowledgments

We would like to thank the reviewers for their valuable feedback, which helped shape this paper in the right direction. We also extend our gratitude to the members of the CFILT lab for their support and insightful discussions. Finally, we sincerely thank the annotators for their effort and contribution to this work.

References

American Psychiatric Association. 2025. *Diagnostic and statistical manual of mental disorders, fifth edition, text revision (DSM-5-TR®)*. American Psychiatric Association Publishing, Arlington, TX.

Azizi Ab Aziz, Michel C. A. Klein, and Jan Treur. 2009. An agent model of temporal dynamics in relapse and recurrence in depression. In *Next-Generation Applied Intelligence, 22nd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2009, Tainan, Taiwan, June 24-27, 2009. Proceedings*, volume 5579 of *Lecture Notes in Computer Science*, pages 36–45. Springer.

Azizi Ab Aziz, Michel C. A. Klein, and Jan Treur. 2010. An integrative ambient agent model for unipolar depression relapse prevention. *J. Ambient Intell. Smart Environ.*, 2(1):5–20.

AT Beck. 1967. Depression: clinical, experimental, and theoretical aspects. *Harper & Row google schola*, 2:103–113.

Stephanie L Burcusa and William G Iacono. 2007. Risk for recurrence in depression. *Clinical psychology review*, 27(8):959–985.

Gillian Butler and Andrew Mathews. 1983. Cognitive processes in anxiety. *Advances in behaviour research and therapy*, 5(1):51–62.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485–1497, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

P L de Zwart, B F Jeronimus, and P de Jonge. 2019. Empirical evidence for definitions of episode, remission, recovery, relapse and recurrence in depression: a systematic review. *Epidemiol. Psychiatr. Sci.*, 28(5):544–562.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Lesley Dwyer. 2019. Modelling Depression Recurrence Through Analysis of Electronic Health Records. Technical report, City, University of London.

Ellen Frank, Robert F. Prien, Robin B. Jarrett, Martin B. Keller, David J. Kupfer, Philip W. Lavori, A. John Rush, and Myrna M. Weissman. 1991. Conceptualization and rationale for consensus definitions of terms in major depressive disorder: Remission, recovery, relapse, and recurrence. *Archives of General Psychiatry*, 48(9):851–855.

Felan Carlo C. Garcia, Ayumi Hirao, Aran Tajika, Toshi A. Furukawa, Kazushi Ikeda, and Junichiro Yoshimoto. 2021. Leveraging longitudinal lifelog data using survival models for predicting risk of relapse among patients with depression in remission. In *43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC 2021, Mexico, November 1-5, 2021*, pages 2455–2458. IEEE.

Ian H. Gotlib and Jutta Joormann. 2010. Cognition and depression: Current status and future directions. *Annual Review of Clinical Psychology*, 6:285–312.

Ian H Gotlib and C Douglas McCann. 1984. Construct accessibility and depression: An examination of cognitive and affective factors. *J. Pers. Soc. Psychol.*, 47(2):427–439.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and Ahmad Al-Dahle. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Rashmi Gupta and Bhoomika R Kar. 2008. Interpretive bias: Indicators of cognitive vulnerability to depression. *Ger. J. Psychiatr.*, 11(3):98–102.

Rashmi Gupta and Bhoomika R Kar. 2012. Attention and memory biases as stable abnormalities among currently depressed and currently remitted individuals with unipolar depression. *Front. Psychiatry*, 3:99.

E Clare Harris and Brian Barraclough. 1997. Suicide as an outcome for mental disorders. *Br. J. Psychiatry*, 170(3):205–228.

Amey Hengle, Atharva Kulkarni, Shantanu Deepak Patankar, Madhumitha Chandrasekaran, Sneha D'silva, Jemima S. Jacob, and Rashmi Gupta. 2024. Still not quite there! evaluating large language models for comorbid mental health diagnosis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16698–16721, Miami, Florida, USA. Association for Computational Linguistics.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *CoRR*, abs/1904.05342.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.

D J Kupfer, E Frank, and J Wamhoff. 1996. Mood disorders: Update on prevention of recurrence. In C Mundt, M J Goldstein, K Hahlweg, and P Fiedler, editors, *Interpersonal factors in the origin and course of affective disorders*, pages 289–302. Gaskell/Royal College of Psychiatrists.

Joelle LeMoult and Ian H. Gotlib. 2019. Depression: A cognitive perspective. *Clinical Psychology Review*, 69:51–66.

G G Lloyd and W A Lishman. 1975. Effect of depression on the speed of recall of pleasant and unpleasant experiences. *Psychol. Med.*, 5(2):173–180.

David E. Losada, Fabio Crestani, and Javier Parapar. 2018. Overview of erisk: Early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings*, volume 11018 of *Lecture Notes in Computer Science*, pages 343–361. Springer.

Daniel M Low, Laurie Rumker, John Torous, Guillermo Cecchi, Satrajit S Ghosh, and Tanya Talkar. 2020. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of medical Internet research*, 22(10):e22635.

Christopher Lucasius, Mai Ali, Deepa Kundur, Marco Battaglia, Peter Szatmari, and John Strauss. 2024. Prediction of relapse in adolescent depression using fusion of video and speech data (short paper). In *Proceedings of Machine Learning for Cognitive and Mental Health Workshop (ML4CMH 2024) Co-located with the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI 2024), Vancouver, BC, Canada, February 26th, 2024*, volume 3649 of *CEUR Workshop Proceedings*, pages 112–117. CEUR-WS.org.

Munn-Sann Lye, Yin-Yee Tey, Yin-Sim Tor, Aisya Farhana Shahabudin, Normala Ibrahim, King-Hwa Ling, Johnson Stanslas, Su-Peng Loh, Rozita Rosli, Khairul Aiman Lokman, Ibrahim Mohammed Badamasi, Asraa Faris-Aldoghachi, and Nurul Asyikin Abdul Razak. 2020. Predictors of recurrence of major depressive disorder. *PLoS One*, 15(3):e0230363.

F Matcham, E Carr, N Meyer, K M White, C Oetzmann, D Leightley, F Lamers, S Siddi, N Cummins, P Annas, G de Girolamo, J M Haro, G Lavelle, Q Li, F Lombardini, D C Mohr, V A Narayan, B W H J Penninx, M Coromina, G Riquelme Alacid, S K Simblett, R Nica, T Wykes, J C Brasen, I Myin-Germeys, R J B Dobson, A A Folarin, Y Ranjan, Z Rashid, J Dineley, S Vairavan, M Hotopf, and RADAR-CNS consortium. 2024. The relationship between wearable-derived sleep features and relapse in major depressive disorder. *J. Affect. Disord.*, 363:90–98.

Augusto R. Mendes and Helena Caseli. 2024. Identifying fine-grained depression signs in social media posts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8594–8604, Torino, Italia. ELRA and ICCL.

Andrew S Moriarty, Nicholas Meader, Kym Ie Snell, Richard D Riley, Lewis W Paton, Carolyn A Chew-Graham, Simon Gilbody, Rachel Churchill, Robert S Phillips, Shehzad Ali, and Dean McMillan. 2021. Prognostic models for predicting relapse or recurrence of major depressive disorder in adults. *Cochrane Database Syst. Rev.*, 5(5):CD013491.

Andrew S Moriarty, Nicholas Meader, Kym IE Snell, Richard D Riley, Lewis W Paton, Sarah Dawson, Jessica Hendon, Carolyn A Chew-Graham, Simon Gilbody, Rachel Churchill, et al. 2022. Predicting relapse or recurrence of depression: systematic review of prognostic models. *The British Journal of Psychiatry*, 221(2):448–458.

Muhammad Muzammel, Alice Othmani, Himadri Mukherjee, and Hanan Salam. 2021. Identification of signs of depression relapse using audio-visual cues: A preliminary study. In *34th IEEE International Symposium on Computer-Based Medical Systems, CBMS 2021, Aveiro, Portugal, June 7-9, 2021*, pages 62–67. IEEE.

Zhi Nie, Pinghua Gong, and Jieping Ye. 2016. Predict risk of relapse for patients with multiple stages of treatment of depression. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1795–1804. ACM.

Susan Nolen-Hoeksema. 1991. Responses to depression and their effects on the duration of depressive episodes. *J. Abnorm. Psychol.*, 100(4):569–582.

Rogerio Salvini, Rodrigo da Silva Dias, Beny Lafer, and Inês Dutra. 2015. A multi-relational model for depression relapse in patients with bipolar disorder. In *MEDINFO 2015: eHealth-enabled Health - Proceedings of the 15th World Congress on Health and Biomedical Informatics, São Paulo, Brazil, 19-23 August 2015*, volume 216 of *Studies in Health Technology and Informatics*, pages 741–745. IOS Press.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Gemma Team. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022. Identifying moments of change from longitudinal user text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4647–4660, Dublin, Ireland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Endre Visted, Jon Vøllestad, Morten Birkeland Nielsen, and Elisabeth Schanche. 2018. Emotion regulation in current and remitted depression: A systematic review and meta-analysis. *Front. Psychol.*, 9.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, and Huan Lin. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Wenyi Yin, Chenghao Yu, Pianran Wu, Wenxuan Jiang, Youzhe Liu, Tianqi Ren, and Weihui Dai. 2022. An intelligent mobile system for monitoring relapse of depression. In *Computer Supported Cooperative Work and Social Computing - 17th CCF Conference, ChineseCSCW 2022, Taiyuan, China, November 25-27, 2022, Revised Selected Papers, Part I*, volume 1681 of *Communications in Computer and Information Science*, pages 182–193. Springer.

A A Note on Terminologies

Depression relapse literature seems to have no consensus regarding the exact definitions of the terms related to the phases a person undergoes during the treatment of depression as depicted in Figure 1. Frank et al. (1991) attempted to standardize the definitions but left it open for empirical validations. Later de Zwart et al. (2019) in their systemic review of relevant literature found that the only consensus reached till now is regarding the minimum duration required for defining an episode.

Thus in this work we take the liberty to use the term *relapse* as an umbrella term to refer to both *relapse* as well as *recurrence*. For similar reasons we use the term *remission* to refer to both *remission* as well as *recovery*. More concretely, we consider a user is in remission when we come across any post in which she mentions feeling better and/or asymptomatic. Conversely, we consider a user in relapse when she explicitly mentions about the return of symptoms and feeling worse than before. Moreover, this simplifying assumption also makes it easier to work with social media posts.

B Pattern List

B.1 Depression

Following is a partial list; the full list can be accessed here[5].

`“depression”, “major depression”, “depressive disorder”, “dysthymia”, “premenstrual dysphoric disorder”, “chronic depression”, “clinical depression”, “depressions”, “depressive”, “depressive illness”, “depressive neuroses”, “depressive neurosis”, “depresssion”, “disorder depressive”, “disorder dysphoric premenstrual”, “disorder premenstrual dysphoric”, “disthymia”, “distimea”, “distimia”, “double depression”, “dpression”, “dysthymia”, “dysthymia disorder”, “dysthymic”, “dysthymic dis”, “dysthymic disorder”, “dystimea”, “dystimia”, “late luteal phase dysphoric disorder”, “llpdd”, “major depression”, “major depression disorder”, “major depressive dis”, “mdd”, “major depressive illness”, “premenstrual dysphoric syndrome”, “reactive depression”, “recurrent depressive disorder”, “sever depression”, “severe depression”, “severe depressive”, “unipolar depression”, “unipolar depressive disorder”, “unipolar major depression”, “Mild mood disturbance”, “Borderline clinical depression”, “Moderate depression”, “Extreme depression”, “Minimal depression”, “Mild depression”, “Moderately severe depression”, “mildly depressed”, “moderately depressed”, “mildly depressive”, “moderately depressive”`

B.2 Diagnosed

`“diagnose”, “diagnosed”, “diagnosing”, “diagnosis”, “diagnoses”, “prescribed”, “prescribe”, “prescription”, “treated”, “treatment”, “treating”`

C Prompts

Llama 3.2 3B

```
system_prompt = """
### Task: Please help me find the
     following information only about the
     author from its post.
[output format]:
Question 1. What is the main topic of
     the post?
    Answer: [Main topic identified from
     the post content.]
    Reason: [Reasoning based on how the
     main topic is inferred from the post
     content.]
Question 2. Does the author mention any
     mental health conditions if they are
     dealing with (only author's mental
     health conditions)?
    Answer: mental health condition1,
     mental health condition2, or Not
     mentioned
    Reason: [Reasoning based on how
     these mental health conditions are
     inferred from the post content.]
Question 3. Give mental health symptoms
     the author has experienced?
    Answer: [List of symptoms] or Not
     mentioned.
    Reason: [Reasoning based on how
     these symptoms are inferred from the
     post content.]
Question 4. Has the author shared any
     life experiences or events that have
     positively impacted their mental
     health?
    Answer: [List of experiences/events]
     or Not mentioned
    Reason: [Reasoning based on how
     these events are linked to the
     authors mental health.]
Question 5. Has the author shared any
     life experiences or events that have
     negatively impacted their mental
     health?
    Answer: [List of experiences/events]
     or Not mentioned
    Reason: [Reasoning based on how
     these events are linked to the
     authors mental health.]
```

[5]Details available at https://github.com/saprativa/ReDepress

```
Question 6. Does the author state
     whether they have been diagnosed
     with these conditions?
    Answer: mental health
     condition1[Yes/No or Not mentioned],
     mental health condition2[Yes/No or
     Not mentioned]
    Reason: [Reasoning based on how the
     diagnosis is inferred from the post
     content.]
Question 7. Was the diagnosis of author
     for mental health conditions very
     recent from the post creation date?
    Answer: mental health
     condition1[Yes/No or Not mentioned],
     mental health condition2[Yes/No or
     Not mentioned]
    Reason: [Reasoning based on how the
     recency of the diagnosis is inferred
     from the post content.]
Question 8. Has the author mentioned any
     therapy, counseling, or support
     programs they are undergoing for
     mental health conditions?
    Answer: Yes[Type of therapy/support
     program] or Not mentioned
    Reason: [Reasoning based on how
     therapy or counseling is inferred
     from the post content.]
Question 9. Has the author mentioned any
     medication they are taking for
     mental health conditions?
    Answer: Yes[Name or type of
     medication] or Not mentioned
    Reason: [Reasoning based on how
     medication use is inferred from the
     post content.]
Question 10. Has the author mentioned
     recovery of any mental health
     condition?
    Answer: mental health
     condition1[Yes/No or Not mentioned],
     mental health condition2[Yes/No or
     Not mentioned]
    Reason: [Reasoning based on how
     recovery is inferred from the post
     content.]
Question 11. Has the author mentioned
     supportive factors or positive
     influences for the recovery in their
     mental health condition?
    Answer: [List of supportive factors
     or positive influences] or Not
     mentioned
    Reason: [Reasoning based on how the
     Supportive factors or positive
     influences for recovery are inferred
     from the post content.]
Question 12. Has the author experienced
     a relapse of any mental health
     condition?
    Answer: mental health
     condition1[Yes/No or Not mentioned],
     mental health condition2[Yes/No or
     Not mentioned]
    Reason: [Reasoning based on how the
     relapse is inferred from the post
     content.]
```

```
Question 13. Has the author mentioned
     triggers or reasons for the relapse?
    Answer: [List of triggers] or Not
     mentioned
    Reason: [Reasoning based on how the
     triggers or reasons for relapse are
     inferred from the post content.]
[End of output format]
Please provide the answers and reasoning
     of these 13 questions only and
     exactly as in the [output format]."""
```

Llama 3.1 70B

```
system_prompt = """
### Task: Please help me find the
     following information only about the
     author from its post.
[output format]:
Question 1. What is the main topic of
     the post?
    Answer: [Main topic identified from
     the post content.]
    Reason: [Reasoning based on how the
     main topic is inferred from the post
     content.]
    Post Part: [Give that part of post
     content.]
Question 2. Does the author mention any
     mental health conditions if they are
     dealing with (only author's mental
     health conditions)?
    Answer: mental health condition1,
     mental health condition2, or Not
     mentioned
    Reason: [Reasoning based on how
     these mental health conditions are
     inferred from the post content.]
    Post Part: [Give that part of post
     content.]
Question 3. Give mental health symptoms
     the author is currently experiencing?
    Answer: [List of symptoms] or Not
     mentioned.
    Reason: [Reasoning based on how
     these symptoms are inferred from the
     post content.]
    Post Part: [Give that part of post
     content.]
Question 4. Has the author shared any
     life experiences or events that have
     positively impacted their mental
     health?
    Answer: [List of positive
     experiences or events] or Not
     mentioned
    Reason: [Reasoning based on how
     these positive experiences linked to
     the authors mental health.]
    Post Part: [Give that part of post
     content.]
Question 5. Has the author shared any
     life experiences or events that have
     negatively impacted their mental
     health?
    Answer: [List of negative
     experiences or events] or Not
```

```
     mentioned
     Reason: [Reasoning based on how
     these negative experiences linked to
     the authors mental health.]
     Post Part: [Give that part of post
     content.]
Question 6. Does the author state
     whether they have been diagnosed
     with these conditions?
     Answer: mental health
     condition1[Yes/No or Not mentioned],
     mental health condition2[Yes/No or
     Not mentioned]
     Reason: [Reasoning based on how the
     diagnosis is inferred from the post
     content.]
     Post Part: [Give that part of post
     content.]
Question 7. Was the diagnosis of author
     for mental health conditions very
     recent from the post creation date?
     Answer: mental health
     condition1[Yes/No or Not mentioned],
     mental health condition2[Yes/No or
     Not mentioned]
     Reason: [Reasoning based on how the
     recency of the diagnosis is inferred
     from the post content.]
     Post Part: [Give that part of post
     content.]
Question 8. Has the author mentioned any
     therapy, counseling, or support
     programs they are currently
     undergoing for mental health
     condition?
     Answer: Yes[Type of therapy/support
     program] or Not mentioned
     Reason: [Reasoning based on how
     therapy, counseling, or support
     programs is inferred from the post
     content.]
     Post Part: [Give that part of post
     content.]
Question 9. Has the author mentioned any
     medication they are currently taking
     for mental health conditions?
     Answer: Yes[Name or type of
     medication] or Not mentioned
     Reason: [Reasoning based on how
     medication use is inferred from the
     post content.]
     Post Part: [Give that part of post
     content.]
Question 10: Has the author mentioned
     any current changes to their
     medication regimen (e.g., stopping
     medication, starting a new
     medication, changing medication,
     increasing dosage, decreasing
     dosage)?
     Answer: Yes [description of change]
     or Not mentioned
     Reason: [Reasoning based on how
     these changes are inferred from the
     post content.]
     Post Part: [Give that part of post
     content.]
Question 11. Is the author currently
     experiencing (not in past) recovery
     or remission of any mental health
     condition?
     Answer: mental health
     condition1[Yes/No or Not mentioned],
     mental health condition2[Yes/No or
     Not mentioned]
     Reason: [Reasoning based on how
     recovery or remission is inferred
     from the post content.]
     Post Part: [Give that part of post
     content.]
Question 12. Has the author mentioned
     any current supportive factors or
     positive influences aiding their
     recovery or remission of their
     mental health condition?
     Answer: [List of supportive factors
     or positive influences] or Not
     mentioned
     Reason: [Reasoning based on how the
     Supportive factors or positive
     influences are inferred from the
     post content.]
     Post Part: [Give that part of post
     content.]
Question 13. Is the author currently
     experiencing (not in past) relapse
     or recurrence of any mental health
     condition?
     Answer: mental health
     condition1[Yes/No or Not mentioned],
     mental health condition2[Yes/No or
     Not mentioned]
     Reason: [Reasoning based on how the
     current relapse or recurrence is
     inferred from the post content.]
     Post Part: [Give that part of post
     content.]
Question 14. Has the author mentioned
     any triggers or reasons for the
     current relapse or recurrence?
     Answer: [List of triggers or
     reasons] or Not mentioned
     Reason: [Reasoning based on how the
     triggers or reasons are inferred
     from the post content.]
     Post Part: [Give that part of post
     content.]
[End of output format]
Please provide the answers and reasoning
     of these 14 questions only and
     exactly as in the [output format].
     Answer only what is asked and do not
     provide any extra questions,
     information, explanations."""
```

D Annotation Guidelines

Reproduced below is an abridged version of the annotation guidelines for the sake of brevity.

Each one of you will be provided with a set of excel files. Each file corresponds to one Reddit user and has the following columns:

1. post_id: A unique alphanumeric identifier for

the post.

2. author: Name of the Reddit user (mapped to a unique number for privacy preservation).
3. created_readable: The date and time when the post was created.
4. title: The title of the post.
5. selftext: The body of the post.
6. Remission/Relapse: A drop-down with the following options:
   (a) Remission
   (b) Recovery
   (c) Relapse
   (d) Other
7. Memory bias: A drop-down with the following options:
   (a) Positive
   (b) Negative
   (c) No bias
8. Attention bias: A drop-down with the following options:
   (a) Positive
   (b) Negative
   (c) No bias
9. Interpretation bias: A drop-down with the following options:
   (a) Positive
   (b) Negative
   (c) No bias
10. Rumination: A drop-down with the following options:
    (a) Brooding
    (b) Reflection
    (c) No rumination

- Each row in a file corresponds to a separate post. The posts are arranged in chronological order starting from old to recent.
- Your task is to look at each post title and post body and then annotate the columns Remission/Relapse, Memory bias, Attention bias, Interpretation bias and Rumination.
- Take the Dominant Label Approach: Read the entire post and decide which type of Label is most prominent overall. Select only one label.

Summary of Terms

- **Rumination (Thinking Over and Over)**
  - **No Rumination:** The person doesn't keep thinking about events.
  - **Brooding:** The person is stuck in negative thoughts (e.g., "Why do bad things always happen to me?").
  - **Reflection:** The person analyzes past events to understand and learn (e.g., "What can I do differently next time?").
- **Memory Bias (What We Remember)**
  - **Positive:** The person remembers things in a happy or hopeful way.
  - **Negative:** The person remembers things in a sad or pessimistic way.
  - **No bias:** The person recalls things in a neutral way.
- **Attention Bias (What We Notice More at Present)**
  - **Positive:** The person focuses more on the good in a situation.
  - **Negative:** The person focuses more on the bad.
  - **No bias:** The person gives equal attention to good and bad things.
- **Interpretation Bias (How We Interpret Events)**
  - **Positive:** Interprets ambiguous situations in an optimistic way.
  - **Negative:** Interprets ambiguous situations as negative or unfavorable.
  - **No bias:** Interprets ambiguous situations realistically, without excessive optimism or pessimism.

E Examples of Remission and Relapse Posts

A typical remission post looks like the following (paraphrased for privacy reasons):

> *I've been dealing with depression for a decade, but over the past few months, things have started to improve. This year has still been rough and delayed a lot of my goals by at least half a year, but my emotional responses to setbacks have*

become more manageable. For instance, I recently got a fine I can't afford, but instead of spiraling, I just thought, "this is frustrating, but I'll handle it." Even though my boss can be overbearing, I just let it go and remind myself I only interact with her once a week.

It's strange not to feel overwhelmed by everyday stressors. Is this how mentally healthy people normally cope? It's kind of unsettling to respond to difficulties without the usual emotional chaos. That internal voice of depression and anxiety still tries to sound the alarm, like "you should be panicking right now," but it's becoming easier to quiet that voice—even if it still feels oddly unfamiliar.

While, a relapse post looks like this:

A bit of context: I began therapy and medication about three years ago to address depression, PTSD, and social anxiety. It took some trial and error, but eventually we found a combination that worked—Seroquel XR (twice a day), Viibryd in the morning, and Klonopin as needed.

Things improved significantly over the next two years, and we gradually reduced my medications to just a nightly dose of Seroquel XR. Around that time, we shifted focus to managing my ADHD. I tried Concerta without much success, and some doses of Adderall XR that worked mildly. I finally voiced that something still didn't feel quite right.

We then switched from Adderall XR to IR for better control. For the past three months, I've been taking 15mg three times a day. It's been helping a lot with focus—sometimes I even skip the third dose.

But about two weeks ago, I began feeling a low mood creeping back in. Now it's hit hard—I've been emotionally withdrawn from my family, eating irregularly, and isolating myself after work. I mostly distract myself with games, movies, or music while tuning everything else out.

I'm wondering if the Adderall could be contributing to this downturn. It's been effective for my ADHD, and I'd hate to start the search for a new med again. I'm not entirely opposed to restarting an SSRI, but I remember feeling emotionally flat when I was on Viibryd, even though it kept the depression in check.

F Distribution of Relapse Cases

Relapse rates of 50% after first, 80% after second and 90% after third episodes of depression are based on clinical studies. These clinical studies focus specifically on relapse rates in clinical settings, often involving hospitalized patients or those actively engaged in clinical treatments. Our dataset, derived from social media posts, reflects self-disclosed mental health statuses from a non-clinical, online environment. Consequently, our reported relapse rate (approximately 40% in the annotated dataset: 83 relapse vs. 121 non-relapse cases) emerges organically from users' self-disclosures and subsequent clinical annotations. This rate, while similar to clinical findings, is also expected to deviate from clinical studies due to the fact that social media users are not obligated to self-report every recovery and relapse. And this is true for all social media based mental health datasets including ours. We discuss in detail the implications of using social media data for depression relapse prediction in the limitations section of our paper.

Notwithstanding the above, we took extreme care during the manual filtering stages of our data curation pipeline to include only those cases which were corresponding to the first relapse after remission from depression (as much as was possible with social media self disclosures) to get unbiased data.

G Normality Test

To determine the appropriate statistical tests for Table 3, we conducted normality tests on the variables. As shown in Table 11, the Shapiro-Wilk test p-values for all features were less than 0.05, indicating significant deviations from normality. Given this non-normal distribution, we used the non-parametric Mann-Whitney U test as the primary method to assess group differences.

H Average

The averaging process ensures a balanced representation of cognitive parameter scores by incor-

Feature	Shapiro-Wilk stat	Shapiro-Wilk p-value
Attention Bias	0.967189	0.000109
Interpretation Bias	0.963059	0.000036
Memory Bias	0.963968	0.000046
Rumination	0.983268	0.015915

Table 11: Shapiro-Wilk normality test results for cognitive features. All features significantly deviate from normality, except Rumination which shows marginal deviation.

Feature	Label	Mapped Value
Memory Bias	Positive	1
	Negative	-1
	No bias	0
Attention Bias	Positive	1
	Negative	-1
	No bias	0
Interpretation Bias	Positive	1
	Negative	-1
	No bias	0
Rumination	Reflection	1
	Brooding	-1
	No Rumination	0

Table 12: Mapping values for different cognitive parameters.

porating multiple human annotations. For each parameter, individual annotators provided qualitative labels, which were mapped to numerical values based on predefined mappings (see Table 12). The final score for each of dimensions was computed as the average of available human annotations. In the majority setting, only the majority agreeing annotations are considered for averaging.

I Hyperparameters

For traditional machine learning models, we performed grid search over Random Forest, Gradient Boosting, XGBoost, Logistic Regression, SVM, KNN, and Neural Network classifiers trained on aggregated cognitive features. The optimal settings were chosen using 5-fold cross-validation based on F1-score. Table 13 reports the best hyperparameters along with the corresponding feature aggregation strategies and class balancing methods (random oversampling or undersampling).

For transformer-based temporal models, we implemented a transformer encoder over post-level embeddings (MentalBERT, MentalRoBERTa, MPNet) enriched with cognitive markers. A wide grid of parameters was explored, including hidden dimension size, number of heads, number of layers, dropout, learning rate, batch size, weight decay, and early stopping patience. The best settings for each embedding configuration (text only, text + markers, joint embeddings) were identified via early stopping on the validation set and are summarized in Table 14.

J Information Imbalance

Table 15 provides the full ablation results for each type of bias and aggregation (min, max, mean, median) for the Information Imbalance metric.

K Cognitive Classification Prompt

```
# User prompt with post content
user_prompt = f"### Post creation date:
    {row['created_readable']} \n###
    Post: {row['title']}\n
    {row['selftext']}\nEnd of the post"

# System prompt for cognitive dimension
    classification
system_prompt = """
### Task: Please help me classify the
    following Reddit post into cognitive
    dimensions.
Focus only on the authors perspective
    expressed in the post.
Assign exactly one label per dimension.

### output format:
Attention Bias:
    Answer: positive / negative / no bias
    Reason: [Reasoning based on how the
    bias is inferred from the post
    content.]

Memory Bias:
    Answer: positive / negative / no bias
    Reason: [Reasoning based on how the
    bias is inferred from the post
    content.]

Interpretation Bias:
    Answer: positive / negative / no bias
    Reason: [Reasoning based on how the
    bias is inferred from the post
    content.]

Rumination:
    Answer: reflection / brooding / no
    rumination
    Reason: [Reasoning based on how
    rumination type is inferred from the
    post content.]

[End of output format]

Please provide the answers and reasoning
    of these 4 dimensions only and
    exactly as in the output format.
Answer only what is asked and do not
    provide any extra questions,
    information, or explanations.
"""
```

Model	Aggregations	Num Features	Balancing	Best Hyperparameters
RandomForest	mean	4	random_oversample	clf__bootstrap=True, clf__max_depth=2, clf__min_samples_leaf=4, clf__min_samples_split=2, clf__n_estimators=100
GradientBoosting	mean+min+max+median	16	smote	clf__learning_rate=0.1, clf__max_depth=3, clf__min_samples_leaf=2, clf__min_samples_split=2, clf__n_estimators=50, clf__subsample=0.4
XGBoost	mean+min	8	smote	clf__colsample_bytree=0.8, clf__learning_rate=0.01, clf__max_depth=3, clf__n_estimators=50, clf__subsample=0.8
LogisticRegression	mean+min	8	none	clf__C=1, clf__penalty=l1, clf__class_weight=balanced,
SVM	mean+max	8	smote	clf__C=0.001, clf__kernel=rbf, clf__degree=2, clf__gamma=0.01, clf__shrinking=True, clf__class_weight=None
KNN	mean+min+median	12	random_oversample	clf__n_neighbors=5, clf__algorithm=auto, clf__leaf_size=10, clf__metric=euclidean, clf__p=1, clf__weights=uniform
NeuralNetwork	min+max+median	12	random_oversample	clf__hidden_layer_sizes=(128,64,32), clf__activation=tanh, clf__alpha=1e-07, clf__learning_rate_init=1e-05, clf__momentum=0.9, clf__solver=adam

Table 13: Best hyperparameter settings, feature aggregation strategies, and balancing methods for different models.

Embedding Model	Configuration	Balancing	Best Hyperparameters
mpnet	No CM	No Sampling	d_model=256, nhead=8, num_layers=2, dropout=0.0 learning_rate=0.001, batch_size=16, weight_decay=1e-4, epochs=100, early_stopping_patience=10
mentalbert	No CM	RandomUnderSampler	d_model=512, nhead=8, num_layers=2, dropout=0.5 learning_rate=0.0001, batch_size=32, weight_decay=1e-4, epochs=100, early_stopping_patience=10
mentalroberta	No CM	No Sampling	d_model=128, nhead=8, num_layers=2, dropout=0.0 learning_rate=0.0001, batch_size=16, weight_decay=1e-5, epochs=100, early_stopping_patience=20
mpnet	CM	RandomOverSampler	d_model=256, nhead=8, num_layers=2, dropout=0.0 learning_rate=0.001, batch_size=32, weight_decay=1e-4, epochs=100, early_stopping_patience=10
mentalbert	CM	RandomUnderSampler	d_model=128, nhead=4, num_layers=4, dropout=0.0 learning_rate=0.0001, batch_size=32, weight_decay=1e-4, epochs=100, early_stopping_patience=10
mentalroberta	CM	No Sampling	d_model=512, nhead=8, num_layers=4, dropout=0.0 learning_rate=0.0001, batch_size=16, weight_decay=1e-5, epochs=100, early_stopping_patience=20
mpnet	CM-Emb	RandomUnderSampler	d_model=128, nhead=4, num_layers=2, dropout=0.0 learning_rate=0.001, batch_size=8, weight_decay=1e-4, epochs=100, early_stopping_patience=15
mentalbert	CM-Emb	RandomUnderSampler	d_model=256, nhead=4, num_layers=2, dropout=0.0 learning_rate=0.001, batch_size=16, weight_decay=1e-5, epochs=100, early_stopping_patience=20
mentalroberta	CM-Emb	RandomUnderSampler	d_model=512, nhead=8, num_layers=2, dropout=0.0 learning_rate=0.0001, batch_size=32, weight_decay=1e-5, epochs=100, early_stopping_patience=15

Table 14: Best hyperparameter settings and balancing strategies for transformer-based temporal sequence models under different input configurations. Classification was performed using the last hidden state. Configuration short forms: No CM = No Cognitive Markers, CM = Cognitive Markers, CM-Emb = Cognitive Markers Embedding.

Feature	Statistic	$\mathbf{II}_{\text{Ablated}}$	$\mathbf{\Delta}$ II	Interpretation
Rumination	Mean	0.611	+0.014	Most important: removing increases II
Interpretation Bias	Mean	0.605	+0.008	Some importance
Attention Bias	Mean	0.605	+0.007	Some importance
Attention Bias	Median	0.602	+0.004	Some importance
Memory Bias	Mean	0.598	-0.000	Neutral/redundant
Rumination	Median	0.595	-0.003	Neutral/redundant
Rumination	Min	0.595	-0.003	Neutral/redundant
Interpretation Bias	Min	0.593	-0.005	Neutral/redundant
Memory Bias	Max	0.588	-0.010	Slightly redundant
Interpretation Bias	Max	0.587	-0.011	Slightly redundant
Memory Bias	Median	0.586	-0.011	Slightly redundant
Memory Bias	Min	0.584	-0.014	Slightly redundant
Attention Bias	Min	0.583	-0.015	Slightly redundant
Rumination	Max	0.578	-0.020	Slightly redundant
Attention Bias	Max	0.578	-0.020	Slightly redundant
Interpretation Bias	Median	0.573	-0.025	Slightly redundant

Table 15: Information Imbalance ablation analysis for each feature. Positive ΔII indicates the feature is highly informative for relapse prediction; negative or near-zero values indicate redundancy or low unique contribution.