CTMC-LSTM: A Markov-Based Hybrid Model for Depression Severity Modelling with an Expert-Annotated Longitudinal Dataset

Anonymous ACL submission

Abstract

Depression severity is a clinically important indicator for assessing mental health status and guiding treatment, yet it remains challenging to infer reliably from user-generated text. Existing NLP research has largely focused on binary depression detection or isolated symptom classification, with limited attention to modelling severity progression over time. We present the DepSy Severity dataset, a novel resource for modelling depression severity from social media, fully annotated by psychologists with weekly severity scores for users who self-report a depression diagnosis. To address this task, we propose CTMC-LSTM, a hybrid model that combines LSTM-based predictions with temporal dynamics modelled through a Continuous-Time Markov Chain (CTMC), capturing user-level severity progression over time. We frame severity estimation as both regression and multi-class classification, evaluating a range of architectures and feature combinations. Our experiments show that models incorporating structured features outperform text-only baselines, and the CTMC-LSTM model yields the highest performance in severity classification, particularly for underrepresented classes such as moderate and severe depression. These results highlight the importance of integrating temporal context for robust mental health modelling.

1 Introduction

Depression is a major global health concern, affecting hundreds of millions of individuals and contributing to substantial emotional, social, and economic burdens. The severity of depressive episodes—ranging from mild emotional distress to severe impairment—plays a critical role in shaping clinical outcomes and determining the urgency and type of intervention needed. Despite growing efforts in computational mental health, the task of modelling depression severity remains largely overlooked in natural language processing (NLP) research.

While prior work has primarily relied on structured clinical data such as electronic health records and diagnostic tools (Kim et al., 2020; Pradier et al., 2021; Mesbah et al., 2021; de Oliveira et al., 2021; Ignashina et al., 2025), social media offers a complementary perspective by capturing spontaneous, real-world expressions of psychological states. These longitudinal and context-rich narratives provide an opportunity to analyse how depression progresses over time. Yet, building robust severity prediction models remains a challenge due to the absence of publicly available, longitudinal datasets that are rigorously annotated by domain experts.

Most existing NLP studies in this space focus on binary classification—detecting whether a person is depressed or not—or identifying transitions to crisis states like suicidal ideation (De Choudhury et al., 2016; Gong et al., 2019; Sawhney et al., 2020; Kour and Gupta, 2022; Baghdadi et al., 2022; Khafaga et al., 2023; Adarsh et al., 2023). Far fewer studies attempt to quantify or track how severe an individual's depressive state is over time. This represents a critical gap in mental health NLP, where timely detection of increasing severity could improve early intervention and support strategies.

This study aims to fill the identified gap by enhancing early detection and intervention strategies through improved datasets and model development. Our key contributions can be summarised as follows:

- The first English dataset of depression severity in a longitudinal format using textual posts of depressed users that is **fully annotated by psychologists**[1].

[1]The DepSy severity dataset and all codes will be made available upon paper acceptance

- Empirical work comparing multiple predictive models (based on LSTM, BERT, RoBERTa, and MentalBERT) built using our dataset for the task of predicting depression severity score from posts chunks (1-week posts).
- CTMC-LSTM: a novel hybrid model, which integrates temporal dynamics through Markov chains.

2 Related Work

Research on mental health from social media has primarily followed two directions: severity classification and longitudinal monitoring. Severity classification focuses on assigning discrete depression levels at specific points in time, while longitudinal approaches aim to capture temporal dynamics by modelling changes in mental health status across user timelines.

The eRisk shared tasks (Losada et al., 2017, 2018, 2019, 2020, 2021) provide benchmark datasets for user-level detection of depression and related conditions using Reddit data. While later editions included severity labels, these are typically based on a subset of posts, with the remainder left unreviewed. All Posts are given a label based on one post annotation, and annotations lack clinical expert involvement, limiting their use for fine-grained severity tracking. Despite these limitations, eRisk remains a key benchmark for early detection research.

Beyond the eRisk shared tasks, a growing body of work has focused on depression severity classification using social media, with increasing attention to clinical alignment in both annotation and modelling. Coppersmith et al. (2014) analysed linguistic patterns preceding suicide attempts using word embeddings and LSTMs. Gong et al. (2019) explored PHQ-9 scores from health records to examine links between general depressive symptoms and suicidal ideation. Several datasets have since been introduced to support severity classification. DepSign (Sampath and Durairaj, 2022) categorises posts into four severity levels and uses traditional models with data augmentation to mitigate imbalance. The DsD dataset (Naseem et al., 2022) introduces ordinal labels based on clinical guidelines and employs a hierarchical attention model. DepTweet (Kabir et al., 2023) builds on DSM-5 and PHQ-9 to label over tweets with severity levels and confidence scores. Further, Zhang et al. (2023) proposed a sentiment-guided transformer trained on DsD and DepSign, and Ahmed et al. (2024) used an ensemble of BERT variants to classify Reddit posts by severity. Recent work has also leveraged sentence embeddings (Qasim et al., 2025) and sentiment signals (Ogunleye et al., 2024) to enhance severity detection.

In parallel, longitudinal approaches have aimed to capture changes in mental health status over time. Sawhney et al. (2020) proposed a time-aware transformer for suicidal risk detection using tweet sequences, extended in Sawhney et al. (2021) by modelling emotional progression via Plutchik's wheel. Chiu et al. (2021) applied a multi-modal model on Instagram combining image, text, and behavioural data to track depressive patterns. Tsakalidis et al. (2022b) introduced a dataset of annotated user timelines to detect "moments of change" using sequential modelling. More recent approaches incorporate temporal structure more explicitly: Hills et al. (2023, 2024) added a Hawkes process-inspired layer to a hierarchical transformer, while Song et al. (2024) proposed a hybrid TH-VAE and LLaMA-2 model to summarise mental health trajectories, evaluated against expert and clinician assessments.

Most severity classification studies rely on post-level labels, assigning a severity score to individual posts in isolation. However, clinical assessment of depression severity depends on observing symptom patterns over time, making single-post labelling an unreliable proxy. At the same time, existing longitudinal mental health studies primarily model mood or risk trajectories, not depression severity. As a result, prior work has not combined clinically grounded depression severity labels with user-level temporal modelling. This study addresses that gap by introducing a time-aware, psychologist-annotated dataset and a model designed to capture severity progression over time.

3 Dataset

We use the DepSy dataset, collected from users on X (formerly Twitter) who self-reported a clinical depression diagnosis, focusing on posts written after the point of diagnosis (Alhamed et al., 2024b). We annotated this dataset for both depressive symptoms and severity using an annotation scheme informed by clinically validated assessment tools (Alhamed et al., 2024a). Each post was labelled for the presence of one or more depressive symptoms, selected from the following set: poor appetite or eating disturbances, feeling down and depressed, crying, concentration problems, feeling tired or

having little energy, feelings of failure, sleep disturbances, loss of interest, self-blame and shame, loneliness, and suicidal thoughts. Posts may contain none, one, or multiple symptoms. While this paper focuses on depression severity, the depression symptoms of posts are annotated and will be used as input features, specifically symptom counts, to support the prediction of severity levels.

3.1 Depression Severity Annotation

To enable severity modelling, we grouped user posts into weekly chunks in line with the clinical practice of validated questionnaires, where users are asked about symptoms of depression faced in the last week. Each chunk contains up to 10 posts of at least eight words. We limited sampling to 10 posts per week per user to manage annotation load and ensure consistency across users with varying posting frequencies. This approach upholds a balance between capturing sufficient user activity and maintaining feasible annotation volumes. Similar sampling strategies have been adopted in prior social media studies that analyse user behaviour over time (Zafar et al., 2024; Boyraz et al., 2015; Heaton et al., 2024), supporting the practicality of this design. Psychologists were asked to assess the overall severity of depression expressed within each chunk, rating it on a 10-point scale (0–9), where 0 indicates no depression and 9 indicates severe depression, following the guidelines in (Alhamed et al., 2024a).

This process yielded 4,000 annotated weekly chunks, creating a longitudinal resource for studying depression severity trajectories in social media.

3.2 Annotation Procedure

Five experienced psychologists conducted the annotation, each with over three years of clinical experience. Annotations were carried out using a customised interface in LabelStudio[2]. Inter-annotator reliability was evaluated using Cohen's kappa (κ), with a pairwise score of 0.67 computed over a 10% subset, reflecting substantial agreement (Landis and Koch, 1977). Measuring inter-annotator agreement on 10% of the data is a widely adopted strategy in annotation studies, balancing reliability checks with practical annotation costs. This practice has been followed in several recent works, including (Abirami et al., 2024; Sanchez-Montero et al., 2025; Jiménez-Zafra et al., 2020; Bastos and Farkas, 2019). In our case, the 10% subset was randomly sampled from the full pool of annotated posts and includes posts from a broad set of users. We ensured annotation consistency by involving a third expert annotator to resolve all cases of disagreement. Annotators were also provided with detailed guidelines and underwent training to align their understanding of the annotation scheme. The dataset, annotated for both depressive symptoms and severity, will be made publicly available upon paper acceptance.

[2] https://labelstud.io/

4 Data Analysis

To gain insights into the temporal and behavioural patterns captured in the dataset, we conducted a descriptive analysis of depression severity trends, user transitions, symptom distributions, and class imbalance.

4.1 Temporal Patterns in Severity

We analysed how depression severity evolves over time at both individual and population levels. Appendix Figure 1 illustrates changes in severity for selected users across multiple months, highlighting the dynamic nature of symptom expression. To investigate broader patterns, we computed the average severity score per user per year and mapped it to one of four discrete levels: no depression (0), mild (1–3), moderate (4–6), and severe (7–9), based on the annotation scheme. A heatmap of year-over-year transitions (Appendix Figure 2) shows how users moved between categories. Most remained in the same or adjacent severity levels—for instance, 47% of those initially categorised as "mild" stayed in the same group the following year. Sharp transitions, such as a direct shift from no depression to severe, were rare, suggesting that severity generally changes gradually over time. We also examined whether severity followed seasonal trends. Appendix Figure 3 shows the monthly distribution of user severity across the dataset. While we expected potential increases during winter or holiday months, no clear or consistent seasonal patterns emerged, suggesting limited calendar-based variation in severity.

4.2 Symptom Trends Across Severity Levels

We further analysed how the frequency of depressive symptoms varies with severity levels. Appendix Figure 4 shows the average occurrence of each symptom across severity scores from 0 to 9.

Some symptoms showed strong correlations with increasing severity. For example, "Feeling down and depressed" steadily increased and appeared in over 90% of posts at severity level 9. "Crying" followed a similar upward trend but plateaued at moderate levels. Other symptoms, such as "Feeling tired or having little energy" and "Loneliness," rose modestly with severity, while symptoms like "Concentration problems," "Self-blame," and "Poor appetite or eating disturbance" remained relatively stable and low in frequency across all levels. "Suicidal thoughts" became more prevalent at higher severity levels, particularly from level 6 onward, reaching 40% at level 9.

These findings suggest that some symptoms, especially emotional and affective ones, are more predictive of increasing severity, while others contribute less distinct signal. This has implications for model design, as certain symptoms may serve as stronger indicators in predicting fine-grained severity levels.

4.3 Severity Class Distribution

Table 1 presents the distribution of samples across severity classes in both the four-class and ten-class settings. The dataset is highly imbalanced, with most samples falling into the non-depressed or mild categories, and relatively few representing severe depression. This observation is consistent with population-level mental health data in both the UK and the US, where the majority of depression cases fall within the mild to moderate range, with fewer individuals meeting the clinical threshold for severe depression (Villarroel and Terlizzi, 2020; Parker et al., 2014). The distribution observed in our dataset mirrors these trends, supporting its representativeness.

Severity Class / Level	Number of Samples
Ten-Class Setting	
0 - No Depression	2675
1	464
2	351
3	149
4	66
5	43
6	23
7	14
8	6
9 - Severe Depression	2
Four-Class Setting	
No Depression	2675
Mild	964
Moderate	132
Severe	22

Table 1: Distribution of samples across severity classes and severity levels in the ten-class and four-class settings based on DepSy dataset.

5 Depression Severity Monitoring

This study aims to predict depression severity from user-generated text, where severity is represented as an integer on a `0-9` scale, with `0` indicating no depression and `9` indicating the most severe level. Given a weekly chunk of posts $P_i = \{p_{i,1}, \ldots, p_{i,n}\}$, the model $f(\cdot)$, parameterised by θ, predicts a severity score $S_i \in \{0, \ldots, 9\}$.

We explore three formulations of this task. The first treats it as **regression**, where the model outputs a real-valued score that is rounded to the nearest integer. The second frames it as **ten-class classification**, assigning each input to a discrete severity level. The third simplifies the task to a **four-class classification** problem, grouping levels into `0` (No Depression), `1-3` (Mild), `4-6` (Moderate), and `7-9` (Severe). Grouping severity levels into broader categories reduces granularity but often improves classification performance and aligns with clinical practice, where diagnoses are typically assigned as mild, moderate, or severe rather than on a fine-grained scale. We apply several models under each formulation to assess how different approaches capture severity and to examine trade-offs between fine-grained prediction and classification performance.

6 CTMC-LSTM Hybrid Model for Depression Severity Prediction

To improve temporal consistency and robustness in predicting depression severity, we propose a hybrid architecture that combines neural predictions from an LSTM with probabilistic reasoning derived from a Continuous-Time Markov Chain (CTMC). We use CTMC to model transitions between depression severity states over time. We estimate the transition rate matrix from the available severity annotations, which capture how frequently and quickly one severity level shifts to another. This hybrid approach will allow us to leverage both chunk-level classification and temporal information in prediction. This model is applied to a four-class severity classification task, derived from a more granular ten-class annotation scheme. The ten-class scheme was found to be difficult to model reliably due to its sparsity, so the categorised four-class version is

used for evaluation.

6.1 Model Architecture

The proposed hybrid model combines an LSTM classifier with a Continuous-Time Markov Chain (CTMC) to predict depression severity from user posts. The base classifier is an LSTM model trained to predict one of four severity levels—0 (No), 1 (Mild), 2 (Moderate), or 3 (Severe)—from individual posts. In parallel, a CTMC is estimated from the training data to capture transitions between severity states over time. The transition rate matrix is learned via maximum likelihood estimation using sequences of severity labels. At inference time, both components produce probability distributions over severity classes. The LSTM generates a distribution via softmax, denoted as `lstm_probs`, while the CTMC produces a probability vector `markov_probs` based on the previous predicted state and the elapsed time. These two distributions are combined through a weighted average, where a hyperparameter $\alpha \in [0, 1]$ controls the contribution of each source. This integration allows the model to balance direct textual signals with temporal progression patterns in severity. The final prediction is computed as:

$$\texttt{probs} = \alpha \cdot \texttt{lstm\_probs} + (1-\alpha) \cdot \texttt{markov\_probs}$$

Inspired by Gao et al. (2020); Zawbaa et al. (2024); Liu et al. (2019), after obtaining the final probability distribution, a thresholding strategy is applied to prioritise high-severity decisions. This mechanism, tuned empirically on a validation set to optimise recall for higher severity levels while preserving balanced performance, increases the likelihood of detecting moderate and severe cases even when they are not the top-scoring class. The empirically derived thresholds are `Severe: 0.12, Moderate:0.25, Mild:0.40`.

7 Models and Experiments

LSTM Long Short-Term Memory (LSTM) networks were utilised for this task, given their well-established efficacy in handling longitudinal and sequential data.

BERT, RoBERTa, and MentalBERT BERT-based models were employed to predict depression severity across all three tasks. The models' specifications and hyperparameter settings are detailed in Appendix B

Algorithm 1: Markov-Neural hybrid inference procedure for severity prediction

Input: Post x_t, previous severity label s_{t-1}, time gap Δt, CTMC matrix Q, weight α
Output: Predicted severity class y_t

`// Step 1: LSTM/BERT prediction`
$p_{\text{lstm}} \leftarrow$ softmax output from neural model on x_t
`// Step 2: CTMC-based prediction`
$P \leftarrow \exp(Q \cdot \Delta t)$ `// Transition probability matrix over time`
$p_{\text{markov}} \leftarrow P[s_{t-1}]$ `// Row for previous state`
`// Step 3: Combine model and CTMC`
$p_{\text{final}} \leftarrow \alpha \cdot p_{\text{lstm}} + (1-\alpha) \cdot p_{\text{markov}}$
`// Step 4: Threshold-based decision rule`

$$y_t = \begin{cases} \max\{k \in \{1,2,3\} : p_{\text{final}}[k] > \theta_k\} & \text{if } k \text{ exists} \\ \arg\max(p_{\text{final}}) & \text{otherwise} \end{cases}$$

return y_t

7.1 Experiments

We conducted a series of experiments in order to establish our benchmark. We used 5-fold cross-validation to evaluate performance based on accuracy, macro-averaged precision, recall, and F1 scores. Our implementation utilizes Scikit-learn (Pedregosa et al., 2011). For each of the three task approaches, we explored different input features to assess their impact on model performance.

- **Features_1:** A chunk of posts (over 1 week) is used as the input.
- **Features_2:** A chunk of posts, along with the sum of symptoms, where the sum of the symptoms represents the total of all symptoms identified in the posts within the respective chunk.
- **Features_3:** A chunk of posts is used, along with a sequence of the previous three chunks, [3] to capture longitudinal severity trends (longitudinal).
- **Features_4:** A set of 12 numerical features: 11 represent symptom occurrences, with each representing the total count of a specific symptom across all posts in the chunk, plus one feature for the overall symptom sum.
- **Features_5:** This incorporates Features_4 for the current and previous three chunks (longitudinal).

Each combination of input features was tested to determine the optimal configuration for predicting depression severity.

[3] A sequence of 3 is selected inline with other works (Suhara et al., 2017; Rónai and Polner, 2021)

Approach	Input	Model	RMSE (↓)	Macro F1 (↑)	Acc (↑)	CI 95%
Regression	Features 1	LTSM	1.310	-	-	[1.192, 1.429]
		BERT	1.215	-	-	[1.098, 1.333]
		RoBERTa	1.211	-	-	[1.091, 1.348]
		MentalBERT	1.212	-	-	[1.084, 1.344]
	Features 2	LSTM	0.868	-	-	[0.766, 0.969]
		BERT	1.194	-	-	[1.062, 1.315]
		RoBERTa	1.128	-	-	[1.016, 1.241]
		MentalBERT	1.195	-	-	[1.061, 1.324]
	Features 3	LSTM	1.461	-	-	[0.985, 1.217]
	Features 4	LSTM	**0.66**	-	-	[0.564, 0.771]
	Features 5	LSTM	1.445	-	-	[1.301, 1.590]
Multi-Class classification (9 classes)	Features 1	LTSM	-	0.13	0.629	[0.597, 0.665]
		BERT	-	0.14	0.591	[0.557, 0.626]
		RoBERTa	-	0.14	0.671	[0.640, 0.704]
		MentalBERT	-	0.16	0.653	[0.619, 0.688]
	Features 2	LSTM	-	0.15	0.654	[0.621, 0.685]
		BERT	-	0.17	0.638	[0.603, 0.676]
		RoBERTa	-	0.15	0.661	[0.627, 0.696]
		MentalBERT	-	0.23	0.725	[0.693, 0.755]
	Features 3	LSTM	-	0.09	0.698	[0.666, 0.729]
	Features 4	LSTM	-	0.27	**0.778**	[0.749, 0.808]
	Features 5	LSTM	-	0.09	0.626	[0.593, 0.660]
Multi-Class classification (4 classes)	Features 1	LSTM	-	0.33	0.689	[0.657, 0.721]
		BERT	-	0.34	0.685	[0.653, 0.715]
		RoBERTa	-	0.34	0.706	[0.672, 0.737]
		MentalBERT	-	0.30	0.654	[0.621, 0.688]
	Features 2	LSTM	-	0.32	0.702	[0.668, 0.734]
		BERT	-	0.34	0.714	[0.682, 0.747]
		RoBERTa	-	0.36	0.722	[0.689, 0.752]
		MentalBERT	-	0.30	0.691	[0.659, 0.722]
	Features 3	LSTM	-	0.21	0.698	[0.667, 0.731]
	Features 4	LSTM	-	0.58	**0.877**	[0.855, 0.901]
		CTMC-LSTM	-	**0.72**	0.867	[0.836, 0.878]
	Features 5	LSTM	-	0.24	0.630	[0.597, 0.660]

Table 2: Results for models on classifying depression severity from our DepSy dataset. Macro F1 is used to account for class imbalance, as it gives equal importance to each severity level regardless of frequency. This is especially important for evaluating performance on the severe class, which is underrepresented but clinically critical.

Severity	LSTM (Features_4)			LSTM (Features_2)			BERT			RoBERTa			MentalBERT		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
0	0.88	0.99	0.93	0.79	0.85	0.82	0.81	0.75	0.78	0.81	0.87	0.84	0.80	0.84	0.82
1	0.24	0.18	0.21	0.17	0.14	0.15	0.14	0.35	0.20	0.22	0.11	0.15	0.11	0.08	0.10
2	0.46	0.30	0.36	0.21	0.16	0.18	0.26	0.12	0.17	0.21	0.28	0.24	0.23	0.23	0.23
3	0.29	0.20	0.24	0.08	0.07	0.07	0.22	0.07	0.10	0.07	0.03	0.04	0.24	0.27	0.25
4	0.50	0.27	0.35	0.09	0.09	0.09	0	0	0	0	0	0	0	0	0
5	0.33	0.29	0.31	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 3: Precision, recall, and F1-scores per severity class across models

Model	No Depression			Mild			Moderate			Severe		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baseline	0.70	**1**	**0.83**	0	0	0	0	0	0	0	0	0
LSTM Features4	0.92	**0.94**	**0.93**	**0.75**	0.76	**0.75**	**0.72**	0.54	**0.62**	0	0	0
BERT	0.80	0.80	0.80	0.40	0.42	0.41	0.17	0.12	0.14	0	0	0
RoBERTa	0.82	0.81	0.82	0.44	0.46	0.45	0.13	0.08	0.10	0	0	0
MentalBERT	0.81	0.74	0.77	0.37	0.50	0.42	0	0	0	0	0	0
CTMC-LSTM	**0.95**	0.87	0.91	0.68	**0.84**	**0.75**	0.62	**0.62**	**0.62**	**1**	**0.43**	**0.60**

Table 4: Precision, recall, and F1-scores per severity class (categorized to 4 classes) across models. The baseline model always predicts the “No Depression" class, achieving high recall but failing to detect mild, moderate, or severe cases.

8 Results

Our results in Table 2 indicate that the LSTM model with numerical features (`Features_4`) consistently achieved the best performance across all tasks. It obtained the lowest RMSE (0.66) in the regression task and the highest F1-scores in multi-class classification, with 0.778 for the 9-class setting and 0.877 for the 4-class setting.

We initially hypothesised that incorporating longitudinal severity trends would enhance prediction performance by providing additional contextual information. However, the results did not fully support this assumption. Modelling past severity states (`Features_3` and `Features_5`) directly as input features did not lead to significant improvements in performance. Instead, the best results were obtained using numerical symptom features combined with temporal dynamics through our CTMC-LSTM model, which incorporated severity sequences via a Markovian transition framework. This suggests that depression severity estimation in this setting benefits from incorporating broader temporal patterns at the decision level rather than relying on direct sequential modelling of input features.

Furthermore, textual information alone did not contribute significantly to improving performance. While BERT-based models demonstrated strong predictive capabilities, they did not surpass models that leveraged structured numerical symptom features. This highlights the importance of explicit symptom representations in this task. The findings suggest that immediate symptom patterns play an important role in predicting severity, challenging the assumption that historical severity states provide additional predictive value.

A more detailed examination of per-class performance, as shown in Table 3, reveals critical insights beyond the overall metrics. All models perform best on class 0 (non-depressed), with high F1-scores above 0.80, particularly LSTM with Features_4 (numerical), achieving the highest at 0.93. However, performance drops substantially across all other severity classes. Classes 1, 2 and 3 are weakly detected, with only LSTM (numerical) showing some capability (F1-scores of 0.21, 0.36 and 0.24, respectively). The results show that while models can effectively detect the lowest severity level (class 0), their ability to identify moderate to high severity cases (classes 4 to 9) remains limited, with performance metrics dropping to near zero in most cases. This is a critical limitation, as missing severe cases may prevent timely support for individuals in urgent need. Despite using high-quality annotations provided by psychologists and clinically grounded labels, the models struggle to capture these less frequent classes. This might be an outcome of the naturally imbalanced distribution of depression severity in the population, where high severity cases are relatively rare. These findings emphasise the difficulty of this task and the need for more effective modelling strategies that can detect clinically important but infrequent severity levels without compromising overall performance.

Table 4 presents the performance of five models across four severity levels. All models perform well on the no-depression class (0), with F1-scores above 0.77. However, performance declines consistently with increasing severity. Mild and moderate levels (1 and 2) show lower precision and recall, and none of the models were able to identify any instances of the severe class (3), except for the CTMC-LSTM model. This highlights the persistent difficulty of detecting high-severity depression from text, even with reliable annotations and a realistic class distribution. The proposed CTMC-LSTM model achieved the highest F1-score for the

severe class (0.60), where all other models failed. It also matched the best performance on the moderate class (F1 = 0.62) and slightly improved recall compared to LSTM alone. For mild cases, CTMC-LSTM achieved the highest F1-score (0.75), mainly through increased recall. BERT-based models underperformed across all classes beyond no depression. In contrast, the LSTM model produced more balanced results, and CTMC fusion further improved its performance without reducing accuracy on the non-depressed class. CTMC-LSTM demonstrated the most consistent results across all severity levels. We hypothesise that the combination of temporal smoothing from CTMC, correction of model uncertainty, and threshold-based prioritisation of minority classes allowed CTMC-LSTM to outperform other models that rely solely on immediate chunk-level classification.

9 Discussion

The results of our depression severity prediction experiments highlight the difficulty of this task, particularly when relying solely on post-level textual features. Most models struggled to identify moderate and severe cases, with no model achieving reliable performance on these classes in the absence of symptom input. Only our proposed CTMC-LSTM model achieved an F1-score of 0.72 when provided with the full set of ground-truth symptoms. This performance gap may be attributed to several factors. First, the linguistic signals distinguishing higher severity levels are often subtle and inconsistent, making them difficult to learn. Second, the distribution of severity levels in the dataset is highly imbalanced, with "Severe" cases accounting for less than 10% of the data. Such imbalance can hinder the model's ability to generalise to underrepresented classes, a limitation also noted in related tasks such as suicidal risk detection (Tsakalidis et al., 2022a) and post-level depression severity prediction (Kabir et al., 2023).

Model predictions were consistently biased towards the majority class, a common issue in imbalanced learning scenarios. Although we applied class-balancing strategies during training to support minority class learning, evaluation was always conducted on the original distribution to preserve the dataset's real-world representativeness and deployment relevance.

Interestingly, sequence modelling did not lead to improved results. In fact, models using only current post features often performed better than sequential models. This may be due to the lack of consistent temporal patterns in severity progression, as suggested by the user-level trajectories in Figure 1.

We also hypothesised that symptom frequency would help predict higher severity levels, particularly given the strong association between symptoms such as suicidal thoughts and severe depression (Figure 4). However, the benefit of symptom features was limited, likely due to the scarcity of severe examples, which restricted the model's ability to learn these associations effectively.

Finally, predicting severity on a ten-point scale proved particularly challenging. This may be explained by the uneven contribution of symptoms across severity levels: while some symptoms show a clear progression, others remain stable or infrequent. As a result, consecutive severity levels often contain overlapping or indistinguishable symptom patterns, reducing the granularity of available signal and making fine-grained severity classification difficult.

10 Conclusions and Future Work

This paper introduced the DepSy Severity dataset and a set of modelling approaches for depression severity prediction from longitudinal social media data. To our knowledge, this is the first English-language dataset combining chunk-based severity annotations in a longitudinal format. We explored both regression and classification formulations of severity prediction, incorporating textual, symptom-based, and temporal features. Our experiments showed that models leveraging structured features, particularly symptoms, outperformed purely text-based models. The LSTM model trained with these features achieved consistently strong performance across task settings. Sequential models incorporating previous chunks did not yield improvements. To introduce temporal consistency into predictions, we proposed a hybrid CTMC-LSTM model that integrates LSTM predictions with severity transition probabilities derived from a Continuous-Time Markov Chain. This hybrid approach improved classification performance, particularly for underrepresented classes such as moderate and severe depression. Despite these advances, predicting depression severity remains a challenging task, especially for fine-grained levels and minority classes.

11 Limitations

While this study contributes a novel dataset and modelling approach for depression severity prediction, several limitations remain. First, we were unable to validate model generalisability on external datasets, due to lack of similar datasets. As a result, the evaluation is limited to the DepSy dataset. Second, although DepSy is annotated by expert psychologists, the reliance on publicly available social media posts may introduce self-presentation bias, limiting coverage of the broader population affected by depression. Finally, severity scores are based on weekly post samples and do not account for external factors—such as life events or clinical context—that may influence depression but are not observable in text, potentially limiting the accuracy of severity estimation from social media alone.

Ethical Consideration

This study has received ethics approval from XXXXXX [4] (Reference: 21IC7222). The dataset contains only publicly available posts from X, and we are committed to following ethical practices to protect the privacy and anonymity of the users. To ensure this, the author's usernames, which could contain sensitive information related to the names or locations of the user, are not saved or used. Instead, the information was pre-processed and replaced with user IDs. Social media data is often sensitive, particularly when it is related to mental health, and we take great care to ensure that our dataset is handled responsibly. Since the dataset is related to mental disorders, it might trigger some people, thus, annotators were advised to take breaks during annotation and were given plenty of time.

References

AM Abirami, Wei Qi Leong, Hamsawardhini Rengarajan, D Anitha, R Suganya, Himanshu Singh, Kengatharaiyer Sarveswaran, William Chandra Tjhi, and Rajiv Shah. 2024. Aalamaram: A large-scale linguistically annotated treebank for the tamil language. In *Proceedings of the 7th Workshop on Indian Language Data: Resources and Evaluation*, pages 73–83.

V Adarsh, P Arun Kumar, V Lavanya, and GR Gangadharan. 2023. Fair and explainable depression detection in social media. *Information Processing & Management*, 60(1):103168.

Tasnim Ahmed, Shahriar Ivan, Ahnaf Munir, and Sabbir Ahmed. 2024. Decoding depression: Analyzing social network insights for depression severity assessment with transformers and explainable ai. *Natural Language Processing Journal*, 7:100079.

Falwah Alhamed, Rebecca Bendayan, Julia Ive, and Lucia Specia. 2024a. Monitoring depression severity and symptoms in user-generated content: An annotation scheme and guidelines. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 227–233.

Falwah Alhamed, Julia Ive, and Lucia Specia. 2024b. Classifying social media users before and after depression diagnosis via their language usage: A dataset and study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3250–3260.

Nadiah A Baghdadi, Amer Malki, Hossam Magdy Balaha, Yousry AbdulAzeem, Mahmoud Badawy, and Mostafa Elhosseini. 2022. An optimized deep learning approach for suicide detection through arabic tweets. *PeerJ Computer Science*, 8:e1070.

Marco Bastos and Johan Farkas. 2019. "donald trump is my president!": The internet research agency propaganda machine. *Social Media+ Society*, 5(3):2056305119865466.

Maggie Boyraz, Aparna Krishnan, and Danielle Catona. 2015. Who is retweeted in times of political protest? an analysis of characteristics of top tweeters and top retweeted users during the 2011 egyptian revolution. *Atlantic Journal of Communication*, 23(2):99–119.

Chun Yueh Chiu, Hsien Yuan Lane, Jia Ling Koh, and Arbee L.P. Chen. 2021. Multimodal depression detection on instagram considering time interval of posts. *Journal of Intelligent Information Systems*, 56(1):25–47.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. *Conference on Human Factors in Computing Systems - Proceedings*, pages 2098–2110.

Joseigla Pinto de Oliveira, Karen Jansen, Taiane de Azevedo Cardoso, Thaíse Campos Mondin, Luciano Dias de Mattos Souza, Ricardo Azevedo da Silva, and Fernanda Pedrotti Moreira. 2021. Predictors of conversion from major depressive disorder to bipolar disorder. *Psychiatry Research*, 297(January):113740.

[4] masked for anonymity

Shang Gao, Wenlu Dong, Ke Cheng, Xibei Yang, Shang Zheng, and Hualong Yu. 2020. Adaptive decision threshold-based extreme learning machine for classifying imbalanced multi-label data. *Neural Processing Letters*, 52:2151–2173.

Jue Gong, Gregory E. Simon, and Shan Liu. 2019. Machine learning discovery of longitudinal patterns of depression and suicidal ideation. *PLoS ONE*, 14(9):1–15.

Dan Heaton, Jeremie Clos, Elena Nichele, and Joel E Fischer. 2024. "the chatgpt bot is causing panic now–but it'll soon be as mundane a tool as excel": analysing topics, sentiment and emotions relating to chatgpt on twitter. *Personal and Ubiquitous Computing*, pages 1–20.

Anthony Hills, Adam Tsakalidis, and Maria Liakata. 2023. Time-aware predictions of moments of change in longitudinal user posts on social media. In *International Workshop on Advanced Analytics and Learning on Temporal Data*, pages 293–305. Springer.

Anthony Hills, Talia Tseriotou, Xenia Miscouridou, Adam Tsakalidis, and Maria Liakata. 2024. Exciting mood changes: A time-aware hierarchical transformer for change detection modelling. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12526–12537.

Mariia Ignashina, Paulina Bondaronek, Dan Santel, John Pestian, and Julia Ive. 2025. Llm assistance for pediatric depression. *Preprint*, arXiv:2501.17510.

Salud María Jiménez-Zafra, Roser Morante, M Teresa Martín-Valdivia, and L Alfonso Urena Lopez. 2020. Corpora annotated with negation: An overview. *Computational Linguistics*, 46(1):1–52.

Mohsinul Kabir, Tasnim Ahmed, Md Bakhtiar Hasan, Md Tahmid Rahman Laskar, Tarun Kumar Joarder, Hasan Mahmud, and Kamrul Hasan. 2023. Deptweet: A typology for social media texts to detect depression severities. *Computers in Human Behavior*, 139:107503.

D Sami Khafaga, Maheshwari Auvdaiappan, K Deepa, Mohamed Abouhawwash, and F Khalid Karim. 2023. Deep learning for depression detection using twitter data. *Intelligent Automation & Soft Computing*, 36(2):1301–1313.

Hyewon Kim, Yuwon Kim, Ji Hyun Baek, Maurizio Fava, David Mischoulon, Andrew A. Nierenberg, Kwan Woo Choi, Eun Jin Na, Myung Hee Shin, and Hong Jin Jeon. 2020. Predictive factors of diagnostic conversion from major depressive disorder to bipolar disorder in young adults ages 19–34: A nationwide population study in South Korea. *Journal of Affective Disorders*, 265(December 2019):52–58.

Harnain Kour and Manoj K Gupta. 2022. An hybrid deep learning approach for depression prediction from user tweets using feature-rich cnn and bidirectional lstm. *Multimedia Tools and Applications*, 81(17):23649–23685.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Jiaxiang Liu, Shuohuan Wang, and Yu Sun. 2019. Olenet at semeval-2019 task 9: Bert based multi-perspective models for suggestion mining. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 1231–1236.

David E Losada, Fabio Crestani, and Javier Parapar. 2017. Overview of erisk: Early risk prediction on the internet. In *Working Notes of CLEF 2017—Conference and Labs of the Evaluation Forum*, volume 1866 of *CEUR Workshop Proceedings*.

David E Losada, Fabio Crestani, and Javier Parapar. 2018. Overview of erisk 2018: Early risk prediction on the internet. In *Working Notes of CLEF 2018—Conference and Labs of the Evaluation Forum*, volume 2125 of *CEUR Workshop Proceedings*.

David E Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of erisk 2019: Early risk prediction on the internet. In *Working Notes of CLEF 2019—Conference and Labs of the Evaluation Forum*, volume 2380 of *CEUR Workshop Proceedings*.

David E Losada, Fabio Crestani, and Javier Parapar. 2020. Overview of erisk 2020: Early risk prediction on the internet. In *Working Notes of CLEF 2020—Conference and Labs of the Evaluation Forum*, volume 2696 of *CEUR Workshop Proceedings*.

David E Losada, Fabio Crestani, and Javier Parapar. 2021. Overview of erisk 2021: Early risk prediction on the internet. In *Working Notes of CLEF 2021—Conference and Labs of the Evaluation Forum*, volume 2936 of *CEUR Workshop Proceedings*.

Rahele Mesbah, Nienke de Bles, Nathaly Rius-Ottenheim, A. J.Willem van der Does, Brenda W.J.H. Penninx, Albert M. van Hemert, Max de Leeuw, Erik J. Giltay, and Manja Koenders. 2021. Anger and cluster B personality traits and the conversion from unipolar depression to bipolar disorder. *Depression and Anxiety*, (August 2020):1–11.

Usman Naseem, Adam G Dunn, Jinman Kim, and Matloob Khushi. 2022. Early identification of depression severity levels on reddit using ordinal classification. In *Proceedings of the ACM web conference 2022*, pages 2563–2572.

Bayode Ogunleye, Hemlata Sharma, and Olamilekan Shobayo. 2024. Sentiment informed sentence BERT-ensemble algorithm for depression detection. *Big Data Cogn. Comput.*, 8(9):112.

Gordon Parker, Kathryn Fletcher, Amelia Paterson, Josephine Anderson, and Michael Hong. 2014. Gender differences in depression severity and symptoms across depressive sub-types. *Journal of affective disorders*, 167:351–357.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Melanie F. Pradier, Michael C. Hughes, Thomas H. McCoy, Sergio A. Barroilhet, Finale Doshi-Velez, and Roy H. Perlis. 2021. Predicting change in diagnosis from major depression to bipolar disorder after antidepressant initiation. *Neuropsychopharmacology*, 46(2):455–461.

Amna Qasim, Gull Mehak, Nisar Hussain, Alexander Gelbukh, and Grigori Sidorov. 2025. Detection of depression severity in social media text using transformer-based models. *Information*, 16(2):114.

Levente Rónai and Bertalan Polner. 2021. Getting the blues: negative affect dynamics mediate the within-person association of maladaptive emotion regulation and depression.

Kayalvizhi Sampath and Thenmozhi Durairaj. 2022. Data set creation and empirical analysis for detecting signs of depression from social media postings. In *International Conference on Computational Intelligence in Data Science*, pages 136–151. Springer.

Alec M Sanchez-Montero, Gemma Bel-Enguix, Sergio-Luis Ojeda-Trueba, and Gerardo Sierra Martínez. 2025. Disagreement in metaphor annotation of mexican spanish science tweets. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 155–164.

Ramit Sawhney, Harshit Joshi, Lucie Flek, and Rajiv Shah. 2021. Phase: Learning emotional phase-aware representations for suicide ideation detection on social media. In *Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics: main volume*, pages 2415–2428.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2020. A Time-Aware Transformer Based Model for Suicide Ideation Detection on Social Media. pages 7685–7697.

Jiayu Song, Jenny Chim, Adam Tsakalidis, Julia Ive, Dana Atzil-Slonim, and Maria Liakata. 2024. Combining hierachical vaes with llms for clinically meaningful timeline summarisation in social media. *arXiv preprint arXiv:2401.16240*.

Yoshihiko Suhara, Yinzhan Xu, and Alex'Sandy' Pentland. 2017. Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In *Proceedings of the 26th International Conference on World Wide Web*, pages 715–724.

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. 2022a. 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: Mental Health in the Face of Change*.

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022b. Identifying moments of change from longitudinal user text. *arXiv preprint arXiv:2205.05593*.

Maria A Villarroel and Emily P Terlizzi. 2020. Symptoms of depression among adults: United states, 2019.

Anas Zafar, Danyal Aftab, Rizwan Qureshi, Yaofeng Wang, and Hong Yan. 2024. Multi-explainable temporalnet: An interpretable multimodal approach using temporal convolutional network for user-level depression detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2258–2265.

Hossam Zawbaa, Wael Rashwan, Sourav Dutta, and Haytham Assem. 2024. Improved out-of-scope intent classification with dual encoding and threshold-based re-classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8708–8718.

Tianlin Zhang, Kailai Yang, and Sophia Ananiadou. 2023. Sentiment-guided transformer with severity-aware contrastive learning for depression detection on social media. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 114–126.

A Users Depression Severity Analysis

B Models Hyper-parameters for Extracting Depression Severity

LSTM .

Tokenizer for text input: BERT CLS

```
Epochs: 50
Batch_size: 16
Learning_rate:0.01
Hidden_size:128
Optimizer:Adam
Loss: CrossEntropy
```

BERT .

```
Model_card: "bert-base-uncased"
Epochs: 64
Batch_size: 8
Learning_rate:5e-5
Hidden_size:128
Optimizer:Adam
Loss: CrossEntropy
```

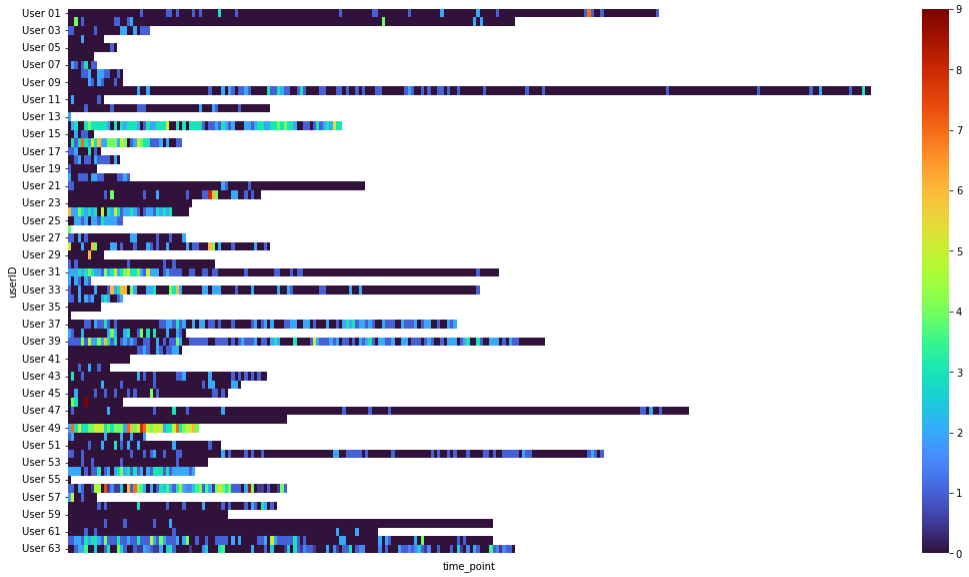

Figure 1: Changes in depression severity for users in our dataset. Colours from blue to red reflect the severity of the depression score 0-9, where blue is mild and red is severe depression.

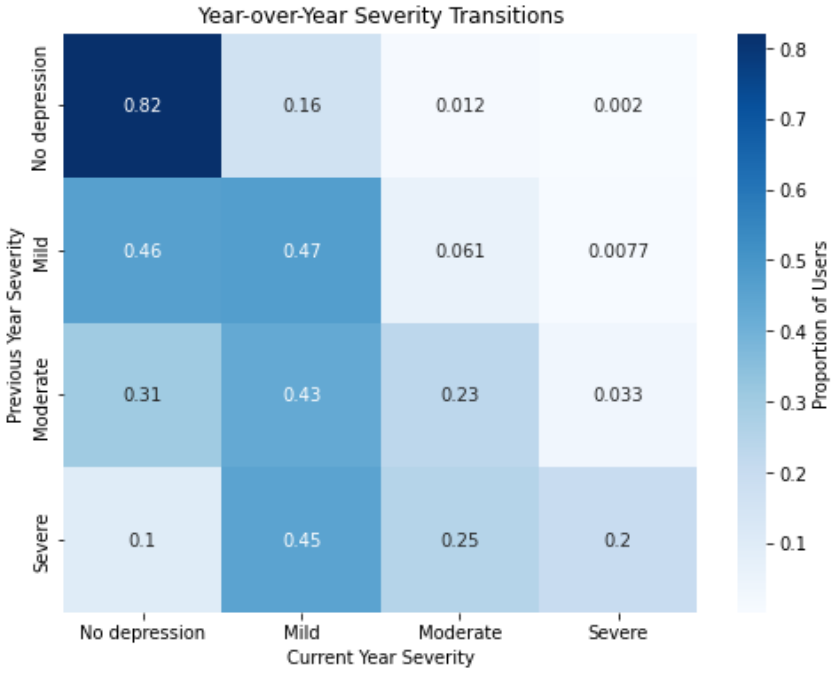

Figure 2: Severity transition year over year heatmap

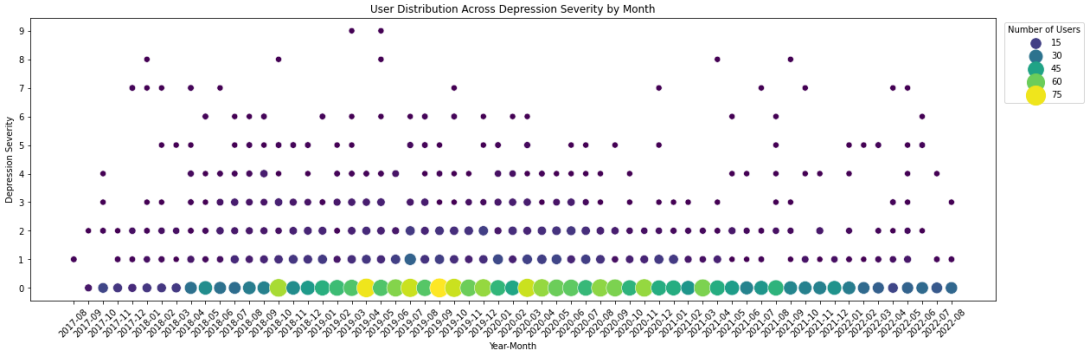

Figure 3: User distribution across depression severity by month

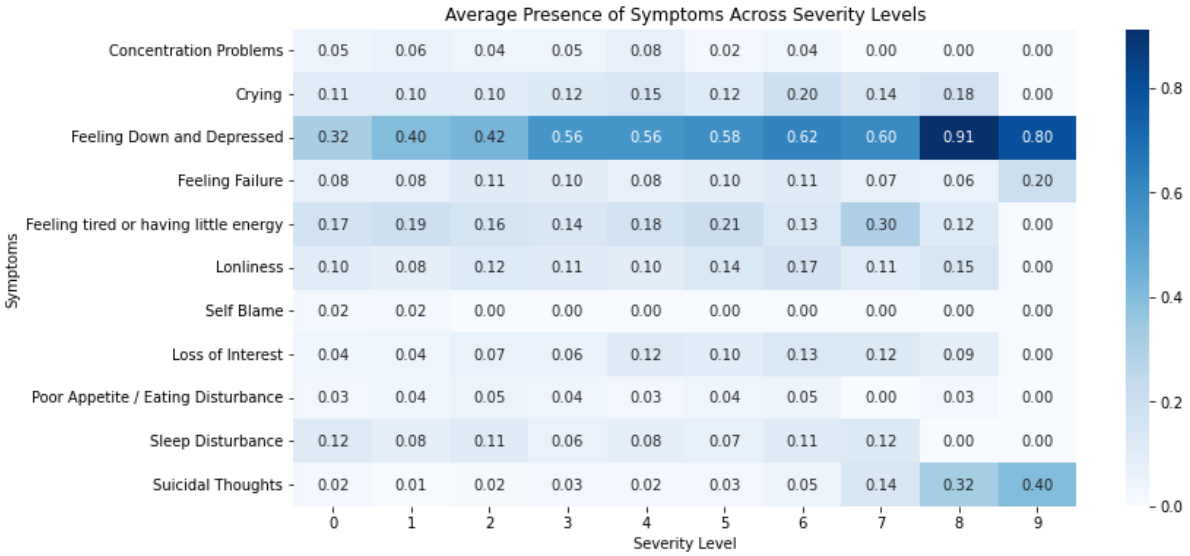

Figure 4: Frequency of symptoms associated with different severity levels

RoBERTa .

```
Model_card: "roberta-base"
Epochs: 64
Batch_size: 8
Learning_rate:2e-5
Hidden_size:128
Optimizer:Adam
Loss: CrossEntropy
```

MentalBERT .

```
Model_card: "mental-bert-base-uncased"
Epochs: 64
Batch_size: 8
Learning_rate:5e-5
Hidden_size:128
Optimizer:Adam
Loss: CrossEntropy
```