A Robust Baseline for User-Level Depression Detection using Pretrained Sentence-Transformers

Anonymous ACL submission

Abstract

User-level depression detection from social media requires robust models that can generalize across different platforms and user behaviors. This paper presents a rigorous, large-scale empirical study to establish a new state-of-the-art baseline for this task. We systematically evaluate a powerful architecture, combining pretrained sentence embeddings with a global self-attention mechanism, on three public benchmark datasets: MDL, CLPsych15, and eRisk18. Our results demonstrate that this approach not only significantly outperforms previous methods but also shows strong generalization across heterogeneous data environments. Through in-depth introspection and error analysis, we identify key limitations of our approach, such as information loss from random sampling in users with extensive post histories. By providing a new high-water mark for performance and a deeper understanding of the task's complexities, this work offers a robust benchmark for future research in computational mental health.

1 Introduction

Major depressive disorder (MDD), also known simply as depression, is among the most prevalent psychiatric disorders globally (Le and Boyd, 2006; Kessler et al., 2009). As described in the World Health Organization's Comprehensive Mental Health Action Plan 2013-2030 (World Health Organization, n.d.), depression affects approximately 280 million people and is one of the largest single causes of disability worldwide. The Institute of Medicine identified depression as a highly preventable disorder (Muñoz et al., 1996), and numerous studies have shown that early recognition and treatment can significantly mitigate its negative impacts (Rost et al., 2004; Halfin, 2007; Picardi et al., 2016). Therefore, developing methods for the early identification of individuals suffering from depression is vital for timely intervention and for reducing the broader public health burden.

Text classification (TC) from social media has emerged as a powerful and scalable tool for this purpose, with the field seeing rapid evolution from traditional machine learning to advanced deep learning techniques (Tahir et al., 2025). While early machine learning models laid the groundwork, the advent of deep learning, particularly transformer-based architectures like BERT (Devlin, 2018) and its successors, revolutionized the field. By capturing deep contextual nuances in language, these models have surpassed traditional approaches in a wide range of NLP tasks. Indeed, recent advancements using fine-tuned Large Language Models have achieved near-perfect classification of individual posts, reaching an F1-score of 0.975 on the MDL dataset (Shah et al., 2025).

However, detecting depression is not a simple post-level classification task. A user's depressive state is often reflected not in a single, explicit post, but in a subtle, distributed pattern of language use across their entire history. These linguistic cues can be inconsistent and spread thinly across hundreds or thousands of posts, posing a significant challenge for models that treat posts in isolation or use simple aggregation methods.

While modern architectures have shown promise, the field lacks a comprehensive study that validates a single, powerful approach across the primary benchmark datasets to establish its true generalizability. This work aims to fill that gap. Our primary contributions are:

1. We conduct a large-scale, systematic evaluation of a powerful user-level classification architecture, leveraging modern pretrained sentence embeddings and a global self-attention mechanism, across three distinct and widely-used benchmarks.

2. We establish new state-of-the-art (SOTA) performance on all three datasets (MDL, CLPsych15, and eRisk18), providing a robust

and highly effective baseline for future comparative studies.

3. Through detailed model introspection and error analysis, we provide critical insights into the model's decision-making process and identify a key limitation of the proposed method related to information loss for high-volume users.

2 Related Work

The task of detecting depression from social media has evolved significantly, moving from classical machine learning techniques to sophisticated deep learning architectures, as comprehensively surveyed by (Tahir et al., 2025). This evolution can be understood as a progressive effort to better capture the complex, user-level linguistic patterns indicative of mental health status.

Early approaches relied on classical machine learning models such as Support Vector Machines (SVM) and Logistic Regression, coupled with hand-crafted features (Lai et al., 2015; Resnik et al., 2015; Kayalvizhi and Thenmozhi, 2022). These features typically included bag-of-words (BoW) representations, n-grams, and TF-IDF scores, which primarily capture lexical frequencies. While effective to a degree, these methods are inherently limited as they disregard word order and semantic context, making them incapable of identifying the subtle linguistic patterns crucial for depression detection.

The first wave of deep learning introduced models like Convolutional Neural Networks (CNNs) for capturing local patterns (e.g., key phrases) and Recurrent Neural Networks (RNNs) (Lai et al., 2015), including LSTMs, for modeling sequential information in text (Zhang et al., 2015; Kim et al., 2020). Architectures like C-LSTM (Zhou et al., 2015) combined these strengths.

The advent of transformer-based models, particularly BERT (Devlin, 2018) and RoBERTa (Liu, 2019), marked a paradigm shift. Their self-attention mechanism allows them to generate deep, context-aware embeddings for each post, leading to a superior understanding of individual texts.

Recent work by Shah et al. (2025) demonstrated high efficacy in post-level analysis by employing fine-tuned versions of GPT-3.5 and LLaMA2-7B to classify individual tweets from the MDL dataset. Their approach achieved an impressive F1-score of 0.975 in a few-shot setting, effectively solving the task of identifying explicitly depressive content in single posts. This success, however, crystallizes the shift in the research challenge: from understanding a single post to synthesizing signals across a user's entire history. This is precisely the gap our work addresses.

Many recent studies have leveraged transformers as powerful post encoders, often feeding the resulting embeddings into a subsequent classifier like an LSTM or CNN (Haque et al., 2021; Zogan et al., 2021; Thekkekara et al., 2024). However, this approach shifts the core challenge from post-level understanding to effectively synthesizing the user's history. Simply averaging post embeddings or using a standard RNN risks diluting the critical, subtle textual cues spread across a user's posts. The key research question becomes: how can we combine hundreds or thousands of rich post embeddings in a way that amplifies these distributed depressive indicators?

An early and influential approach to this challenge was the Hierarchical Attention Network (HAN) (Han et al., 2022), which first builds representations for individual posts and then combines them to form a user representation. This hierarchical paradigm remains a prominent direction in recent research. For example, (Zhang et al., 2024) proposed a Hierarchical Transformer Network (HTN) which utilizes a Transformer to encode individual posts, followed by another Transformer and an LSTM-attention layer for user-level classification on a large Chinese-language dataset.

Our work distinguishes itself from prior art by providing the rigorous, cross-platform validation of a modern approach that leverages powerful, pre-trained sentence embeddings. While many studies focus on architectural novelty within a single data environment (e.g., Zhang et al. (2024) on Chinese Weibo), our contribution is the systematic and large-scale empirical evaluation across three standard benchmarks from Twitter and Reddit. This allows us to establish a strong, generalizable SOTA baseline and, more importantly, to analyze its performance and limitations across heterogeneous data sources, thereby yielding broader insights into the challenges of the task itself.

3 Datasets and Experimental Setup

To ensure a comprehensive and robust evaluation, we test our model on three widely-used benchmark datasets, each presenting distinct characteristics

Dataset	Source	Depressed Users	Control Users	Total Posts
CLPsych15	Twitter	477	872	3,110,414
MDL	Twitter	2,516	5,153	4,745,869
MDL-Filtered	Twitter	2,464	5,153	4,730,184
eRisk18	Reddit	214	1,493	1,076,582

Table 1: General details of the benchmark datasets used for training and evaluation.

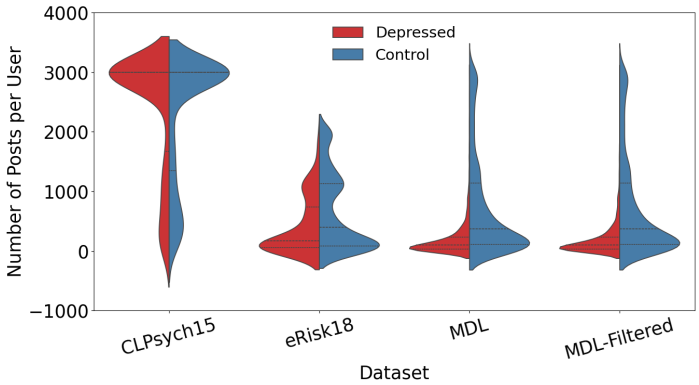

Figure 1: Distribution of post counts per user across datasets by class.

and challenges in terms of data source, labeling methodology, and content style.

3.1 Benchmark Datasets

The first dataset is the publicly accessible Twitter dataset known as **MDL** (Shen et al., 2017). It uses a keyword-based labeling heuristic: users who have posted tweets containing phrases like "I'm depressed" or "I've been diagnosed with depression" are labeled as positive. While this method allows for a large-scale dataset, it can introduce label noise. We also evaluate on **MDL-Filtered**, a more challenging variant where we remove tweets containing these explicit depression-related sentences, forcing the model to rely on more subtle linguistic cues. It is important to note that the specific statistics of the MDL dataset can vary between different studies due to ongoing updates.

The second dataset, **CLPsych 2015** (Coppersmith et al., 2015), was introduced for a shared task and also consists of Twitter posts. However, it employs a more reliable labeling strategy, where the positive class comprises users who have publicly self-disclosed a clinical diagnosis of depression. This provides a higher-quality, though smaller, benchmark.

The third dataset, **eRisk 2018** (Losada et al., 2018), is sourced from Reddit and presents a different set of challenges. It contains longer, more diverse posts, and its labels, while initially based on keywords, are manually reviewed for reliability. The long-form, context-rich nature of Reddit content makes this dataset particularly suitable for evaluating a model's ability to process complex user histories.

Table 1 provides a summary of these datasets.

3.2 Data Characteristics and Challenges

A key challenge in user-level depression detection is the high variance in user activity, as illustrated in Figure 1.

The violin plots reveal significant differences in the distribution of post counts across datasets. For example, while both CLPsych15 and the MDL datasets show a long-tail distribution, the CLPsych15 dataset's tail is notably "heavier," with a larger proportion of "power users" contributing thousands of posts, as indicated by the high density in the upper range of the violin plot. In contrast, the activity in MDL is more sharply concentrated at lower post counts, with a high concentration of users having few posts, despite the presence of a few highly active users.

The eRisk18 dataset is bimodal, with distinct groups of low- and high-activity users. This heterogeneity underscores a critical problem: a robust model must be capable of extracting meaningful textual patterns from both sparse user his-

tories (where every post is crucial) and extremely dense ones (where informative cues might be diluted by irrelevant content). This directly motivates our choice of an architecture designed to handle such varied input without being skewed by outliers or a few dominant posts.

3.3 Data Statement and Licensing

Intended Use and Licensing. The datasets used in this work (MDL, CLPsych15, eRisk18) are established benchmarks created by the research community for the specific task of mental health analysis from social media. Our use of these artifacts is consistent with their intended research purpose. The pre-trained sentence-transformer model, *multi-qa-mpnet-base-dot-v1*, is distributed under the Apache 2.0 license, which permits its use in this research. The datasets themselves were released for research purposes, and access is typically governed by data use agreements with the original creators.

Demographic Characteristics. A significant limitation of these benchmark datasets is the absence of demographic information about the user populations. Characteristics such as age, gender, geographic location, and race are not provided. This makes it impossible to assess the representativeness of the data or to analyze model performance across different demographic groups. As linguistic expressions of mental health can vary significantly across these groups, our model may contain hidden biases and may not generalize equitably to populations that are underrepresented in the original data. This knowledge gap underscores that our model should be considered a research contribution and not be deployed in any real-world application without further rigorous fairness and bias evaluation on representative data.

3.4 Validation Strategy

All three datasets exhibit a significant class imbalance between depressed and control users (see Table 1). To mitigate potential bias during training, we employ an under-sampling strategy, where for each training batch, we randomly sample the majority (control) class to match the number of samples from the minority (depressed) class. We evaluate model performance using a 5-fold cross-validation scheme. In each fold, we split the data into 80% for training and 20% for validation, ensuring that our final results are averaged over different, randomly selected subsets of the data.

4 Model Architecture

To address the core challenge of identifying distributed depressive patterns, we employ an end-to-end deep learning architecture focused on effective aggregation. Our approach treats a user's post history as an unordered set of semantic embeddings and learns a specialized function to synthesize these points into a single, discriminative user representation. The architecture, illustrated in Figure 2, consists of three main stages: post encoding, contextual aggregation, and classification.

1. Post Encoding. Each raw post x_i from a user U is first converted into a high-dimensional vector representation. We selected *multi-qa-mpnet-base-dot-v1* (Microsoft, 2021b) as our primary sentence-transformer based on preliminary experiments (see Appendix A.1 for a comparison). This transforms the user's collection of posts into a matrix $U' \in \mathbb{R}^{n \times 768}$, where n is the number of posts.

2. Contextual Aggregation via Self-Attention. To aggregate the post embeddings, we employ a self-attention mechanism (Vaswani, 2017). Unlike simple averaging, which treats all posts equally, self-attention allows the model to dynamically weigh the importance of each post embedding in the context of all other posts by the same user. This is crucial for identifying posts that are most indicative of depression, whether they are explicit cries for help or subtle linguistic markers, while down-weighting irrelevant content.

3. Feature Refinement and Classification. The contextually-weighted embeddings from the attention layer are then passed through a sequence of two depthwise 1D convolutional blocks. We employ depthwise 1D convolutions with a kernel size of 1 as an efficient method for learning channel-wise interactions. This acts as a lightweight, feature-wise fully connected layer, allowing the model to refine the aggregated user representation by learning complex patterns across the embedding dimensions. Each block uses the GeLU activation function (Hendrycks and Gimpel, 2016) for its smooth gradient properties, along with batch normalization to stabilize training and improve performance (Ioffe, 2015). Finally, a fully connected linear layer followed by a a sigmoid activation function maps the refined user representation to a final probability score for the "depressed" class. To handle the high variance in the number of posts per user, we set a maximum input sequence length of 1024 posts. For users with more than 1024 posts,

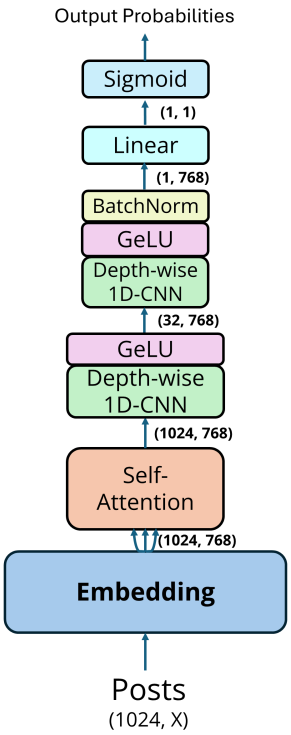

Figure 2: The proposed architecture, showing the flow from individual post embeddings to a final user-level classification.

we selected a random sample of 1024 posts. For users with fewer, we pad the input. For each user, this sampling is performed once before training begins for a given experimental run, ensuring the same set of posts is used across all epochs within that run.

4.1 Training Details

We trained the model for up to 250 epochs using the Adam optimizer. We froze the weights of the sentence-transformer encoder during training to preserve its rich, pre-trained semantic knowledge. The model demonstrated stable convergence across all datasets. For complete hyperparameter and training specifications, please refer to Appendix A.2.

5 Experimental Results

We evaluated our proposed model against prior state-of-the-art (SOTA) methods on the three benchmark datasets, demonstrating the consistent and significant performance improvements.

We report our results, averaged 50 runs of 5-fold cross-validation, in Table 2. We initialized each cross-validation experiment using a different random seed to ensure robustness. We report our model's performance as mean F1-score ± standard deviation. For the MDL datasets, where the comparative SOTA study (Han et al., 2022) provided sufficient statistical information, we performed Welch's t-test to assess statistical significance. For CLPsych15 and eRisk18, the original SOTA papers reported only mean scores without variance, precluding a direct significance test.

5.1 Performance on MDL Datasets

On the main MDL dataset, our model achieves a mean F1-score of 0.992, which is a statistically significant improvement over the previous SOTA of 0.973 ($p < 0.01$, Welch's t-test). To ensure a direct and fair comparison, we also evaluated our model on the specific train/validation/test split used by (Han et al., 2022), denoted as MDL*. Even on this split, our model demonstrates a clear improvement (0.987 vs. 0.973).

Crucially, on the more challenging MDL-Filtered dataset, from which we removed explicit depression-related keywords, our model again achieves a statistically significant performance gain, increasing the F1-score from 0.928 to 0.945 ($p < 0.05$, Welch's t-test). This result underscores our model's ability to identify depression from more subtle and nuanced linguistic cues, rather than relying solely on obvious keywords.

5.2 Performance on CLPsych15 and eRisk18 Datasets

On the CLPsych15 dataset, which uses self-disclosed diagnoses for labeling, our model achieves an F1-score of 0.910 ± 0.01, While this result significantly outperforms the previous benchmark reported by (Orabi et al., 2018), To our knowledge, this result remains the state-of-the-art for the user-level classification task on this dataset, as subsequent works have focused on the distinct task of post-level classification (Yang et al., 2023; Ji et al., 2023). The result primarily demonstrates our model's effectiveness on this high-quality, diagnosis-based dataset.

We observe significant improvement on the eRisk18 dataset, which is sourced from Reddit and features longer, more complex user posts. Our model achieves an F1-score of 0.88 ± 0.03. This represents a substantial leap of over 22 percentage points compared to the current deep learning-based SOTA of 0.655, set by the DeepBoSE model (Lara et al., 2021). This pronounced gain validates our approach of modeling the entire user history as a

Dataset	SOTA Method	SOTA F1	Our Model F1
MDL	HAN (Han et al., 2022)	0.973	**0.992 ± 0.002**
MDL*	HAN (Han et al., 2022)	0.973	**0.987 ± 0.002**
MDL-Filtered	HAN (Han et al., 2022)	0.928	**0.945 ± 0.005**
CLPsych15	CNN + BiLSTM (Orabi et al., 2018)‡	0.869	**0.910 ± 0.010**
eRisk18	DeepBoSE (Lara et al., 2021)	0.655	**0.88 ± 0.030**

Table 2: F1-scores of our model compared to previous SOTA. Our results are reported as mean ± standard deviation over 5 folds. The asterisk (*) denotes the specific data split from (Han et al., 2022) as described at section 5. ‡To our knowledge, this is the most recent SOTA for the user-level classification task on this dataset; subsequent works have focused on the task of post-level classification (Yang et al., 2023; Ji et al., 2023).

set to capture distributed textual patterns.[1]

6 Model Introspection and Error Analysis

To move beyond performance metrics and understand the internal workings of our model, we conducted a multi-faceted analysis. This section explores how our model transforms raw post embeddings into meaningful representations and investigates the primary causes of misclassifications.

6.1 Model Introspection: Visualizing Feature Transformation

To visualize how our architecture learns to distinguish between depressed and non-depressed users, we used Principal Component Analysis (PCA) to project the average user embeddings into a 3D space at different stages of the model. Figure 3 illustrates this transformation process on the CLPsych15 dataset.

Initially, the raw sentence-transformer embeddings (a) show significant overlap between the two classes, indicating that a simple aggregation of posts is insufficient for accurate classification. After the self-attention mechanism is applied (b), the clusters become more distinct. This demonstrates the effectiveness of the attention layer in re-weighting the post embeddings to emphasize salient markers and suppress noise. Finally, after passing through the depthwise convolutional layers (c), the user representations achieve near-complete linear separability. This visualization confirms that our architecture successfully learns to map a complex, entangled input space into a feature space that is highly discriminative.

[1]Other works, such as (Ji et al., 2021), have reported high F1-scores on eRisk18. However, their reported dataset statistics differ significantly from the official release, suggesting the use of a non-standard version of the dataset and precluding a direct comparison.

6.2 Error Analysis: Why Does the Model Fail?

While the model demonstrates strong performance, analyzing its errors provides critical insights. We conducted a systematic error analysis by quantifying each post's contribution to the final prediction (its "post weight"), measured by the change in output probability upon its removal. This allowed us to categorize misclassifications into two primary types: those driven by a single dominant post and those resulting from an aggregation of weaker, distributed cues.

As Figure 4 shows, the vast majority of misclassifications fall into the "spread-out attention" category, where no single post dominates the decision. This suggests that the model's primary challenge lies not in misinterpreting a single strong textual cue, but in correctly aggregating numerous subtle ones.

6.2.1 Errors Driven by a Dominant Post

In a smaller subset of errors, a single post with a high weight (> 0.5) was responsible for the misclassification. We identified two main causes for these errors:

- **Contextual Ambiguity:** Posts contained trigger words (e.g., "depressed," "anxious") but used them in a non-personal context (e.g., discussing a news article about depression). The model, in these cases, over-relied on the keyword, leading to false positives. For instance, a control user was misclassified from the post "I truly believe, diagnosed depression can never be truly fixed," where the model focused on keywords ("diagnosed depression") while missing the non-personal context. This failure to distinguish topic commentary from personal experience highlights the need for

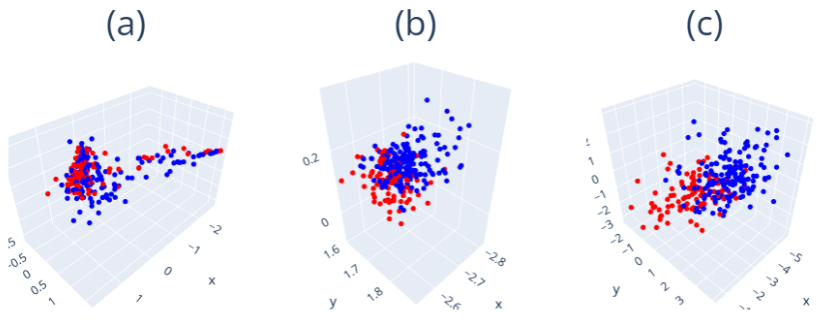

Figure 3: PCA visualization of average user embeddings from the CLPsych15 dataset at three key stages: (a) raw sentence embeddings, (b) after self-attention, and (c) the final representation before classification.

more sophisticated pragmatic understanding.

- **Label Noise:** We found evidence of mislabeled data in the ground truth, accounting for approximately 16% of misclassifications in the MDL and CLPsych datasets. For example, users with explicit self-reports of a diagnosis were sometimes labeled as control, leading to false negatives.

6.2.2 Errors Driven by Distributed Cues

This more common type of error occurs when the model must aggregate cues across many posts. Our analysis, presented in Figure 5, reveals a nuanced relationship between post count and error type.

As expected, users with very few posts (< 100) are difficult to classify due to data sparsity. More surprisingly, a large cluster of errors, primarily false positives, occurs for users with a high volume of posts ($\sim$3000). This analysis reveals a key source of error in our sampling strategy: for users with a high volume of posts, important signals can be difficult to extract. For users exceeding the 1024-post limit, our random sampling may inadvertently omit critical depressive indicators (causing false negatives) or aggregate transient negative sentiments that are not indicative of clinical depression (causing false positives). The model's high confidence in its incorrect predictions (see Appendix A.3, Figure 7) further suggests that for these high-volume users, the aggregated "noise" can be misinterpreted as a strong, albeit incorrect, pattern.

6.3 Implications of Analysis

Our analysis indicates that the primary obstacles to improving user-level depression detection are no longer about encoding individual posts, but about modeling the user as a whole. While our model can see past simple keywords (as shown on the MDL-Filtered dataset), it struggles with two deeper user-level challenges: **pragmatic context** (distinguishing personal disclosure from general commentary) and **signal representation** (effectively sampling from long post histories without losing critical information). These findings suggest a necessary shift in the field's focus: from engineering better post encoders to developing more sophisticated user modeling techniques. Future work must prioritize signal-driven sampling or summarization strategies that can robustly handle the extreme variance in user activity.

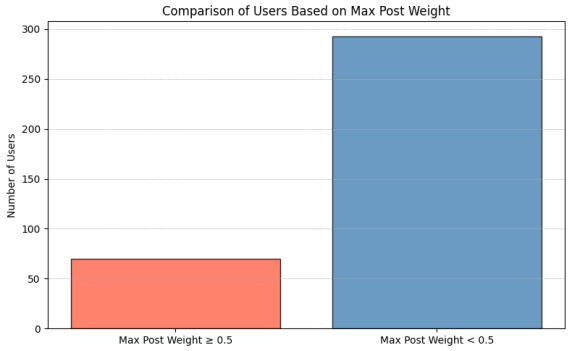

Figure 4: Distribution of misclassified users based on the maximum weight of any single post. Most errors occur when attention is spread out across many posts rather than being dominated by a single, high-weight post.

7 Ablation Study

To rigorously verify the importance of our architectural choices, we conducted ablation studies on

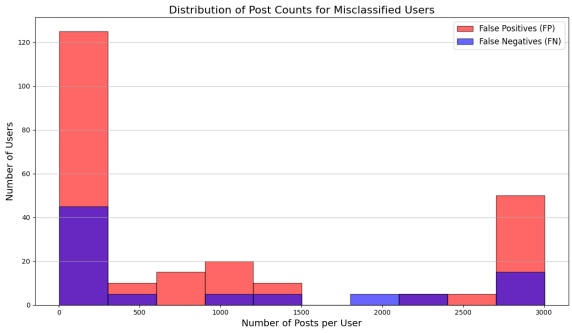

Figure 5: Distribution of False Positives (FP) and False Negatives (FN) by user post count. Errors cluster at both low and high post counts, indicating challenges with both data sparsity and signal loss in dense histories.

the CLPsych15 and eRisk18 datasets. We systematically removed or replaced key components from our full model and measured the impact on F1-score. We present the results in Table 3.

Model Configuration	CLPsych15 F1	eRisk18 F1
Our Full Model	**0.91**	**0.88**
(1) w/o Self-Attention	0.71	0.80
(2) w/o Depthwise Conv.	0.84	0.86
(3) Using basic S-BERT	0.83	0.86

Table 3: Ablation study results showing the drop in F1-score after removing or replacing key model components. Results are the mean F1 over 5-fold cross-validation.

The results strongly validate our design. Removing the **self-attention mechanism** (1) led to the most significant performance degradation across both datasets. Notably, the drop was most pronounced on the higher-quality CLPsych15 dataset (a 20-point drop in F1-score), confirming that the attention mechanism is critical for identifying genuine, subtle linguistic cues, not just for filtering noise. Removing the **depthwise convolutional layers** (2) also resulted in a clear drop in performance, underscoring their importance in refining the aggregated user representation. Finally, using a smaller, more generic sentence-transformer (3) instead of our chosen encoder further demonstrates that a powerful base encoder is essential for achieving state-of-the-art results.

8 Conclusions and Future Work

This paper presented a large-scale empirical study that establishes a new and robust state-of-the-art baseline for user-level depression detection. By systematically applying a modern architecture leveraging powerful pretrained sentence embeddings and a global self-attention mechanism, we have demonstrated significant performance gains across three major public benchmarks. Our primary contribution is the rigorous cross-dataset validation and the subsequent in-depth analysis. Our work provides both a new performance benchmark for the field and critical insights to guide future research.

Our analysis revealed a key limitation of our method centered on its handling of users with extreme variations in post volume. For those with sparse histories, the lack of data is a persistent challenge. More critically, for highly active users, we found that our random sampling strategy can omit crucial linguistic markers or over-amplify transient noise. This highlights a difficult trade-off between capturing a full user history and managing the signal-to-noise ratio, a core problem that must be addressed in subsequent research.

Promising directions for future research include:

1. **Intelligent Post Selection:** Moving beyond random sampling by using signal-driven sampling or summarization techniques to better handle long user histories.

2. **Incorporating Temporal Dynamics:** Modeling the evolution of language over time, rather than treating posts as an unordered set, to detect the onset of depressive episodes.

3. **Improving Data Curation:** Addressing the label noise found in our error analysis through more rigorous validation of user labels to create a cleaner ground truth.

4. **Assessing Generalizability and Fairness:** Evaluating model performance on diverse multilingual and cross-cultural data to ensure it is both accurate and equitable.

9 Limitations

While our model achieves state-of-the-art results, we acknowledge several limitations that define the scope of our claims and highlight important directions for future work.

Assumptions and Robustness. A core assumption of our method is that a fixed-size random sample of 1024 posts is a sufficient representation for users with extensive histories. Our own error analysis directly challenges this assumption, identifying information loss in long post histories as

a key limitation of our approach. This indicates our results are not robust to violations of this assumption; the model's performance likely degrades when critical, sparse cues are omitted by the random sampling process in highly active users. This implies that while effective, our model's aggregation mechanism is sensitive to the quality of its input sample, and its performance on "power users" may be overstated. Future work must explore signal-driven, rather than random, sampling to mitigate this vulnerability. Furthermore, our model assumes that the linguistic patterns learned are generalizable across datasets with different labeling heuristics (keyword-based vs. self-disclosure), which may not hold true and could be a source of noise. We also acknowledge that datasets labeled via keyword-matching, such as MDL, may contain significant label noise and may not accurately reflect clinical diagnoses, potentially impacting the model's ability to learn genuine psychological signals.

Scope of Claims and Generalizability. Our empirical validation is conducted on three English-language datasets from two platforms (Twitter and Reddit). Therefore, our claims of SOTA performance are strictly limited to this context. The model's efficacy on other languages, cultural contexts, or social media platforms remains unevaluated. The expression of mental health can vary significantly across cultures, and our model may fail to generalize. The performance is reported as the mean and standard deviation over 5-fold cross-validation, which provides a measure of stability. However, the variance across different runs does not account for the variance that would arise from different datasets or demographic subgroups, which are known to influence language use.

Potential for Negative Societal Impact. Our model is trained on data where the ground truth is imperfect and may contain demographic biases, as discussed in Section 3.3. The model could inadvertently learn to associate the linguistic patterns of certain demographic groups with depression more than others, leading to inequitable performance. It is critical that this research tool is not misconstrued as a diagnostic instrument, as misclassification could lead to significant harm or stigmatization.

10 Ethical Considerations and Potential Risks

The development of automated systems for mental health monitoring carries significant ethical responsibilities. The primary concern is user privacy. All data used in this study were publicly available at the time of collection. To protect privacy, all user identifiers were replaced with anonymous hashes, and no personal metadata was retained during our analysis.

Furthermore, we acknowledge that social media data often contains sensitive or offensive content. We did not filter this content, as it can form part of the linguistic signal relevant to mental health expression, but this means the datasets contain potentially disturbing text.

It is crucial to emphasize that models like the one proposed here are intended as research tools for understanding linguistic patterns, not as diagnostic instruments. A prediction from the model should never be interpreted as a clinical diagnosis. Any real-world application would require strict protocols, including user consent, data security, and oversight by qualified mental health professionals to prevent potential harm, misinterpretation, or stigmatization. The risk of dual use is significant; a misuse of this technology could lead to automated, non-consensual flagging of individuals.

Finally, as noted in Section 3.3, the lack of demographic information means our model may not perform equitably across different groups. The deployment of such a system without rigorous fairness audits could perpetuate or amplify existing societal biases, potentially leading to certain populations being over- or under-diagnosed, causing significant harm.

Statement on AI-Assisted Writing

We utilized AI-powered language models (Gemini and ChatGPT) to assist with text rephrasing and formatting during manuscript preparation. All research ideas, experimental design, analysis, and conclusions were conceived and developed solely by the authors.

References

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings*, pages 31–39.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv Preprint arXiv:1810.04805*.

Aron Halfin. 2007. Depression: The benefits of early and appropriate treatment. *American Journal of Managed Care*, 13(4):S92.

Sooji Han, Rui Mao, and Erik Cambria. 2022. Hierarchical attention network for explainable depression detection on twitter aided by metaphor concept mappings. *arXiv preprint arXiv:2209.07494*.

Ayaan Haque, Viraaj Reddi, and Tyler Giallanza. 2021. Deep learning for suicide and depression identification with unsupervised label correction. In *Springer*, pages 436–47.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv Preprint arXiv:1606.08415*.

Sergey Ioffe. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv Preprint arXiv:1502.03167*.

S Ji, T Zhang, L Ansari, J Fu, P Tiwari, and E Cambria. 2021. Mentalbert: publicly available pretrained language models for mental healthcare. arxiv. *Preprint posted online on December*, 29.

Shaoxiong Ji, Tianlin Zhang, Kailai Yang, Sophia Ananiadou, Erik Cambria, and Jörg Tiedemann. 2023. Domain-specific continued pretraining of language models for capturing long context in mental health. *arXiv preprint arXiv:2304.10447*.

S Kayalvizhi and D Thenmozhi. 2022. Data set creation and empirical analysis for detecting signs of depression from social media postings. *Archives of Computational Methods in Engineering*.

Ronald C Kessler, Sergio Aguilar-Gaxiola, Jordi Alonso, Somnath Chatterji, Sing Lee, Johan Ormel, T Bedirhan Üstün, and Philip S Wang. 2009. The global burden of mental disorders: An update from the who world mental health (wmh) surveys. *Epidemiology and Psychiatric Sciences*, 18(1):23–33.

J Kim, J Lee, E Park, and J Han. 2020. A deep learning model for detecting mental illness from user content on social media. *Scientific Reports*, 1:1–6.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Proceedings*, volume 29.

Juan S Lara, Mario Ezra Aragón, Fabio A González, and Manuel Montes-y Gómez. 2021. Deep bag-of-sub-emotions for depression detection in social media. In *International Conference on Text, Speech, and Dialogue*, pages 60–72. Springer.

Huynh-Nhu Le and Rhonda C Boyd. 2006. Prevention of major depression: Early detection and early intervention in the general population. *Clin Neuropsychiatry*, 3(1):6–22.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv Preprint arXiv:1907.11692*, 364.

D E Losada, F Crestani, and J Parapar. 2018. Overview of erisk: Early risk prediction on the internet. In *International conference of the Cross-Language Evaluation Forum for European*, pages 343–361.

Microsoft. 2020. all-mpnet-base-v2: Pretrained sentence transformer. https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

Microsoft. 2021a. all-MiniLM-L6-v2: Sentence embedding model for semantic search. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

Microsoft. 2021b. multi-qa-mpnet-base-dot-v1: Sentence embedding model for semantic search. https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1.

Ricardo F Muñoz, Patricia J Mrazek, and Robert J Haggerty. 1996. Institute of medicine report on prevention of mental disorders: Summary and commentary. *American Psychologist*, 51(11):1116.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of twitter users. In *Proceedings*, pages 88–97.

A Picardi, I Lega, L Tarsitani, M Caredda, G Matteucci, MP Zerella, R Miglio, A Gigantesco, M Cerbo, and A Gaddini. 2016. A randomised controlled trial of the effectiveness of a program for early detection and treatment of depression in primary care. *Journal of Affective Disorders*, 198:96–101.

P Resnik, W Armstrong, L Claudino, and T Nguyen. 2015. The university of maryland clpsych 2015 shared task system. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 54–60.

Kathryn Rost, Jeffrey L Smith, and Miriam Dickinson. 2004. The effect of improving primary care depression management on employee absenteeism and productivity: A randomized trial. *Medical Care*, 42(12):1202–10.

Shahid Munir Shah, Syeda Anshrah Gillani, Mirza Samad Ahmed Baig, Muhammad Aamer Saleem, and Muhammad Hamzah Siddiqui. 2025. Advancing depression detection on social media platforms through fine-tuned large language models. *Online Social Networks and Media*, 46:100311.

G Shen, J Jia, L Nie, F Feng, C Zhang, T Hu, T S Chua, and Zhu. 2017. Depression detection via harvesting social media: A. *IJCAI*, pages 44883–44893.

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*, volume 33, pages 16857–67.

Waleed Bin Tahir, Shah Khalid, Sulaiman Almutairi, Mohammed Abohashrh, Sufyan Ali Memon, and Jawad Khan. 2025. Depression detection in social media: A comprehensive review of machine learning and deep learning techniques. *IEEE Access*.

Joel Philip Thekkekara, Sira Yongchareon, and Veronica Liesaputra. 2024. An attention-based cnn-bilstm model for depression detection on social media text. *Expert systems with applications*, 249:123834.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

World Health Organization. n.d. Comprehensive mental health action plan 2013–2030. https://www.who.int/initiatives/mental-health-action-plan-2013-2030.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. *arXiv preprint arXiv:2304.03347*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, 28.

Zhenwen Zhang, Jianghong Zhu, Zhihua Guo, Yu Zhang, Zepeng Li, and Bin Hu. 2024. Natural language processing for depression prediction on sina weibo: method study and analysis. *JMIR Mental Health*, 11:e58259.

Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis Lau. 2015. A c-lstm neural network for text classification. *arXiv Preprint arXiv:1511.08630*.

Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. 2021. Depressionnet: Learning multimodalities with user post summarization for depression detection on social media. In *Proceedings*, pages 133–42.

A Appendix

A.1 Encoder Selection

The choice of the sentence-transformer encoder is critical for generating high-quality post embeddings. We evaluated several strong candidates:

- **multi-qa-mpnet-base-dot-v1** (Microsoft, 2021b): An encoder based on MPNet (Song et al., 2020) and fine-tuned on numerous Q&A datasets, designed for semantic search tasks.
- **all-mpnet-base-v2** (Microsoft, 2020): A general-purpose model fine-tuned on a massive and diverse dataset, often serving as a strong baseline.
- **all-MiniLM-L6-v2** (Microsoft, 2021a): A compact and efficient sentence-transformer (22.7M parameters) known for its excellent balance of speed and performance, making it a strong candidate for general-purpose text embedding tasks.

While *all-mpnet-base-v2* provided strong performance, we observed a consistent, albeit slight, improvement using *multi-qa-mpnet-base-dot-v1*, which we therefore selected as our primary encoder. The smaller *all-MiniLM-L6-v2* model, while efficient, did not yield any performance gains over the MPNet-based models.

We present the validation F1-scores for each encoder on two key datasets in Table 4.

Encoder	CLPsych15 F1	eRisk18 F1
multi-qa-mpnet-base-dot-v1	**0.912**	**0.896**
all-mpnet-base-v2	0.853	0.878
all-MiniLM-L6-v2	0.827	0.877

Table 4: Comparison of encoder performance on the validation set (mean F1-score over 5 folds).

A.2 Hyperparameter and Training Details

We performed a grid search to determine the optimal hyperparameters, conducting tests on the first fold of the validation set for each dataset. The search space included batch size (16, 32, 64), number of posts per user (512, 1024, 2048), and number of attention heads (4, 8, 16). The final configuration that yielded the best and most stable performance was a **batch size of 32**, a maximum of **1024 posts per user**, and **16 attention heads** in the self-attention block. We trained the model for up to 250 epochs using the Adam optimizer with an initial learning rate of 1e-4 and a dropout rate of 0.2 in the self-attention block. To ensure robust and reproducible results, we initialized each cross-validation experiment with a different random seed.

Computational Resources and Software. All experiments were conducted on a single NVIDIA GeForce RTX 3070 Ti GPU with 8GB of memory. The sentence-transformer encoder (*multi-qa-mpnet-base-dot-v1*) has approximately 110 million parameters, the embeddings are calcualted before training to reduce training time. The trainable part

of our model (the aggregation head) consists of approximately 3 million parameters. The total computational budget for all 5-fold cross-validation runs across all datasets amounted to approximately 25 GPU hours. The implementation was done using Python 3.10, PyTorch 2.2.1, scikit-learn 1.4.1, Numpy 1.26.4 and the ‘sentence-transformers‘ library (v2.6.1).

A.3 Additional Analysis Figures

This section contains supplementary figures referenced in the main text. Figure 6 shows the training and validation loss curves, demonstrating stable model convergence. Figure 7 visualizes the distribution of the model’s predicted probabilities for misclassified users, supporting the error analysis in Section 6.

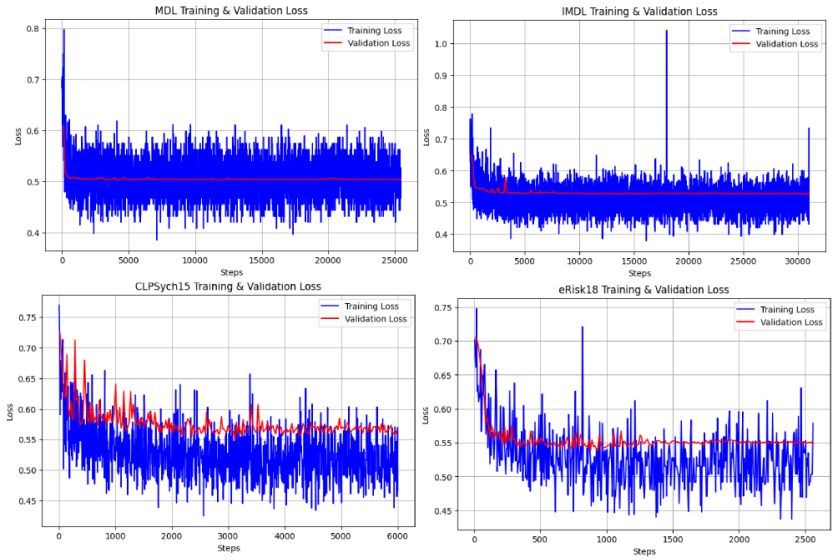

Figure 6: Training and validation loss curves across the four datasets. The model demonstrates stable convergence in all cases.

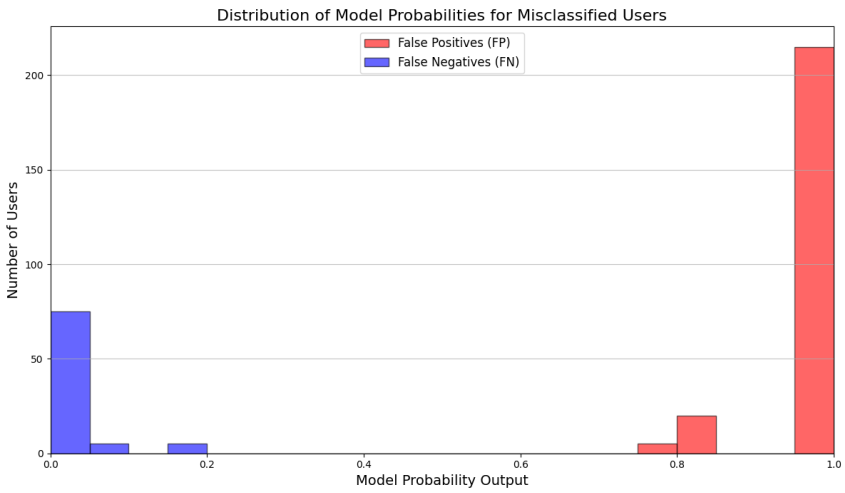

Figure 7: Distribution of the model's predicted probabilities for misclassified users. The model is often highly confident in its incorrect predictions for both FPs and FNs.