Therapist by Chance: Investigating ChatGPT's Emotional and Mental Health Support via Sentiment Analysis on Social Networks

Smita Ghosh[1], Xiaochen Luo[1], Jared Maeyama[1*], Shiv Jhalani[1*], CJ Oshiro[1†], Tharun Venkatesh[2†], and Rushil Patel[1]

[1] Santa Clara University, Santa Clara, CA, USA,
{sghosh3, xluo, jmaeyama, sjhalani, coshiro, rmpatel}@scu.edu
[2] University of Waterloo, Ontario, Canada,
tvenkate@uwaterloo.ca

Abstract. As large language models (LLMs) increasingly engage in emotionally sensitive interactions, their unintended therapeutic roles demand systematic investigation. We examine ChatGPT's perceived capacity to provide emotional and mental health support by analyzing user-generated content from social networks through relevance classification and sentiment analysis. Guided by three research questions, we quantify public sentiment toward ChatGPT in therapeutic contexts. To identify posts suggesting therapeutic use of ChatGPT, we introduce two methods: *SemReC*, a supervised relevance classification, and *PASS*, an unsupervised similarity-based approach. Both methods demonstrate high accuracy and consistent performance across the dataset. We further assess the performance of existing pre-trained sentiment analysis models to benchmark their effectiveness. To capture affective sentiment propagation in multi-turn interactions, we propose two tree-structured methods—*HierSent* and *AggSent*—which model emotional dynamics within threaded conversations. Empirical results validate the effectiveness of our methods and reveal a predominance of positive sentiment toward using ChatGPT for therapeutic purposes. These findings highlight the public popularity of the emergent therapeutic use of LLMs and underscore the need to examine their broader implications for mental health.

Keywords: Sentiment Analysis, Social Media Mining, Mental Health Informatics

1 Introduction

The rapid rise of generative AI, exemplified by ChatGPT, has fundamentally altered how individuals interact with technology, prompting new discourse on AI's role in both the functional and emotional facets of human life [1–3]. An

[*] Jared Maeyama and Shiv Jhalani contributed equally.
[†] CJ Oshiro and Tharun Venkatesh contributed equally.

emerging trend indicates that individuals—particularly younger users—are increasingly turning to ChatGPT for emotional support and therapeutic guidance, despite the model not being designed or regulated for mental health applications [4,5]. In response to this trend, OpenAI introduced policy updates in early 2023 that included disclaimers advising users not to rely on ChatGPT for mental health guidance [6–8]. Nonetheless, the continued use of AI tools such as ChatGPT for mental health topics underscores an emergent role that LLMs may play in meeting users' needs for emotional and mental health support, highlighting a persistent gap between the huge need for mental health care and the lack of accessibility to affordable, timely services amid a global mental health crisis [5,9,10].

Although prior work has highlighted the potential of AI to augment mental health support, significant concerns persist regarding its capacity for empathetic engagement, user safety, and the ethical safeguards required for responsible deployment [11–13]. This paper examines public sentiment towards ChatGPT's perceived therapeutic role by analyzing user-generated content on Reddit and X (formerly Twitter). Rather than directly assessing AI's responses or therapeutic effectiveness, we focus on how the general public perceive and discuss ChatGPT in emotionally supportive or therapeutic roles. Through relevance classification and sentiment analysis, we evaluate the polarity of these discussions to determine whether ChatGPT is perceived as helpful, neutral, or potentially harmful in the domain of mental health and emotional support.

Aligned with the framing of ChatGPT as a "therapist by chance," this data-driven analysis examines whether large language models are increasingly perceived by the general public as unintended providers of emotional and mental health support, raising important questions about user expectations and the ethical boundaries of AI-human interactions for mental health.

The key contributions of this work are as follows:

- **Therapeutic Relevance Classification:** To identify posts that meaningfully discussed ChatGPT's therapeutic role in mental health and emotional support, we introduce two methods for filtering relevant content: (i)***SemReC***, a supervised semantic relevance classification leveraging sentence embeddings and k-NN, and (ii)***PASS***, an unsupervised polarity-aligned similarity scoring method, to categorize whether a post is relevant or not on the topic of using ChatGPT for emotional and mental health support.
- **Evaluation of Pre-trained Sentiment Analysis Models:** We conduct a comparative analysis of existing pre-trained sentiment analysis models to assess their ability to capture sentiment in user posts for ChatGPT's role in mental health support.
- **Context-aware Sentiment Modeling via Tree Structures:** We introduce two models—*HierSent* (top-down) and *AggSent* (bottom-up)—that propagate sentiment through conversation trees. These models capture hierarchical context, enabling the detection of sentiment shifts in threaded discussions.

2 Related Work

2.1 Large Language Models as Emotional Support Agents

Large language models (LLMs), such as ChatGPT, are increasingly used for emotional support and mental health advice [1, 2, 14, 15]. Studies show that a significant portion of LLM interactions involve users seeking empathy, reassurance, or emotional support. In these contexts, LLMs can generate contextually appropriate and supportive responses [1, 14, 16–18]. However, their guidance can be inconsistent, sometimes exhibiting biased behavior and risking the spread of misinformation, especially in sensitive contexts [2, 19]. These limitations highlight the need for transparency and safeguards in deploying LLMs for mental health support.

2.2 Multi-Turn Dialogue and Emotion Progression

Tree-structured models have advanced sentiment analysis by modeling syntactic and structural relationships within text. Prior work includes tree communication networks combining graph convolution and recurrence [20], RST-based models that use text structure and nucleus-satellite relations [21], and capsule tree-LSTMs that reduce root node bias using dynamic routing [22]. Hierarchical models have also been applied to aspect-based sentiment by aggregating sentence-level embeddings [23], while recent methods use latent parsing and TreeCRFs to identify sentiment spans as subtrees [24]. Despite these advances, existing work is largely limited to static, single-document or sentence-level trees. To our knowledge, there is little prior research that systematically adapts tree-structured sentiment models to multi-turn conversation trees. This motivates our proposed models, designed to capture evolving sentiment in threaded interactions.

3 Data Collection

The goal of this study is to summarize user-generated content to reflect therapeutic engagement with ChatGPT. To this end, we collected posts from Reddit and X (formerly Twitter) using a query-based retrieval strategy. Targeted keyword queries were designed to find discussions referencing ChatGPT in emotional or supportive contexts, including phrases such as “ChatGPT therapy”, “ChatGPT emotional support,” “ChatGPT friends,” “ChatGPT mental health support,” and “ChatGPT personal advice.” These queries aimed to capture a broad range of interactions in which users implicitly or explicitly framed the model as a source of emotional support, guidance, or informal companionship. Following retrieval, the data was preprocessed to ensure topical relevance, structural consistency, anonymity, and overall integrity.

Using the X API [25] and Reddit API [26], we collected posts from January 1, 2023, to September 31, 2024. Identical keyword queries were applied across both platforms, though the resulting data formats differed substantially. X yielded 7,049 standalone posts accompanied by metadata, including timestamps and full text content. In contrast, Reddit's structure enabled the extraction of 182 complete conversation threads, organized as tree-structured dialogues to support multi-turn interaction analysis. All data were publicly accessible and collected in accordance with platform terms of service and established research ethics guidelines.

4 Proposed Approach for Relevance Classification

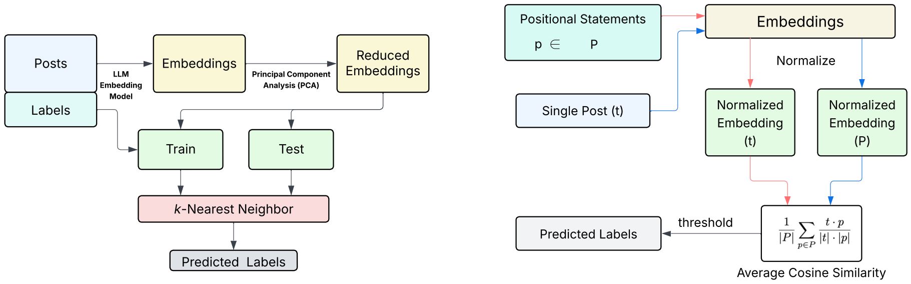
Despite targeted keyword queries, many retrieved posts were irrelevant to ChatGPT's therapeutic role, often matching terms superficially rather than reflecting genuine emotional or mental health support. Some used "therapy" metaphorically or referenced mental health in unrelated contexts, motivating a dedicated relevance classification step. For instance, a post retrieved using the query "ChatGPT therapy" read: "@anonymous_user: I will join the trend. I asked chatgpt to write the most original dad joke ever written. The outcome was: Why did the computer go to therapy? Because it had too many unresolved issues and couldn't find its cache flow!" This illustrates a case where the keyword "therapy" is used with no actual reference to emotional support or mental health, highlighting the need for dedicated relevance filtering.

RQ1: To what extent do user interactions with ChatGPT exhibit characteristics indicative of therapeutic engagement or intent?

In this study, relevance is defined as either the presence of emotional or psychological engagement—or therapeutic effect, where users derive support regardless of intent. However, identifying such relevance at scale presents practical challenges. Manual annotation of the collected dataset was infeasible due to its volume. To establish a ground-truth reference, a representative subset of 180 posts was manually annotated by human raters as either relevant or irrelevant, with relevant posts further labeled by sentiment polarity (positive or negative). To address the limitations of manual annotation and enable scalable relevance classification, we propose two automated methods: a supervised semantic classification based on k-nearest neighbors using sentence embeddings (**SemReC**), and an unsupervised similarity-based scoring approach anchored on polarity-aligned exemplars (**PASS**). These methods provide efficient, interpretable relevance estimates consistent with human annotations.

The **Sem**antic **Re**levance **C**lassification (**SemReC**) embeds each post using a pre-trained sentence transformer model (text-embedding-3-small [27] from OpenAI). Dimensionality was reduced using Principal Component Analysis (PCA) to suppress noise and improve signal fidelity. A k-Nearest Neighbors (k-NN) classifier, trained on the labeled subset described above, predicts relevance labels by exploiting semantic proximity within the dimensionality-reduced embedding space. Figure 1a illustrates the overall architecture of **SemReC**.

The **P**olarity-**A**ligned **S**imilarity **S**coring (**PASS**) method leverages a curated set of polarity-anchored reference sentences. Figure 1b illustrates the PASS



(a) **SemReC architecture. Posts are embedded using a pre-trained LLM, reduced via PCA, and classified with k-NN.**

(b) **PASS architecture. Post embeddings are compared to polarity-anchored exemplars using cosine similarity.**

Fig. 1: **Proposed Therapeutic Relevance Classification Architectures.**

architecture for unsupervised relevance classification. Each input post t is embedded using a pre-trained embedding model [27], alongside a fixed set of polarity-anchored reference statements $\{p \in \mathbf{P}\}$ that represent prototypical therapeutic and non-therapeutic expressions. All embeddings are L2-normalized to enable cosine similarity comparison. The average cosine similarity between the input post (t) and all reference statements $(\mathbf{P})$ is computed as follows:

$$\mathrm{SimScore}(t, \mathbf{P}) = \frac{1}{|\mathbf{P}|} \sum_{p \in \mathbf{P}} \frac{t \cdot p}{\|t\| \, \|p\|} \tag{1}$$

where $t \cdot p$ denotes the dot product between vectors t and p, and $\| \cdot \|$ denotes the Euclidean norm.

$$\mathrm{Label}(t) = \begin{cases} \text{Relevant} & \text{if } \mathrm{SimScore}(t, P) > \tau \\ \text{Irrelevant} & \text{otherwise} \end{cases} \tag{2}$$

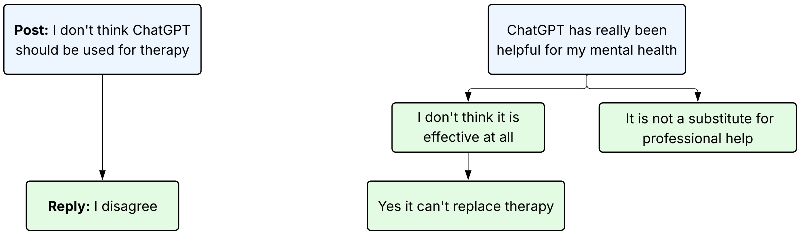
where τ is a user defined similarity threshold.

This approach leverages semantic alignment with curated exemplars to enable scalable, label-free inference of therapeutic relevance.

5 Proposed Approach for Sentiment Analysis

This section presents methods for analyzing ChatGPT's unintended therapeutic role using a relevance-filtered dataset (Section 4). We evaluate pre-trained sentiment models and examine conversation trees to capture therapeutic dynamics and context-aware emotional engagement.

RQ2: How accurately do existing pre-trained sentiment analysis models capture the sentiment expressed in user posts concerning ChatGPT's perceived therapeutic role? To address this question, we evaluate several off-the-shelf sentiment analysis models trained on general-purpose corpora. Performance is assessed and results are reported in Section 6.2. These


Post: I don't think ChatGPT should be used for therapy
Reply: I disagree
ChatGPT has really been helpful for my mental health
I don't think it is effective at all
It is not a substitute for professional help
Yes it can't replace therapy

(a) **Contextual sentiment reversal.** A reply such as “I disagree” appears negative in isolation but implies support when disagreeing with a negative statement.

(b) **Conversation tree with disagreement propagation.** A positive post is followed by critical replies at multiple levels.

Fig. 2: **Examples highlighting the importance of conversational structure in sentiment interpretation.** (a) Reversals in polarity due to disagreement. (b) Multi-turn disagreement dynamics within a conversation tree.

models vary in architecture and design, offering a diverse baseline for assessing sentiment classification performance in emotionally expressive posts referencing ChatGPT. **BERT** [28] learns deep contextual representations via masked language modeling (MLM) and next sentence prediction (NSP), improving performance on downstream NLP tasks. **RoBERTa** [29], a BERT variant, removes NSP, uses dynamic masking, and is trained on more data, yielding superior benchmark results for sentiment classification. **VADER** [30] is a rule-based model tailored for social media, leveraging lexical and syntactic cues to score sentiment intensity, making it effective for short, informal posts. The **Microsoft AI Builder** sentiment model [31] assigns sentiment labels with confidence scores across multiple languages, offering scalable and real-time classification in applied contexts.

RQ3: Can sentiment analysis over conversation trees uncover latent emotional progression in ChatGPT–user interactions on social platforms? To examine emotional progression in ChatGPT–user interactions, conversation threads were modeled as trees, where nodes represent posts or replies and edges denote reply links. This structure preserves conversational flow, allowing sentiment dynamics to be tracked by embedding each node with sentiment scores and modeling dependencies between posts and their replies.

Tree-based representations reveal patterns like sentiment shifts, and peer validation that flat analyses often miss. For instance, a reply that challenges a negative post about ChatGPT may signal support, despite using negative language (Figure 2a). Similarly, positive root posts followed by skeptical replies may lead to an overall negative tone (Figure 2b). Such context-dependent sentiment shifts underscore the need for structured models that embed each post’s sentiment relative to its conversational context. Conversation trees capture these dependencies, enabling sentiment inference based on the emotional progression of an exchange rather than treating each statement in isolation.

To leverage the conversation tree structure, two models are proposed under the framework of **Context-Embedded Sentiment Model**, incorporating both top-down and bottom-up traversal strategies. The top-down approach, termed ***Hier****archical* ***Sent****iment Propagation (****HierSent****)*, models how sentiment flows from the root post to its descendants, capturing the influence of initial emotional framing on subsequent replies. The bottom-up approach, referred to as ***Agg****regated* ***Sent****iment Propagation (****AggSent****)*, aggregates sentiment from lower-level responses to infer the evolving sentiment of higher-level posts.

Hier*archical* ***Sent****iment Propagation (****HierSent****)* is a model proposed to capture contextual sentiment by representing each comment's sentiment as either $+1$ (positive) or -1 (negative) or 0 (neutral). To quantify the influence of conversational structure on sentiment interpretation, each node in the conversation tree is represented as a tuple (s, s^*), where s and s^* denote the initial stance and context-aware sentiment of the node, respectively, with $s, s^* \in \{-1, 0, +1\}$. Stance refers to the sentiment of a post with respect to its parent, capturing agreement, neutrality, or disagreement within the local conversational context. For example, consider the parent post: "ChatGPT should be used as a therapist." A reply like "Yes, it helps" shows agreement, "Maybe, but it needs safeguards" is neutral, and "That's a terrible idea" expresses disagreement.

The tree is modeled as a directed graph rooted at the original post, with edges representing reply relationships. For the root node r, the context-aware sentiment is equal to its initial sentiment: $s_r^* = s_r$. For every other node u with parent node $p(u)$, the context-aware sentiment is defined by a top-down dependency relation: $s_u^* = s_{p(u)}^* \cdot s_u$. This formulation captures how sentiment shifts through agreement or disagreement within the conversational structure.

The overall sentiment of the conversation tree is given by the sum of context-aware sentiment values across all nodes: $S_{\text{tree}} = \sum_{u \in T} s_u^*$, where T denotes the set of all nodes in the tree. This aggregated score provides a structurally informed measure of the conversation's overall emotional polarity. An illustrative example is presented in Figure 3, in which the initial post expresses a negative sentiment (labeled a) and is followed by two replies that disagree with it (labeled b and c). To compute the context-aware classification of a child node, its initial polarity is multiplied by the context-aware classification of its parent.

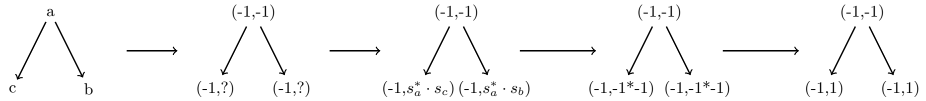


Fig. 3: **Illustration of hierarchical sentiment computation with intermediate transformations. Left: abstract tree structure; right: sentiment propagation using contextual multipliers.**

In the example, given three nodes a, b, c with context-aware sentiment values $s_a^* = -1$, $s_b^* = 1$, and $s_c^* = 1$, the total sentiment is: $S_{\text{tree}} = s_a^* + s_b^* + s_c^* = -1 + 1 + 1 = 1$

If the root post is neutral (0), it is treated as 1 when propagating sentiment to children—especially in open-ended questions like “What are your thoughts on ChatGPT for therapy?” where replies reflect stance. In Figure 3, despite a negative root post, the overall sentiment is positive due to disagreement in replies. This illustrates how the proposed context-aware model captures sentiment reversals shaped by hierarchical context within conversation trees.

Aggregated Sentiment Propagation While *HierSent* captures top-down sentiment influence, ***AggSent*** models bottom-up propagation. Each node is represented by a tuple $(st_{\text{node}}, st_{\text{node},p})$, where st_{node} denotes the aggregated stance of the subtree rooted at that node, and $st_{\text{node},p}$ indicates the stance of the node with respect to its parent p. Let $s_{\text{root}} \in \{-1, 0, +1\}$ denote the sentiment of the entire tree's root. We define the aggregated stance of a node recursively as:

$$st_{\text{node}} = \begin{cases} +1 & \text{if the node is a leaf (base case)} \\ \sum\limits_{c \in C} st_c \cdot st_{c,p} + 1 & \text{otherwise (recursive case)} \end{cases}$$

Here, C denotes the set of children of a node, st_c is the aggregated stance of each child $c \in C$, and $st_{c,p}$ is the stance of child c with respect to its parent p.

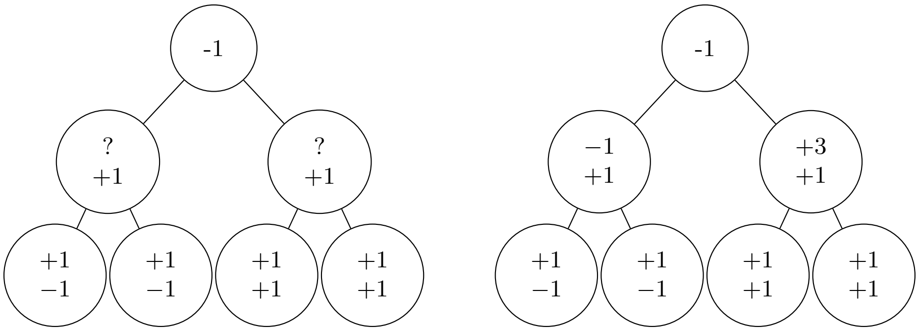
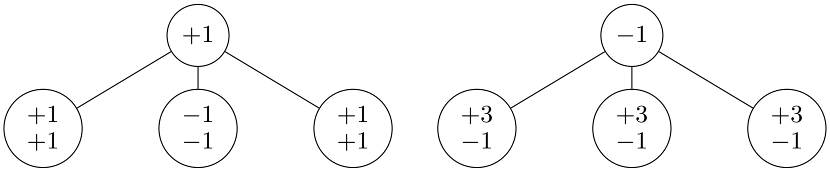


Fig. 4: **Recursive stance aggregation before and after intermediate node resolution. Each node shows st_{node} (top) and $st_{\text{node},p}$ (bottom). The root has sentiment -1, and all nodes have a stance toward their parent. Leaf nodes have $st_{\text{node}} = +1$. Intermediate nodes initially have unresolved stances (“?”) and stance $+1$ toward the root. After aggregating their children's stance-sentiment tuples, they resolve to -1 and $+3$.**

Finally, after recursive aggregation, the tree reduces to a 1-depth structure rooted at node r, and the overall sentiment score of the tree T is computed as:

$$\text{Sen}_T = \sum_{c \in C} st_c \cdot st_{c,p} \cdot s_{\text{root}} + s_{\text{root}}$$



(a) Root (+1) with mixed children

(b) Root (−1) with three (+3, −1) children

Fig. 5: **Sentiment propagation under *AggSent*. (a) A root with sentiment +1 receives mixed input from its children, resulting in a reinforced sentiment of +2. (b) A root with sentiment −1 is opposed by strongly aligned children, producing an aggregated sentiment of +5 and reversing the root's polarity. This highlights how lower-level agreement can override the root's original stance.**

6 Experimental Results

6.1 Relevance Classification using Machine Learning

As described in Section 3, we retrieved 7,049 X posts. To estimate therapeutic relevance, 180 posts were manually annotated, revealing that 46% were irrelevant. The annotated subset served as both a ground-truth reference and an evaluation set for our proposed relevance classification methods.

Evaluation of SemReC The **SemReC** model (Section 4) was evaluated using both an 80/20 train-test split and leave-one-out cross-validation (LOOCV) to balance efficiency with robustness (see Tables 1a and 1b). Due to the small dataset (180 samples), LOOCV helped ensure stable performance estimates. We experimented with $k = 3$ and $k = 5$, as these values offer a balance between local sensitivity and robustness to label noise in sparse, high-dimensional data. Notably, using PCA led to significantly higher accuracy and precision compared to configurations without PCA, highlighting its importance for improving semantic separation in the embedding space.

Table 1: Comparison of **SemReC** performance under two validation strategies:

(a) Performance under 80/20 Split

K	PCA	Accuracy	Precision	Recall	F1
3	No	0.7297	0.6667	1.0000	0.8000
3	Yes	0.8919	0.8333	1.0000	0.9091
5	No	0.6757	0.6250	1.0000	0.7692
5	Yes	0.8649	0.8649	1.0000	0.8889

(b) Performance under LOOCV

K	PCA	Accuracy	Precision	Recall	F1
3	No	0.7569	0.6957	0.9796	0.8136
3	Yes	0.8840	0.8738	0.9184	0.8955
5	No	0.7127	0.6575	0.9796	0.7869
5	Yes	0.8674	0.8364	0.9388	0.8846

To contextualize the performance of the k-Nearest Neighbor classifier, we compared it against several baseline models, including Support Vector Machine

(SVM), Logistic Regression, and XGBoost. All models were trained and evaluated on the same settings to ensure consistency and comparability.

Table 2: Comparison of Metrics Across Models

Model	Acc	Prec	Rec	F1
SVM	0.84	0.89	0.80	0.84
LogReg	0.81	0.78	0.90	0.84
XGB	0.78	0.80	0.80	0.80
KNN	**0.88**	**0.87**	**0.92**	**0.90**

Table 3: Performance of Pre-trained Sentiment Models

Model	PL Acc	HL Acc
BERT	0.53	0.56
RoBERTa	0.77	0.86
VADER	0.73	0.84
MS AI Builder	0.78	0.81

As shown in Table 2, the k-Nearest Neighbor model outperformed all other classifiers across all evaluation metrics, demonstrating superior overall performance and robustness in relevance classification. Based on these results, the $k = 3$ with PCA configuration was selected for final evaluation. When applied to all remaining unlabeled collected posts, **SemReC** classified 54% of posts as therapeutically relevant. This provided a filtered dataset for downstream sentiment analysis.

Evaluation of PASS PASS (Section 4) generated binary relevance labels (`Is_Relevant`) for each post. A post was labeled relevant if its average similarity exceeded a tunable threshold. We selected the value of thresholds by using 5-fold cross validation and the best-performing configuration achieved an accuracy of 80% on the test dataset.

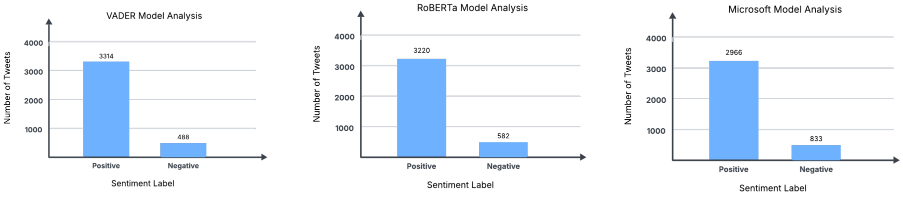
6.2 Evaluation of Sentiment Analysis Models

Models described in Section 5 were assessed for their ability to classify posts as positive, negative, or neutral. To evaluate performance, we used two ground truth labeling strategies: human-annotated labels (**HL**) and large language model generated pseudo-labels (**PL**). A subset of posts were manually annotated to ensure high-fidelity sentiment labels (Section 4), while an LLM was used to generate PLs, leveraging its demonstrated effectiveness in approximating human judgment for subjective tasks [32]. Evaluating against both HL and PL enabled precise, small-scale assessment and scalable, broader validation.

Pre-trained sentiment analysis models produced varied output formats, requiring standardization for consistent evaluation. VADER's compound scores were discretized, and Microsoft's categorical labels were relabeled—grouping neutral and mixed with positive—to reflect the assumption that non-negative sentiment implies a constructive ChatGPT experience. This enabled uniform binary classification across models. Table 3 summarizes the performance of four sentiment analysis models on both pseudo-labeled and human-labeled datasets. Off-the-shelf models like VADER, RoBERTa, and Microsoft's pre-trained model demonstrated strong accuracy on both pseudo-labeled and human-labeled data,

Table 4: Predicted Labels of Posts Using the Microsoft Model

Post	**Sentiment / Confidence Score**
I hate ChatGPT. I simply asked ”Could you say it with more empathy” after asking a mental health question. I was reacting to somebody saying it could be of support to peeps dealing with MH issues. Please, if you need real help, find a human.	Negative, 99%
ChatGPT goes to therapy and finds its ’voice,’ while the rest of us are still looking for the mute button on Zoom. Jokes aside, the rise of AI in emotional support is both fascinating and fraught with challenges.	Neutral, 88%
ChatGPT can be a versatile tool for different aspects of life, from career development to emotional support. Great thread!	Positive, 100%



(a) VADER (b) RoBERTa (c) Microsoft

Fig. 6: Sentiment label distributions predicted by VADER, RoBERTa, and Microsoft sentiment models.

with RoBERTa performing best overall. BERT lagged behind, indicating limited effectiveness in capturing sentiment in this context. Overall, the findings highlight the strength of pre-trained models in capturing sentiment in therapeutic contexts.

Analysis of Predicted Labels Given the comparable performance of the evaluated sentiment models (Table 3), we deployed RoBERTa, VADER and Microsoft models to label the relevance-filtered posts identified in Section 6.1. Figure 6 presents the predicted sentiment counts for these models. Despite architectural differences, these models exhibit similar sentiment distributions, further supporting the consistency of sentiment trends across the dataset.

Table 4 shows predictions from the Microsoft sentiment model on posts about ChatGPT in therapeutic contexts. The model performs well on clearly supportive posts but often labels nuanced or mixed-emotion content as neutral or misclassifies appreciative yet critical tones as negative.

6.3 Enhancing Sentiment Analysis with Conversation Trees

Real-world conversation trees were scraped from Reddit, as described in Section 3, and used for sentiment analysis. For evaluation purposes, a separate set of 500 synthetic conversation trees was generated and labeled by a large language model (LLM). We conducted two evaluations to assess the effectiveness of the proposed models. First, we evaluated model effectiveness by comparing predicted sentiment labels with LLM-generated pseudo-labels. Second, we examined sentiment "flips," where the predicted sentiment of the entire conversation tree diverged from that of the root node. Such flips—e.g., a negative root followed by a predominantly positive thread—demonstrate how sentiment evolves across interactions. This analysis highlights the ability of our proposed models to capture context-dependent emotional progression, underscoring the value of tree-structured sentiment analysis over flat, post-level classification.

HierSent Model Evaluation We evaluated the performance of the HierSent model (Section 4) using methods in [33] for the stance classification and VADER (Section 5) for sentiment analysis. As shown in Table 5, the model achieved high agreement for neutral and positive classes, with minor confusion between negative and neutral labels. The overall accuracy is computed as $\text{Accuracy} = \frac{\text{Correct predictions}}{\text{Total predictions}} = \frac{105+166+168}{500} = 0.878$, indicating the proportion of correctly classified instances across all sentiment categories. This high accuracy showcases the effectiveness of the proposed HierSent model.

Table 5: Confusion Matrix for *HierSent*

True / Pred	**-1**	**0**	**+1**
Negative (-1)	105	61	0
Neutral (0)	0	166	0
Positive (+1)	0	0	168

AggSent Model Evaluation We evaluated the AggSent model (4) using the same set of LLM-generated and labeled conversation trees. As with HierSent, performance was assessed based on the accuracy of predicted sentiment labels relative to the LLM-provided ground truth. AggSent achieved similar accuracy of 87%, further confirming the effectiveness and reliability of tree-structured sentiment propagation in capturing conversation-level sentiment.

Table 6: Predicted Labels and Sentiment Flips on Web-Scraped Reddit Data

(a) Predicted Sentiment Distribution

Model	**Neg(-1)**	**Neutral(0)**	**Pos(+1)**
HierSent	19	16	147
AggSent	19	15	148

(b) Sentiment Flips on Reddit Trees

Initial / Final	**-1**	**0**	**+1**
Negative (-1)	13	2	0
Neutral (0)	1	9	33
Positive (+1)	5	5	114

To summarize the sentiment of the collected real-world data, we applied both HierSent and AggSent to Reddit conversation trees (Section 3). Table 6a presents the predicted sentiment distributions for each model. Both approaches produced similar outputs, with a strong skew toward positive sentiment across threads. This suggests that discussions in the sampled Reddit communities often reflect favorable sentiment toward ChatGPT in therapeutic contexts. To evaluate how sentiment evolves within conversations, we examined sentiment "flips". Table 6b quantifies these shifts, comparing the initial sentiment at the root with the aggregated sentiment of the full thread. While many conversations preserve the tone set by the root, a substantial number deviate. These dynamics highlight the limitations of single-turn sentiment models and emphasize the importance of tree-aware approaches in capturing sentiment transitions across multi-turn interactions.

6.4 Summary of Results

The proposed relevance classification approaches—**SemReC** and PASS, demonstrated high effectiveness, achieving accuracy of 90% and 80% when evaluated against manually annotated ground truth. These results validate the utility of both supervised and unsupervised methods for identifying therapeutically relevant content at scale.

In parallel, pre-trained sentiment models such as RoBERTa, VADER and Microsoft models showed strong performance, with accuracies exceeding 80%. These findings confirm that off-the-shelf models, despite being trained on general-domain corpora, can reliably capture sentiment in mental health–related tasks. Additionally, the proposed HierSent and AggSent models achieved 87% accuracy, demonstrating its robustness in capturing sentiment dynamics across threaded multi-turn discussions.

When applied to the collected unlabeled data, all models consistently predicted a predominance of positive sentiment, both at the level of individual posts and across entire conversation trees. These findings indicate that user sentiment toward ChatGPT's therapeutic potential is largely favorable. Moreover, they demonstrate the effectiveness of our proposed methods for performing scalable, context-aware sentiment analysis.

7 Conclusion

This study investigated ChatGPT's unintended therapeutic role using social media data, focusing on how users express positive or negative sentiment toward using ChatGPT therapeutically for mental health or emotional support. We proposed two relevance classification methods—**SemReC** (supervised) and PASS (unsupervised)—to identify posts referencing ChatGPT in therapeutic contexts. For sentiment analysis, we benchmarked pre-trained models. To capture sentiment dynamics in threaded conversations, we developed two tree-structured models: HierSent (top-down) and AggSent (bottom-up). These approaches revealed nuanced emotional patterns often missed by flat classifiers.

Our empirical findings show that sentiment toward ChatGPT's therapeutic role is predominantly positive, underscoring the importance of examining its social and psychological impacts. Future work may expand this analysis across platforms, incorporate real-time interactions, and refine models to gain a deeper understanding of user perceptions of AI-driven emotional and mental health support.

References

1. Jason Wei et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.
2. Sebastien Bubeck et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.
3. Minh Vu et al. Chatgpt and the future of ai in healthcare: A scoping review. *Journal of Medical Internet Research*, 2023.
4. Health Staff. People are using chatgpt as therapy—is it safe?, 2023.
5. Vice Staff. We spoke to people who started using chatgpt as their therapist, 2023.
6. Benedict Carey. Human therapists prepare for battle against a.i. pretenders. *The New York Times*, 2025. Published February 24, 2025.
7. Ismail Dergaa, Feten Fekih-Romdhane, Souheil Hallit, Alexandre Andrade Loch, Jordan M Glenn, Mohamed Saifeddin Fessi, Mohamed Ben Aissa, Nizar Souissi, Noomen Guelmami, Sarya Swed, et al. Chatgpt is not ready yet for use in providing mental health assessment and interventions. *Frontiers in Psychiatry*, 14:1277756, 2024.
8. Charlotte Blease and John Torous. Chatgpt and mental healthcare: balancing benefits with risks of harms. *BMJ Ment Health*, 26(1), 2023.
9. Rawan AlMakinah, Andrea Norcini-Pala, Lindsey Disney, and M. Abdullah Canbaz. Enhancing mental health support through human-ai collaboration: Toward secure and empathetic ai-enabled chatbots. *arXiv preprint arXiv:2410.02783*, 2024.
10. Sergio Triscari, Sebastiano Battiato, Luca Guarnera, and Pasquale Caponnetto. Ai chatbots for mental health: A scoping review of effectiveness, feasibility, and applications. *Applied Sciences*, 14(13):5889, 2024.
11. Roshini Salil, Binny Jose, Jaya Cherian, and Nisha Vikraman. Digitalized therapy and the unresolved gap between artificial and human empathy. *Frontiers in Psychiatry*, 15:1522915, 2025.
12. Ruosi Shao. An empathetic ai for mental health intervention: Conceptualizing and examining artificial empathy. In *Proceedings of the 2nd Empathy-Centric Design Workshop*, pages 1–6, 2023.
13. Matan Rubin, Hadar Arnon, Jonathan D Huppert, Anat Perry, et al. Considering the role of human empathy in ai-driven therapy. *JMIR Mental Health*, 11(1):e56529, 2024.
14. Md Fazle Rabbi. Chatgpt and emotional support: a reflection on modern needs. *The Daily Observer*, 2025. Available online: observerbd.com.
15. Romal Thoppilan et al. Lamda: Language models for dialog applications. In *Proceedings of NeurIPS*, 2022.

16. Nazish Imran, Aateqa Hashmi, and Ahad Imran. Chat-gpt: opportunities and challenges in child mental healthcare. *Pakistan Journal of Medical Sciences*, 39(4):1191, 2023.
17. Vera Sorin, Dana Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. Large language models and empathy: Systematic review. *Journal of Medical Internet Research*, 26:e52597, 2024.
18. Hao Zhou et al. Emotional chatting machine: Emotional conversation generation with internal and external memory. *AAAI*, 2018.
19. D. Ganguli et al. Opportunities and risks of llms for scalable deliberation with polis. *arXiv preprint arXiv:2306.11932*, 2022.
20. Yuan Zhang and Yue Zhang. Tree communication models for sentiment analysis. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3518–3527, 2019.
21. Erfaneh Gharavi and Hadi Veisi. Using rst-based deep neural networks to improve text representation. *Signal and Data Processing*, 20(1):181–197, 2023.
22. Jin Wang, Liang-Chih Yu, K Robert Lai, and Xuejie Zhang. Investigating dynamic routing in tree-structured lstm for sentiment analysis. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3432–3437, 2019.
23. Sebastian Ruder, Parsa Ghaffari, and John G Breslin. A hierarchical model of reviews for aspect-based sentiment analysis. *arXiv preprint arXiv:1609.02745*, 2016.
24. Chengjie Zhou, Bobo Li, Hao Fei, Fei Li, Chong Teng, and Donghong Ji. Revisiting structured sentiment analysis as latent dependency graph parsing. *arXiv preprint arXiv:2407.04801*, 2024.
25. X Corp. X API: Introduction, n.d.
26. Reddit Inc. Reddit api documentation, n.d.
27. OpenAI. text-embedding-3-small - openai platform documentation, 2024.
28. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.
29. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
30. Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225, 2014.
31. Microsoft. AI Builder Sentiment Analysis. https://learn.microsoft.com/en-us/ai-builder/prebuilt-sentiment-analysis, 2023. https://learn.microsoft.com/en-us/ai-builder/prebuilt-sentiment-analysis.
32. Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023.
33. Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41, 2016.