Detecting Early and Implicit Suicidal Ideation via Longitudinal and Information Environment Signals on Social Media

Anonymous ACL submission

Abstract

On social media, many individuals experiencing suicidal ideation (SI) do not disclose their distress explicitly. Instead, signs may surface indirectly through everyday posts or peer interactions. Detecting such implicit signals early is critical but remains challenging. We frame early and implicit SI as a forward-looking prediction task and develop a computational framework that models a user's information environment, consisting of both their longitudinal posting histories as well as the discourse of their socially proximal peers. We adopted a composite network centrality measure to identify top neighbors of a user, and temporally aligned the user's and neighbors' interactions—integrating the multi-layered signals in a fine-tuned DeBERTa-v3 model. In a Reddit study of 1,000 (500 *Case* and 500 *Control*) users, our approach improves early and implicit SI detection by 15% over individual-only baselines. These findings highlight that peer interactions offer valuable predictive signals and carry broader implications for designing early detection systems that capture indirect as well as masked expressions of risk in online environments.

1 Introduction

More than 703,000 individuals die by suicide each year worldwide, making it a major global public health concern (WHO, 2025). Suicidal ideation (SI) refers to a spectrum of cognitive and behavioral manifestations related to suicide, ranging from passive thoughts of death to active planning and engagement in self-injurious actions with intent to die (Joiner, 2005). While early detection of these signs is critical, psychiatry and mental health services have long struggled to detect risk before individuals disclose it explicitly (Calear and Batterham, 2019; Hom et al., 2017). A recent meta-analysis revealed that the estimated prevalence of SI disclosure is only 46.3%, inferring that the majority of people who experience suicidal thoughts do not disclose them, making early detection critical for timely intervention (Hallford et al., 2023).

Online platforms have become vital spaces where signs of SI may surface. People use online mental health communities to share distress, seek help, and connect with peers (De Choudhury and De, 2014). This has created opportunities for both researchers and clinicians to understand suicide risk through language, social interactions, and online behavioral cues. Prior work has shown the value of linguistic cues for identifying mental health risks (Coppersmith et al., 2014; De Choudhury et al., 2013; Guntuku et al., 2017). In particular, within the context of SI, prior work has computationally modeled the language of SI on social media (Burnap et al., 2015; Saha et al., 2019; Alghazzawi et al., 2025; Zhang et al., 2025; Naseem et al., 2025). More recent computational approaches leveraged longitudinal and multimodal data, as well as social network analyses, to anticipate suicide-related behaviors (Shen et al., 2020). Yet, critical limitations persist. Prior research has primarily examined posts where SI is explicitly conveyed, such as direct or indirect references to self-harm or suicidal thoughts within suicide-related forums or discussions (Ji, 2022a; Bloom et al., 2025). These approaches presuppose that individuals articulate their distress, yet, many at-risk individuals neither disclose SI nor exhibit overt warning signs, but instead mask distress within seemingly ordinary discourse (McGillivray et al., 2022; Mérelle et al., 2018; Podlogar et al., 2022). Our work targets the detection of these undisclosed SI, characterized by subtle, contextually obscured, or socially distributed indicators, to potentially enable earlier and more effective intervention.

To address the above gap, our work is guided by the research question (RQ): **How can early signals of SI be detected in social media activity, particularly in the absence of explicit disclosures?** To address the RQ, we conceptualize early and im-

plicit SI as a forward-looking prediction problem that requires modeling both an individual's longitudinal behavior and the broader social context in which that behavior unfolds. We develop a framework that jointly captures temporal dynamics in a user's posting history and the conversations of socially proximal peers, enabling a richer representation of early warning signals. This idea of using socially proximal peers to understand an individual's mental state is derived from many prior acclaimed psychology works (La Greca and Harrison, 2005; Copeland, 2021; Victor et al., 2019). Within this framework, we pursue three aims:

Aim 1: Examine whether users' longitudinal posting patterns, in conjunction with commentary on those posts, reveal early signals of implicit SI.

Aim 2: Examine how the surrounding information environment, including peer interactions and exposure to content, contributes as an additional predictive feature for implicit SI.

Aim 3: Identify the linguistic markers that underlie interaction types, in terms of how specific language cues are associated with these interactions.

We conducted our study on Reddit, focusing on a sample of 1,000 users divided into 500 *Case* and 500 *Control* users (who never participated in mental health conversations). Our predictive framework jointly modeled each user's full posting timeline along with the discourse of their most socially proximal neighbors, identified through network centrality, and fine-tuned a DeBERTa-v3 model (He et al., 2021) to embed both individual and peer interaction signals into a unified representation.

Our findings show that incorporating peer interactions within the information environment, alongside users' longitudinal posting histories, significantly enhances detection of early and implicit SI, improving performance by **15%** compared to models based only on longitudinal user data. Overall, this paper contributes: 1) a framework for detecting implicit SI; 2) a method for systematically integrating environment information through neighbor interactions; and 3) empirical evidence that this integration improves early detection.

2 Related Work

2.1 Suicidal Ideation and Social Context

Psychological theory views SI as the product of intertwined social and cognitive factors. The Interpersonal Theory of Suicide (Joiner, 2005) posits that suicidal desire arises from perceived burdensomeness and thwarted belongingness, while capability develops through repeated exposure to pain or fear. Early computational studies of depression and SI focused on overt signals such as self-disclosure or negative sentiment (Coppersmith et al., 2014; De Choudhury et al., 2013), but later work showed that many at-risk individuals express distress through subtle cues, motivating methods that model temporal and semantic dynamics of behavior (Guntuku et al., 2017; Benton et al., 2017).

A range of methods have been proposed to detect mental health risk signals beyond surface cues. For instance, (Fatima et al., 2021) introduced *DASentimental*, a semi-supervised model combining bag-of-words and semantic networks, while (Trotzek et al., 2018) showed that convolutional networks with linguistic metadata enable earlier detection. More recent work leverages large language models, with GPT-3.5/4 using chain-of-thought prompting on diaries (Shin et al., 2024) and reasoning-guided LLMs improving interpretability (Teng et al., 2025). Temporal dynamics remain critical, from emotional "phases" in user timelines (Sawhney et al., 2021a) to transformer-based models enriched with temporal signals (Sawhney et al., 2020). Social context is equally important: hyperbolic embeddings of user histories and peer interactions enhance prediction (Sawhney et al., 2021b), peer networks and conversational responses influence trajectories (Wyman et al., 2019; De Choudhury and Kiciman, 2017), and longitudinal patterns reveal precursors of SI (De Choudhury et al., 2016). Complementary directions include clinical-domain datasets such as ScAN (Rawat et al., 2022), automated counseling support with PsyGUARD (Qiu et al., 2024), and calls to model underlying intent rather than surface disclosure (Ji, 2022b).

Together, this work affirms that SI arises from psychological distress, temporal dynamics, and social context, demanding models that go beyond surface cues. Yet most approaches still rely on explicit disclosures or static timelines, overlooking how evolving language interacts with peer responses. Our framework addresses this by jointly modeling users' longitudinal histories and post-level commentary, enabling early detection of implicit SI.

2.2 Social Media and Mental Health

Online platforms have become key venues for mental-health self-disclosure (De Choudhury and De, 2014), with communities like Reddit fostering

targeted support and a sense of belonging (Saha and Sharma, 2020; De Choudhury, 2015; Shimgekar et al., 2025). Moderated peer-support spaces reduce isolation and help people discuss stigmatized experiences (Johnson et al., 2022). Social support, especially emotional and informational, has been shown to improve well-being both offline and online (Cutrona and Troutman, 1986; De Choudhury and Kiciman, 2017; Saha and Sharma, 2020). Language plays a central role: psycholinguistic research links specific linguistic markers to mental-health outcomes (Pennebaker and Chung, 2007; Chung and Pennebaker, 2007), and computational studies have used these cues to detect distress and model support dynamics (Chancellor et al., 2016; Guntuku et al., 2017; Chancellor and De Choudhury, 2020). Prior work has also established the construct validity of these measurements (Saha et al., 2022).

Recent NLP work has focused on interpretable and fine-grained modeling of mental health on social media. Symptom-based approaches such as PSYSYM (Zhang et al., 2022; Chen et al., 2023) link online language to clinically meaningful categories of disorders. Depression severity has been quantified through semantic similarity to symptom descriptors (Pérez et al., 2022), while large language models now enable explainable detection with interpretable rationales (Wang et al., 2024). Analyses of pre and post diagnosis language shifts, highlight the temporal dynamics of distress expression (Alhamed et al., 2024). Supportive language marked by adaptability, immediacy, and emotionality predicts better outcomes (Althoff et al., 2016; Saha and Sharma, 2020), and automatic empathy-detection models scale such insights to peer-support settings (Sharma et al., 2020). For SI, machine-learning approaches have identified risk signals in social media language alongside emotional patterns that precede suicide attempts (Coppersmith et al., 2016; De Choudhury et al., 2016).

While online data shows promise for early risk detection, most methods isolate either individual language or specific interactions. Our approach instead models full posting timelines alongside peer influences, capturing risk even without explicit SI disclosures. Unlike prior work that assumes some suicide-related expression in specialized forums, our framework identifies early indicators without requiring references to self-harm or participation in such spaces.

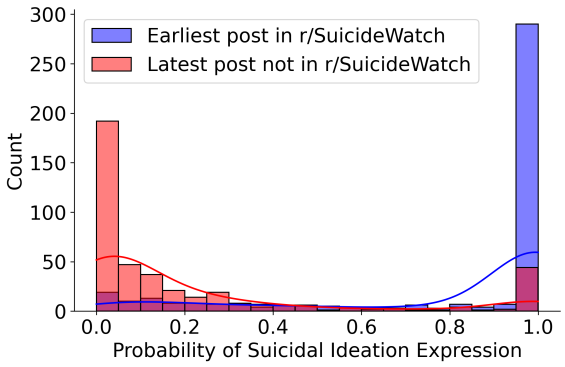


Figure 1: Distribution of SI probability for the same set of users before their last non suicidal post and their first post in `r/SuicideWatch`

3 Data

We used data from *r/SuicideWatch*, a semi-anonymous Reddit community focused on SI, alongside posts and comments from other subreddits. From the PushShift archive (Baumgartner et al., 2020) of April 2019 (18.3M posts, 138.5M comments), the data includes 10,037 posts and 38,130 comments from *r/SuicideWatch*. Prior work has leveraged Reddit for SI (De Choudhury et al., 2016; De Choudhury and Kiciman, 2017; Shimgekar et al., 2025) and broader mental health studies (Sharma and De Choudhury, 2018; Saha et al., 2020; De Choudhury and De, 2014).

Constructing Case and Control datasets. We identified two user cohorts. The first cohort comprises 500 *Case* individuals, defined as individuals who have made at least one post on *r/SuicideWatch*. The second cohort consists of 500 *Control* individuals, who never participated in any subreddit related to mental health. We identified the *Control* users by referencing the taxonomy of mental health-related subreddits from prior work (Sharma and De Choudhury, 2018).

For the *Case* group, all posts and comments before each user's first *r/SuicideWatch* disclosure were labeled positive (1). On average, users made **10.07 posts** before disclosure, transitioning within **1.5 days** of their last non-*r/SuicideWatch* post. To construct a balanced *Control* group, we sampled the first **10 posts** from each user (matching *Case* averages), labeling them negative (0). The dataset was then split 75:25 by users into train and test sets, ensuring no overlap.

A zero-shot NLI model (Laurer et al., 2023) revealed significantly higher SI probabilities in users' first *r/SuicideWatch* posts, indicating that entry marks a key turning point (Figure 1). The model,

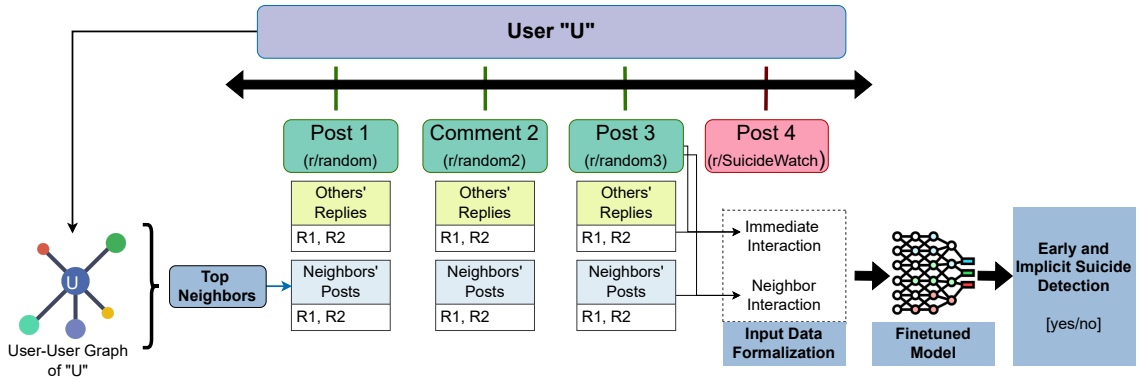


Figure 2: Illustration of user interactions (Immediate/Neighbor): Immediate interactions include users' self-posts, self-comments, and received replies, while neighbor interactions represent top neighbor posts and self-posts. Top neighbors identified via `NeighborScore`

based on DeBERTa-v3-base, was trained on 1.3M hypothesis–premise pairs from eight NLI datasets (e.g., MultiNLI, FEVER-NLI, LingNLI, DocNLI) to capture long-range reasoning. Using zero-shot labels (SI, non-SI), it assigned SI probabilities to posts, showing that users' first *r/SuicideWatch* posts exhibit stronger suicidal self-disclosure than their prior posts. While these probabilities may partly reflect the act of joining *r/SuicideWatch*, the linguistic patterns align more with distress and self-harm intent, supporting their use as an operational proxy for heightened suicidal self-disclosure.

4 Methods

To address our RQ on detecting early signals of SI in social media activity, particularly when explicit disclosures are absent, we conceptualized implicit SI as a forward-looking prediction problem of the likelihood that a user would eventually make a SI disclosure, proxied by their first post on *r/SuicideWatch*. Our framework models a user's longitudinal activity and social context via two interaction types: (1) **immediate interactions**—self-posts, self-comments, and received comments, and (2) **neighbor interactions**—self-posts and top neighbors' posts. This captures both active engagement and passive exposure, reflecting how users interact with and are influenced by content they relate to (De Choudhury and Kiciman, 2017).

Our methodological framework consists of five components—1) Timeline construction, 2) Neighboring user detection, 3) Input data formalization, and 4) implicit SI signal classification with modeling, 5) linguistic analysis of interaction types, which we describe below:

4.1 Timeline Construction

For each user U, we analyzed their full post and comment history for SI signals, ordered chronologically from earliest to latest.

4.2 Neighboring user detection

For a user U, we found its top neighbors by the following four steps:

Step 1: Initial User Collection We first collected all users who interacted with the user U, defined as either U commenting on their posts/comments or them commenting on U's posts/comments. This neighbor-identification procedure was then applied recursively to a maximum depth d=3, ensuring that both direct and indirect neighbors of U were captured.

Step 2: User-User Graph Based on all the initial collected users, we constructed a user-user graph where each **Node** represents an individual user in the network, and an undirected **Edge** connects two users if either of the users has commented on the other's post. The weight of the edge was quantified as the total number of comment-based interactions, with higher weights indicating stronger ties.

Step 3: Top Neighbor Detection From the user-user graph, we identified the top-n (n=10) neighbors for U. We ranked the neighbors using a `NeighborScore`, a combined centrality score $S(n)$, defined as:

$$S(n) = C_{\text{in-degree}}(n) + C_{\text{out-degree}}(n) + C_{\text{closeness}}(n) + C_{\text{eigenvector}}(n) + C_{\text{betweenness}}(n) + C_{\text{PageRank}}(n)$$

$C_{\text{in-degree}}$ and $C_{\text{out-degree}}$ capture normalized connectivity, $C_{\text{betweenness}}$ measures shortest-path centrality, $C_{\text{closeness}}$ denotes proximity, $C_{\text{eigenvector}}$

Table 1: Model performance metrics on different data combinations for Models M1–M4 (epochs=20).

	Data Used	Acc.	F1	Prec.	Rec.
M_1	Self-posts	0.86	0.85	0.88	0.84
M_2	Self-posts + Self-comments	0.81	0.83	0.78	0.91
M_3	Self-posts + Self-comments + Others' comments	0.80	0.79	0.79	0.80
M_4	**Self-posts + Top neighbor posts**	**0.95**	**0.96**	**0.93**	**0.99**

Table 2: Model performance on showing the importance of choosing the best neighbors (epochs=20).

	Data Used	Acc.	F1	Prec.	Rec.
M_4	**Self-posts + top neighbor posts**	**0.95**	**0.96**	**0.93**	**0.99**
M_5	Self-posts + worst neighbor posts	0.75	0.76	0.78	0.75
M_6	Self-posts + non-neighbor posts	0.68	0.71	0.67	0.78

reflects influence via important neighbors, and C_{PageRank} estimates probabilistic importance. The aggregated `NeighborScore` identifies peers with strong direct ties and broader network influence around U. As some centralities are correlated, $S(n)$ approximates overall neighbor prominence rather than precise influence. Performance variation by neighbor depth and count is shown in Appendix Table A9.

4.3 Input Data Formalization

We structured the input to language models to capture U's timeline, others' replies, and peer interactions. The overall data formalization, distinguishing immediate and neighbor interactions, is shown in Figure 2.

Immediate Interaction We aggregated content from a user's timeline, along with others' replies, and then used all of these textual content for fine-tuning the language model. To train our models, we used different combinations of features—1) U's self-posts, 2) U's self-posts and self-comments, and 3) U's self-posts, self-comments, and others' replies to U's posts or comments.

Neighbor Interaction Neighbor Interaction captures signals from proximate peers. For this purpose, we temporally aligned the timelines of U and their top-n neighbors. Then, at each timestamp i, we selected ten posts by the top neighbors closest in time to U's post. We aggregated the neighbors' posts with U's posts and embedded them into a dense vector representation. This approach led to the fourth type of model, where we included features from both immediate interactions above, as well as the top neighbors' posts.

Table 3: Model performance excluding neighbors with *r/SuicideWatch* posts. Results remain comparable to using all neighbors, indicating that broader neighbor interactions capture key signals for implicit SI detection (20 epochs).

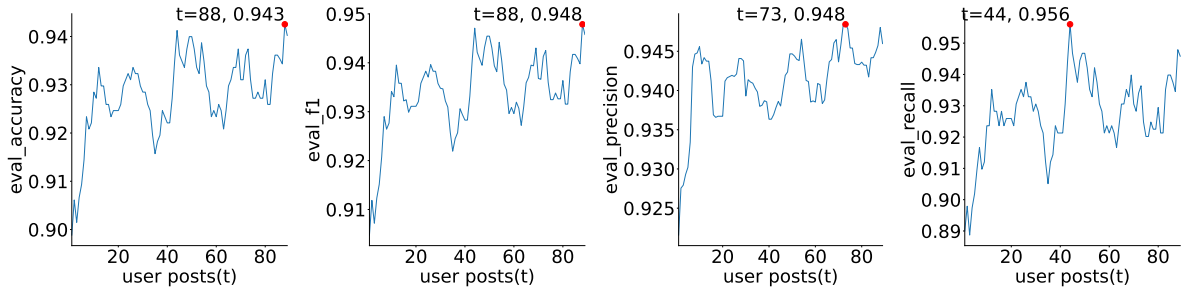
	Data Used	Acc.	F1	Prec.	Rec.
M_4	**Self-posts + top neighbor posts**	**0.95**	**0.96**	**0.93**	**0.99**
M_7	Self-posts + filtered neighbor posts	0.89	0.90	0.87	0.93

4.4 Implicit SI signal classification with modeling

For our study, we leveraged Microsoft's Deberta-v3-large model (He et al., 2021)—a 418M parameter transformer with 24 layers that processes sequences up to 512 tokens. We fine-tuned this model using `CLS` for global representation and `SEP` for segment boundaries on an Nvidia A100 GPU. We framed our problem of detecting implicit SI as a binary classification task—labeling each input x_i as $y_i \in \{0, 1\}$ for absence or presence of risk. We tokenized text into subword embeddings with attention masks, padded or truncated to 512 tokens, and encoded them to obtain a pooled `[CLS]` vector (`1024`-dim). A linear layer then mapped this vector to logits, followed by softmax for class probabilities. We optimized the model with cross-entropy loss using AdamW (learning rate 2×10^{-5}, weight decay 0.01) for 20 epochs with batch size 8, and evaluated after each epoch on a held-out validation set using accuracy, precision, recall, and F1-score. Results from fine-tuning alternative base models are provided in Appendix Table A2.

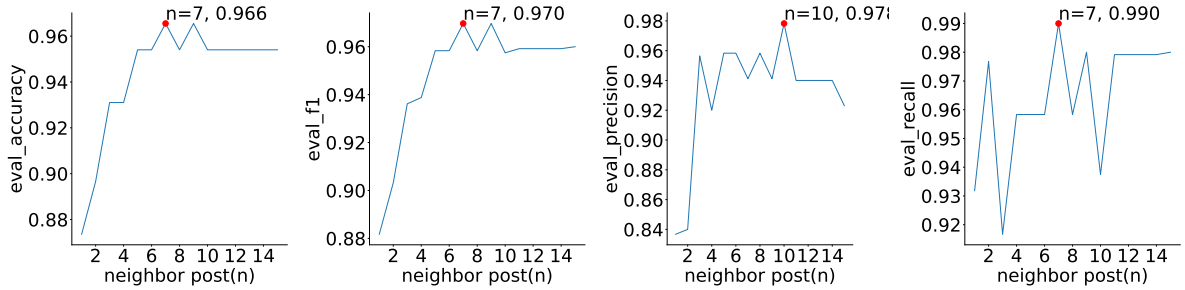
4.5 Linguistic Analysis of Interaction Types

We analyzed lexical, topical, and psycholinguistic patterns across users' self-posts, self-comments, replies, and top-neighbor posts. For lexical analysis, we used the Sparse Additive Generative Model (SAGE) (Eisenstein et al., 2011) to identify discriminative unigrams and bigrams distinguishing immediate and neighbor interactions. SAGE uses multinomial models with adaptive regularization to balance frequent and rare terms. For topical analysis, we employed BERTopic (Grootendorst, 2022) on *Case* and *Control* data, varying topic numbers (k=2–15) and achieving the highest coherence at k=12. After removing one outlier, eleven topics remained (Table 4) whose topic names were assigned by our clinical psychologist coauthor. Their normalized proportions captured topical variation across interaction types Table 5. For psycholinguistic analysis, we examined the occurrences of the af-



(a) Accuracy (b) F1 Score (c) Precision (d) Recall

Figure 3: Performance metrics on varying the number of input user posts (t). The plots show that the performance peaks at t=88.



(a) Accuracy (b) F1 Score (c) Precision (d) Recall

Figure 4: Performance metrics on varying the number of neighbor posts (n). The plots show that the performance peaks at around n=7.

fect category of keywords as per the well-validated Linguistic Inquiry and Word Count (LIWC) lexicon (Tausczik and Pennebaker, 2010).

5 Results

5.1 Model Performance (Immediate Interaction vs. Neighbor Interaction):

Table 1 summarizes the performance comparison of models with combinations of immediate interactions and neighbor interactions feature set. The baseline model using only a target user's posts performs strongly (Accuracy $= 0.86$, F1 $= 0.85$), showing that self-authored text alone captures salient linguistic markers of future SI, consistent with prior temporal analyses on Reddit and Facebook (Coppersmith et al., 2018; De Choudhury et al., 2016). Adding the user's self-comments boosts recall ($0.84 \rightarrow 0.91$) but lowers precision ($0.88 \rightarrow 0.78$), suggesting broader coverage but added noise in terms of precision. For example, a comment from a *Case* user reads, "Are they allowed to hit me? [..] I need to stay strong," including both themes of physical abuse as well as optimism. Additionally, including replies to user's posts/comments further reduces accuracy ($0.81 \rightarrow 0.80$) and F1 ($0.83 \rightarrow 0.79$). Such replies often convey empathy or advice that mask the user's true mindset (Gkotsis et al., 2017), as in one response: "I think talking to a professional would help, because he can understand you and give you advice." We note that the top-neighbor interaction approach achieves the highest performance (Accuracy=0.95, F1=0.96, Precision=0.93, Recall=0.99), with high recall reducing false negatives, which is critical for SI risk detection (Franklin et al., 2017). This suggests that features reflecting social exposure and information environment context provide valuable predictive signals for implicit SI.

5.2 Determining Optimal Post Counts for Robust Prediction

We evaluated how many posts from users (*Case* and *Control*) and their top neighbors optimize implicit SI detection. First, varying the number of neighbor posts (n) showed an improvement in accuracy from 0.87 to 0.97, with near-perfect recall (0.99) at $n = 7$, after which the gains plateaued (Figure 4). With $n = 7$ fixed, varying self-posts (t) revealed unstable performance at low t (e.g., recall 0.87 at $t = 1$), stabilizing for $t \geq 40$; the best results occurred at $t = 88$ (accuracy 0.94, F1 0.95, precision 0.95, recall 0.96) (Figure 3). These findings indicate that combining a user's full history with a few

Table 4: Clinically informed topics identified from posts and comments via topic modeling, with corresponding explanations and keywords.

Topic Theme	Explanation	Keywords
Self-Harm Tools	Mentions of knives, blades, or cutting instruments indicating self-injury ideation	knife, blade, cut, sharpen
Diet / Body Image	Discussions of calories, weight, dieting, or eating habits reflecting body concerns	calori, weight, diet, eat
Physical Pain / Health	Mentions of pain, medication, or suffering over time	pain, aspirin, suffer, longer
War / Military	Discussions related to military, soldiers, or historical conflicts	soldier, german, armi, soviet
Academic / Mental Stress	Mentions of depression, PhD, or mental strain	depress, phd, ggoddamn, eurozon
Cringe / Embarrassment	Mentions of awkwardness, social discomfort, or cringing situations	cri, cringi, crusad, crimin
Burn / Injury	Mentions of burns, cuts, or physical injury	burn, cut, hurt, degre
Self-Harm / Coping Strategies	Mentions of harming oneself or suicide coping mechanisms	harm, self, suicidebyword, cope
Violent Ideation	Mentions of killing or violent intent in extreme contexts	kill, killin, zaibatsu, meeee
Confusion / Uncertainty	Expressions of perplexity or inability to understand situations	confus, percent, stranger, 12ish
Cute / Affection	Positive, playful, or affectionate language	cute, soo, aww, omgggg

Table 5: Normalized topic proportions across different content sources. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Topic	User's Self-Posts	User's Self-Comments	Other's Comments	Top Neighbor's Posts	Kruskal-Wallis H-stat.
Self-Harm Tools	0.002	0.002	0.002	**0.006**	44.08***
Diet & Body Image	**0.005**	0.003	0.003	0.003	46.36***
Physical Pain & Health	0.003	0.002	0.002	**0.003**	51.55***
War & Military	**0.013**	0.003	0.001	0.002	7.45*
Academic Stress	0.001	0.001	0.002	**0.002**	43.81***
Cringe & Embarrassment	0.000	0.001	0.001	**0.002**	39.49***
Burn & Injury	0.000	0.000	0.000	**0.001**	1.84
Coping & Self-Harm	0.001	0.001	0.001	**0.002**	22.81***
Violent Ideation	0.001	0.001	0.000	**0.002**	26.13***
Confusion & Uncertainty	**0.001**	**0.001**	**0.001**	0.000	12.62**
Cute / Affection	0.000	0.001	0.001	**0.002**	22.93***

neighbor posts yields a robust SI predictor.

5.3 Role of Top Neighbor Posts in Capturing SI Signals

To understand the superior performance of top-neighbor-based modeling, we analyzed lexical, topical, and psycholinguistic patterns in users' self-posts, self-comments, replies, and posts of their top neighbors. Others' replies to a user's posts or comments provide a limited discriminative signal: common bigrams such as "feel like", "year old", and "every day" appear across both *Case* and *Control* data, reflecting supportive or neutral discourse (Zirikly et al., 2019). Topic modeling shows that clinically relevant themes (e.g., *Self-Harm Tools*, *Coping/Self-Harm*, *Violent Ideation*) are rare in others' replies, while users' own posts exhibit only moderate and inconsistent signals (Table 4, Table 5). In contrast, top neighbor posts consistently reflect mental-health-related themes that are highly predictive of SI (Kirtley et al., 2021). For *Case* users, neighbors frequently post about *Self-Harm Tools*, *Coping / Self-Harm*, *Violent Ideation*, *Physical Pain / Health*, and *Academic / Mental Stress*, with bigrams such as "mental health", "suicide prevention", and "emotional support" appearing frequently. Neighbors of *Control* users, by comparison, focus on neutral or unrelated topics (Table 4, Table 5). Table 6 shows that differences in affective categories (e.g., negative affect, anger, and sadness) are statistically significant, indicating that including neighbor posts enhances the distinction between *Case* and *Control* users' content. Therefore, top-neighbor posts provide richer social context than replies or isolated user history. Combining them with user history yields a more robust model for early detection of early and implicit SI.

5.4 Robustness Tests:

To ensure that our findings are not artifacts of specific modeling or sampling choices, we conducted additional robustness tests by varying the type of neighbors included—for instance, incorporating non-top neighbors into our models. For this purpose, we built additional models M_5 with the lowest-ranked neighbors and M_6 with non-neighbors—a random selection of other users—of a target user. Table 2 compares the performance—we note both M_4 and M_5 perform significantly poorly. Therefore, the performance boost by including the top neighbor posts is not by chance, and rather an important element in detecting implicit SI (Cero et al., 2024). Moreover, Table 3 demonstrates that the model retains strong predictive performance even after filtering out all top neighbors who have posted in *r/SuicideWatch*. This finding indicates that successful detection does not rely solely on neighbors' direct SI-related language, but rather on broader contextual cues captured from high-quality interactions. Appendix Table A4 shows linguistic analyses aligned with Table 2, Table A3 highlights "Affect" differences linked to Table 3, and addi-

Table 6: Comparison of LIWC Affect categories across M_1–M_4. Columns show normalized category counts for Case and Control groups and their differences. ***p<0.001, *p<0.01, p<0.05.

Category	M_1 (Immediate)				M_2 (Immediate)				M_3 (Immediate)				M_4 (Neighbor)			
	Case	Control	Diff.	H	Case	Control	Diff.	H	Case	Control	Diff.	H	Case	Control	Diff.	H
Pos. Affect	0.049	0.053	-0.004	38.80***	0.080	0.074	0.006	416.94***	0.060	0.057	0.003	45.86***	0.048	0.053	**-0.005**	23.57***
Neg. Affect	0.029	0.025	0.004	74.947***	0.035	0.030	0.005	651.17***	0.029	0.026	0.003	100.07***	0.029	0.021	**0.007**	169.87***
Anxiety	0.004	0.003	**0.001**	125.379***	0.003	0.002	**0.001**	358.217***	0.002	0.002	0.000	176.606***	0.003	0.002	0.000	62.663***
Anger	0.009	0.006	0.003	50.654***	0.012	0.010	0.002	322.855***	0.010	0.008	0.002	77.809***	0.011	0.006	**0.005**	130.857***
Sadness	0.007	0.007	0.000	71.131***	0.008	0.007	0.001	376.545***	0.006	0.005	0.001	136.379***	0.009	0.005	**0.004**	181.551***

tional model comparisons with LIWC results are in Appendix Table A1 (Tables A5–A8).

6 Discussion and Implications

A key takeaway of our study is that, early and implicit SI is best detected by modeling both an individual's longitudinal online activity and their surrounding information environment, capturing the inherently relational nature of SI expressions (Ammerman and Jacobucci, 2023). Neighbor-based modeling is particularly effective, as top neighbors, weighted via a *NeighborScore*, capture direct behavior and network-level influences, consistent with epidemiological evidence and classic effects like Werther and Papageno (Wyman et al., 2019; Phillips, 1974; Niederkrotenthaler et al., 2010; Yuan et al., 2023). Neighbor posts provide complementary signals beyond the user's self-content, showing that exposure to neighbors' expressions of distress or coping is associated with detection.

Neighbor-informed models frame suicidality within a social-ecological perspective (Bronfenbrenner, 1979), reflecting social contagion processes where suicidal behaviors and ideation can spread through networks, especially when peers disclose distress (Wyman et al., 2019; Gould et al., 2003). Online environments may amplify these effects, as individuals in high-risk networks can show subtle precursors before explicit disclosure. Network theories suggest that peers model behavior and influence interpretations of distress (Cero et al., 2024; Bearman and Moody, 2004; Christakis and Fowler, 2025), explaining why neighbor interactions reflect relational signals of implicit SI. A user's longitudinal posting history remains essential, tracing gradual shifts in tone, sentiment, and discourse (De Choudhury et al., 2016). However, interactional content must be selective: including all replies adds noise and weakens risk signals (Schmidt et al., 2024). As shown in Table 6, the difference between the LIWC "Affect" categories is consistently higher in M_4 where neighbors' posts are added. This shows that including posts from the neighbors add a clear distinction signal between the *Case* and *Control* data, making it better for the model to predict (Lu and Tu, 2024; Guo et al., 2012). These findings suggest avenues for theoretical and methodological extensions. Beyond the Interpersonal Theory of Suicide (Joiner, 2005), models such as the Integrated Motivational Volitional Model (O'Connor and Kirtley, 2018) and the Three-Step Theory (Klonsky and May, 2015) may help identify subtle motivational or volitional cues in language and interactions. The strong influence of social context further supports contagion and network theories, indicating that implicit suicidal ideation emerges from both individual cognition and broader interpersonal environments. Overall, this work enhances detection accuracy while framing suicide risk as a relational process in digital social environments. By showing that peer-network interactions substantially improve predictive power, it bridges computational modeling with network-informed prevention strategies.

7 Conclusion

Our study demonstrates that early and implicit SI is best detected when a user's longitudinal activity is interpreted within their social context. Subtle linguistic and interactional cues often emerge well before explicit disclosures, enabling identification of risk trajectories at an early stage. By incorporating posts from socially proximal peers, our framework operationalizes theories of suicide contagion and social influence, yielding substantial gains in predictive performance. Notably, while a user's own posts and comments provide moderate insight, even a small set of strategically selected neighbor posts carries a significantly strong predictive signal. These findings highlight that suicide risk is inherently relational: effective detection requires selectively integrating self-content and socially distributed cues. Beyond improving accuracy, this relational framing supports scalable, context-aware, and ethically responsible early-warning systems that align with both clinical theory and the dynamics of online communities.

8 Limitations

Our study has limitations that also show interesting future directions. Methodologically, our models exclusively rely on textual content, overlooking multimodal signals such as images, videos, emojis, GIFs, or external links that often convey emotional states, coping strategies, or distress. Incorporating these modalities could improve sensitivity to subtle risk signals, enhance interpretability, and provide a more holistic understanding of online behaviors associated with suicidal ideation. Similarly, our current approach treats peer interactions homogeneously, though peers can exert positive, neutral, or negative influences. Modeling these distinctions could capture the functional impact of social context on risk trajectories and guide ethically responsible interventions.

Finally, robustness and generalizability can be improved through replication across platforms (e.g., Twitter, TikTok, Discord), multilingual and cross-cultural settings, and naturalistic populations. Temporal weighting, trajectory-based risk modeling, and interpretability techniques such as attention visualization, SHAP, or counterfactuals could clarify key linguistic and network features while enhancing actionable insights. Future work should integrate multimodal inputs, refine social context modeling, and maintain rigorous ethical oversight to ensure predictive models support vulnerable individuals responsibly and effectively.

9 Ethical Considerations

Reflexivity This paper used publicly accessible social media discussions on Reddit and did not require direct interactions with individuals, thereby not requiring ethics board approval. However, we are committed to the ethics of the research and we followed practices to secure the privacy of individuals in our dataset. Our research team comprises researchers holding diverse gender, racial, and cultural backgrounds, including people of color and immigrants, and hold interdisciplinary research expertise. This team consists of computer scientists with expertise in social computing, NLP, and HCI, and psychologists with expertise in clinical psychology, adolescent depression and suicide, and digital health interventions. One psychologist coauthor specializes in suicide etiology, suicide prevention, and crisis intervention, and another psychologist coauthor is a clinical psychologist with over 16 years of experience spanning adult and adolescent inpatient care and crisis suicide helplines. To ensure validity and prevent misrepresentation, our findings were reviewed and corroborated by our psychologist coauthors. However, our work is not intended to replace the clinical evaluation of an individual undergoing suicidal thoughts and should not be taken out of context to conduct mental health assessments.

Risk of Misinterpretation and Harm with Automated Detection Our study leverages computational methods to identify early and implicit suicidal ideation (SI) in online communities. While these approaches provide opportunities for early intervention, they also present serious ethical challenges. Automated systems may misinterpret subtle linguistic signals or social cues, leading to false positives or false negatives. Misclassification can result in unwarranted labeling of individuals as at risk, potentially causing stigma, anxiety, or social consequences. Conversely, failing to detect genuine risk may deny individuals timely support. We emphasize that suicidal ideation manifests heterogeneously across individuals, and even carefully designed models cannot fully capture the nuanced context of personal distress.

Privacy, Consent, and Data Sensitivity Given the sensitive nature of suicidal thoughts, the privacy and confidentiality of users are paramount. Our approach relies on publicly available text, but the inclusion of social network information, even aggregated, raises concerns about inadvertent exposure of personal interactions. Ethical deployment requires anonymization, secure storage, and strict access controls. Users should be informed about potential data usage and retention, and explicit consent should be prioritized wherever feasible.

Social Context and Potential Misuse By modeling social context and peer interactions, our work operationalizes theories of social influence in suicidal ideation. However, this introduces additional ethical considerations. Insights derived from peer interactions could be misused if exploited for non-clinical purposes, such as targeted advertising or profiling of vulnerable individuals. Careful governance and ethical oversight are necessary to ensure that social context is leveraged solely for supportive, preventive interventions rather than for commercial or punitive applications.

Transparency, Interpretability, and Responsible Use Ethical deployment also requires trans-

parency and interpretability. Systems identifying SI must clearly communicate their scope, limitations, and the fact that outputs are probabilistic, not diagnostic. Stakeholders, including clinicians, platform moderators, and potentially affected users, should understand how predictions are generated, especially when interventions are triggered. Responsible use also involves integrating human-in-the-loop frameworks, where trained professionals augment computational predictions to avoid over-reliance on automated systems.

Future Directions for Ethical AI Future research should explore multimodal detection methods while carefully balancing privacy and ethical constraints. Integrating images, URLs, or videos may improve sensitivity, but must be approached with strict ethical safeguards. Additionally, systematically categorizing peer influence, positive, neutral, or negative, could enhance the interpretability and safety of predictions, ensuring interventions are informed by context rather than raw social activity. Overall, ethical design in computational mental health requires prioritizing user welfare, minimizing harm, and embedding accountability at every stage of model development and deployment.

10 AI Involvement Disclosure

Certain sections of the manuscript were refined using AI-assisted writing tools (e.g., ChatGPT, Grammarly). All analyses, scientific content, and experiments were written solely by the authors.

References

Daniyal Alghazzawi, Hayat Ullah, Naila Tabassum, Sahar K Badri, and Muhammad Zubair Asghar. 2025. Explainable ai-based suicidal and non-suicidal ideations detection from social media text with enhanced ensemble technique. *Scientific Reports*, 15(1):1111.

Falwah Alhamed, Julia Ive, and Lucia Specia. 2024. 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.

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Brooke A Ammerman and Ross Jacobucci. 2023. The impact of social connection on near-term suicidal ideation. *Psychiatry research*, 326:115338.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Peter S Bearman and James Moody. 2004. Suicide and friendships among american adolescents. *American journal of public health*, 94(1):89–95.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multi-task learning for mental health using social media text. *arXiv preprint arXiv:1712.03538*.

Paul Bloom, Isaac Treves, David Pagliaccio, Isabella Nadel, Emma Wool, Hayley Quinones, Julia Greenblatt, Natalia Parjane, Katherine Durham, Samantha Salem, and 1 others. 2025. Identifying suicide-related language in smartphone keyboard entries among high-risk adolescents.

Urie Bronfenbrenner. 1979. *The ecology of human development: Experiments by nature and design*. Harvard university press.

Pete Burnap, Walter Colombo, and Jonathan Scourfield. 2015. Machine classification and analysis of suicide-related communication on twitter. In *Proc. ACM conference on hypertext & social media*.

Alison L Calear and Philip J Batterham. 2019. Suicidal ideation disclosure: Patterns, correlates and outcome. *Psychiatry research*, 278:1–6.

Ian Cero, Munmun De Choudhury, and Peter A Wyman. 2024. Social network structure as a suicide prevention target. *Social psychiatry and psychiatric epidemiology*, 59(3):555–564.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):1–11.

Stevie Chancellor, Zhiyuan Lin, Erica L Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016. Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1171–1184. ACM.

Siyuan Chen, Zhiling Zhang, Mengyue Wu, and Kenny Zhu. 2023. Detection of multiple mental disorders from social media with two-stream psychiatric experts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9071–9084.

Nicholas A Christakis and James H Fowler. 2025. *Connected: The surprising power of our social networks and how they shape our lives*. Hachette+ ORM.

Cindy Chung and James W Pennebaker. 2007. The psychological functions of function words. *Social communication*, pages 343–359.

Molly Copeland. 2021. The long shadow of peers: adolescent networks and young adult mental health. *Social Sciences*, 10(6):231.

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.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the third workshop on computational linguistics and clinical psychology*, pages 106–117.

Carolyn E Cutrona and Beth R Troutman. 1986. Social support, infant temperament, and parenting self-efficacy: A mediational model of postpartum depression. *Child development*, pages 1507–1518.

Munmun De Choudhury. 2015. Social media for mental illness risk assessment, prevention and support. In *Proceedings of the 1st ACM Workshop on Social Media World Sensors*, pages 1–1. ACM.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 71–80.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 128–137.

Munmun De Choudhury and Emre Kiciman. 2017. The language of social support in social media and its effect on suicidal ideation risk. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 32–41.

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. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110.

Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1041–1048.

Asra Fatima, Ying Li, Thomas Trenholm Hills, and Massimo Stella. 2021. Dasentimental: Detecting depression, anxiety, and stress in texts via emotional recall, cognitive networks, and machine learning. *Big Data and Cognitive Computing*, 5(4):77.

Joseph C Franklin, Jessica D Ribeiro, Kathryn R Fox, Kate H Bentley, Evan M Kleiman, Xieyining Huang, Katherine M Musacchio, Adam C Jaroszewski, Bernard P Chang, and Matthew K Nock. 2017. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological bulletin*, 143(2):187.

George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim JP Hubbard, Richard JB Dobson, and Rina Dutta. 2017. Characterisation of mental health conditions in social media using informed deep learning. *Scientific reports*, 7(1):1–11.

Madelyn Gould, Patrick Jamieson, and Daniel Romer. 2003. Media contagion and suicide among the young. *American Behavioral Scientist*, 46(9):1269–1284.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Guibing Guo, Jie Zhang, and Daniel Thalmann. 2012. A simple but effective method to incorporate trusted neighbors in recommender systems. In *International conference on user modeling, adaptation, and personalization*, pages 114–125. Springer.

David John Hallford, Danielle Rusanov, B Winestone, R Kaplan, Matthew Fuller-Tyszkiewicz, and Glenn Melvin. 2023. Disclosure of suicidal ideation and behaviours: A systematic review and meta-analysis of prevalence. *Clinical psychology review*, 101:102272.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Melanie A Hom, Ian H Stanley, Matthew C Podlogar, and Thomas E Joiner Jr. 2017. "are you having thoughts of suicide?" examining experiences with disclosing and denying suicidal ideation. *Journal of Clinical Psychology*, 73(10):1382–1392.

Shaoxiong Ji. 2022a. Towards intention understanding in suicidal risk assessment with natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4057–4067. The Association for Computational Linguistics.

Shaoxiong Ji. 2022b. Towards intention understanding in suicidal risk assessment with natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4057–4067. The Association for Computational Linguistics.

Jazette Johnson, Vitica Arnold, Anne Marie Piper, and Gillian R Hayes. 2022. " it's a lonely disease": Cultivating online spaces for social support among people living with dementia and dementia caregivers. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–27.

Thomas E. Joiner. 2005. *Why People Die by Suicide*. Harvard University Press.

Olivia J Kirtley, Ian Hussey, and Lisa Marzano. 2021. Exposure to and experience of self-harm and self-harm related content: An exploratory network analysis. *Psychiatry Research*, 295:113572.

E David Klonsky and Alexis M May. 2015. The three-step theory (3st): A new theory of suicide rooted in the "ideation-to-action" framework. *International Journal of Cognitive Therapy*, 8(2):114–129.

Annette M La Greca and Hannah Moore Harrison. 2005. Adolescent peer relations, friendships, and romantic relationships: Do they predict social anxiety and depression? *Journal of clinical child and adolescent psychology*, 34(1):49–61.

Moritz Laurer, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. 2023. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, pages 1–33.

Fangcao Lu and Caixie Tu. 2024. The impact of comment slant and comment tone on digital health communication among polarized publics: A web-based survey experiment. *Journal of medical Internet research*, 26:e57967.

Lauren McGillivray, Demee Rheinberger, Jessica Wang, Alexander Burnett, and Michelle Torok. 2022. Non-disclosing youth: a cross sectional study to understand why young people do not disclose suicidal thoughts to their mental health professional. *BMC psychiatry*, 22(1):3.

Saskia Mérelle, Elise Foppen, Renske Gilissen, Jan Mokkenstorm, Resi Cluitmans, and Wouter Van Ballegooijen. 2018. Characteristics associated with non-disclosure of suicidal ideation in adults. *International journal of environmental research and public health*, 15(5):943.

Usman Naseem, Liang Hu, Qi Zhang, Shoujin Wang, and Shoaib Jameel. 2025. Digri: Distorted greedy approach for human-assisted online suicide ideation detection. In *Proceedings of the ACM on Web Conference 2025*, pages 5192–5201.

Thomas Niederkrotenthaler, Martin Voracek, Arno Herberth, Benedikt Till, Markus Strauss, Elmar Etzersdorfer, Brigitte Eisenwort, and Gernot Sonneck. 2010. Role of media reports in completed and prevented suicide: Werther v. papageno effects. *The British Journal of Psychiatry*, 197(3):234–243.

Rory C O'Connor and Olivia J Kirtley. 2018. The integrated motivational–volitional model of suicidal behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1754):20170268.

James W Pennebaker and Cindy K Chung. 2007. Expressive writing, emotional upheavals, and health. *Handbook of health psychology*, pages 263–284.

Anxo Pérez, Neha Warikoo, Kexin Wang, Javier Parapar, and Iryna Gurevych. 2022. Semantic similarity models for depression severity estimation. *arXiv preprint arXiv:2211.07624*.

David P Phillips. 1974. The influence of suggestion on suicide: Substantive and theoretical implications of the werther effect. *American sociological review*, pages 340–354.

Matthew C Podlogar, Peter M Gutierrez, and Thomas E Joiner. 2022. Past levels of mental health intervention and current nondisclosure of suicide risk among men older than age 50. *Assessment*, 29(8):1611–1621.

Huachuan Qiu, Lizhi Ma, and Zhenzhong Lan. 2024. Psyguard: An automated system for suicide detection and risk assessment in psychological counseling. *arXiv preprint arXiv:2409.20243*.

Bhanu Pratap Singh Rawat, Samuel Kovaly, Wilfred R Pigeon, and Hong Yu. 2022. Scan: suicide attempt and ideation events dataset. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2022, page 1029.

Koustuv Saha, Sindhu Kiranmai Ernala, Sarmistha Dutta, Eva Sharma, and Munmun De Choudhury. 2020. Understanding moderation in online mental health communities. In *HCII*. Springer.

Koustuv Saha and Amit Sharma. 2020. Causal factors of effective psychosocial outcomes in online mental health communities. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 590–601.

Koustuv Saha, Benjamin Sugar, John Torous, Bruno Abrahao, Emre Kıcıman, and Munmun De Choudhury. 2019. A social media study on the effects of psychiatric medication use. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Koustuv Saha, Asra Yousuf, Ryan L Boyd, James W Pennebaker, and Munmun De Choudhury. 2022. Social media discussions predict mental health consultations on college campuses. *Scientific reports*, 12(1):123.

Ramit Sawhney, Harshit Joshi, Lucie Flek, and Rajiv Shah. 2021a. 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 Shah. 2020. A time-aware transformer based model for suicide ideation detection on social media. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 7685–7697.

Ramit Sawhney, Harshit Joshi, Rajiv Shah, and Lucie Flek. 2021b. Suicide ideation detection via social and temporal user representations using hyperbolic learning. In *Proceedings of the 2021 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies*, pages 2176–2190.

Henk G Schmidt, Geoffrey R Norman, Silvia Mamede, and Mohi Magzoub. 2024. The influence of context on diagnostic reasoning: A narrative synthesis of experimental findings. *Journal of Evaluation in Clinical Practice*, 30(6):1091–1101.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.

Eva Sharma and Munmun De Choudhury. 2018. Mental health support and its relationship to linguistic accommodation in online communities. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13.

Chen Shen and 1 others. 2020. Suicide risk prediction using social media data. In *Proceedings of AAAI*.

Soorya Ram Shimgekar, Violeta J Rodriguez, Paul A Bloom, Dong Whi Yoo, and Koustuv Saha. 2025. Interpersonal theory of suicide as a lens to examine suicidal ideation in online spaces. *arXiv preprint arXiv:2504.13277*.

Daun Shin, Hyoseung Kim, Seunghwan Lee, Younhee Cho, and Whanbo Jung. 2024. Using large language models to detect depression from user-generated diary text data as a novel approach in digital mental health screening: instrument validation study. *Journal of Medical Internet Research*, 26:e54617.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Shiyu Teng, Jiaqing Liu, Rahul Kumar Jain, Shurong Chai, Ruibo Hou, Tomoko Tateyama, Lanfen Lin, and Yen-wei Chen. 2025. Enhancing depression detection with chain-of-thought prompting: From emotion to reasoning using large language models. *arXiv preprint arXiv:2502.05879*.

Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. 2018. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*, 32(3):588–601.

Sarah E Victor, Alison E Hipwell, Stephanie D Stepp, and Lori N Scott. 2019. Parent and peer relationships as longitudinal predictors of adolescent non-suicidal self-injury onset. *Child and adolescent psychiatry and mental health*, 13(1):1.

Yuxi Wang, Diana Inkpen, and Prasadith Kirinde Gamaarachchige. 2024. Explainable depression detection using large language models on social media data. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 108–126.

WHO. 2025. *Suicide worldwide in 2021: global health estimates*. World Health Organization.

Peter A Wyman, Trevor A Pickering, Anthony R Pisani, Kelly Rulison, Karen Schmeelk-Cone, Chelsey Hartley, Madelyn Gould, Eric D Caine, Mark LoMurray, Charles Hendricks Brown, and 1 others. 2019. Peer-adult network structure and suicide attempts in 38 high schools: Implications for network-informed suicide prevention. *Journal of Child Psychology and Psychiatry*, 60(10):1065–1075.

Yunhao Yuan, Koustuv Saha, Barbara Keller, Erkki Tapio Isometsä, and Talayeh Aledavood. 2023. Mental health coping stories on social media: A causal-inference study of papageno effect. In *Proceedings of the ACM Web Conference 2023*, pages 2677–2685.

Dongsong Zhang, Lina Zhou, Jie Tao, Tingshao Zhu, and Guodong Gao. 2025. Ketch: A knowledge-enhanced transformer-based approach to suicidal ideation detection from social media content. *Information Systems Research*, 36(1):572–599.

Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Kenny Q Zhu. 2022. Symptom identification for interpretable detection of multiple mental disorders. *arXiv preprint arXiv:2205.11308*.

Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. 2019. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 24–33.

A Appendix

Table A1: Model Performance on Remaining Data Combinations (epochs=20).

Model	Data Used	Accuracy	F1	Precision	Recall
M_8	Top neighbor's posts only	0.83	0.82	0.92	0.74
M_9	Target user posts + comments + top neighbor posts	0.91	0.92	0.91	0.92
M_{10}	Target user posts + comments + others' comments + top neighbor posts	0.71	0.64	0.75	0.56
M_{11}	Filtered neighbor posts only (no post/comments in `r/SuicideWatch`)	0.78	0.77	0.73	0.82

Table A2: Performance comparison of transformer models on the classification task.

Model	Accuracy	F1	Precision	Recall
microsoft/deberta-v3-large	**0.955**	**0.946**	**0.945**	**0.945**
roberta-large	0.944	0.931	0.944	0.919
bert-base-uncased	0.933	0.921	0.897	0.946
google/electra-large-discriminator	0.944	0.935	0.900	0.973

Table A3: Comparison of LIWC affect categories between M_4 and M_7 showing how the LIWC Affect category varies when considering all top neighbors vs. filtered top neighbors. Columns show mean proportions for users showing suicidal ideation (SI) and not, their difference, and the Kruskal–Wallis H-statistic. Significance codes: ***p<0.001, **p<0.01, *p<0.05.

Category	M_4				M_7			
	Case	Control	Diff.	H	Case	Control	Diff.	H
posemo	0.048	0.053	**-0.005**	23.568***	0.050	0.053	-0.003	17.482***
negemo	0.029	0.021	**0.007**	169.871***	0.032	0.025	**0.007**	72.531***
anx	0.003	0.002	0.000	62.663***	0.003	0.002	**0.001**	63.084***
anger	0.011	0.006	**0.005**	130.857***	0.010	0.006	0.004	51.436***
sad	0.009	0.005	**0.004**	181.551***	0.008	0.007	0.001	71.131***

Table A4: Comparison of LIWC affect categories across M_4, M_5, and M_6 showing how the Affect LIWC category varies across various neighbors (top, worst, random). Columns show mean proportions for users showing suicidal ideation (SI) and not, their difference, and the Kruskal–Wallis H-statistic. Significance codes: ***p<0.001, **p<0.01, *p<0.05.

Category	M_4				M_5				M_6			
	Case	Control	Diff.	H	Case	Control	Diff.	H	Case	Control	Diff.	H
posemo	0.048	0.053	**-0.005**	23.568***	0.047	0.049	-0.003	38.611***	0.048	0.047	0.001	42.425***
negemo	0.029	0.021	**0.007**	169.871***	0.029	0.025	0.004	74.947***	0.030	0.024	0.006	59.078***
anx	0.003	0.002	0.000	62.663***	0.004	0.002	**0.002**	53.518***	0.003	0.002	0.001	125.397***
anger	0.011	0.006	**0.005**	130.857***	0.009	0.006	0.003	40.154***	0.009	0.006	0.002	138.704***
sad	0.009	0.005	**0.004**	181.551***	0.007	0.007	0.000	71.131***	0.007	0.008	-0.001	36.640***

Table A5: Comparison of LIWC affect categories for M_8. Columns show mean proportions for suicidal and non-suicidal users, their difference, and the Kruskal–Wallis H statistic. Significance codes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Category	M_8			
	Case	Control	Diff	H
posemo	0.053	0.053	0.000	0.409
negemo	0.026	0.025	0.001	4.805*
anx	0.001	0.002	-0.001	3.323
anger	0.009	0.006	0.003	9.523**
sad	0.005	0.007	-0.001	3.638

Table A6: Comparison of LIWC affect categories for M_9. Columns show mean proportions for suicidal and non-suicidal users, their difference, and the Kruskal–Wallis H statistic. Significance codes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Category	M_9			
	Case	Control	Diff	H
posemo	0.050	0.053	-0.003	17.482***
negemo	0.032	0.025	0.007	72.531***
anx	0.003	0.002	0.001	63.084***
anger	0.010	0.006	0.004	51.436***
sad	0.008	0.007	0.001	71.131***

Table A7: Comparison of LIWC affect categories for M_{10}. Columns show mean proportions for suicidal and non-suicidal users, their difference, and the Kruskal–Wallis H statistic. Significance codes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Category	M_{10}			
	Case	Control	Diff	H
posemo	0.070	0.067	0.003	349.268***
negemo	0.031	0.028	0.003	538.074***
anx	0.003	0.002	0.000	403.691***
anger	0.010	0.009	0.001	281.206***
sad	0.007	0.006	0.001	361.992***

Table A8: Comparison of LIWC affect categories for M_{11}. Columns show mean proportions for suicidal and non-suicidal users, their difference, and the Kruskal–Wallis H statistic. Significance codes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Category	M_{11}			
	Case	Control	Diff	H
posemo	0.053	0.053	0.000	0.409
negemo	0.026	0.025	0.001	4.805*
anx	0.001	0.002	-0.001	3.323
anger	0.009	0.006	0.003	9.523**
sad	0.005	0.007	-0.001	3.638

Table A9: Performance metrics for different depth and neighbor configurations. **Depth** refers to the number of recursive depth considered, and **Neighbors** refers to the number of top neighbor posts included.

Depth	**Neighbors**	**Accuracy**	**F1**	**Precision**	**Recall**
1	1	0.98	0.97	0.96	0.98
1	2	0.97	0.96	0.95	0.97
1	3	0.96	0.95	0.95	0.96
1	4	0.96	0.95	0.94	0.96
1	5	0.95	0.94	0.94	0.95
2	1	0.97	0.96	0.95	0.97
2	2	0.96	0.95	0.94	0.96
2	3	0.95	0.94	0.93	0.95
2	4	0.95	0.94	0.93	0.95
2	5	0.94	0.93	0.92	0.94
3	1	0.96	0.95	0.94	0.96
3	2	0.95	0.94	0.93	0.95
3	3	0.94	0.93	0.92	0.94
3	4	0.94	0.93	0.92	0.94
3	5	0.95	0.94	0.93	0.95
4	1	0.95	0.94	0.93	0.95
4	2	0.94	0.93	0.92	0.94
4	3	0.93	0.92	0.91	0.93
4	4	0.93	0.92	0.91	0.93
4	5	0.92	0.91	0.90	0.92
5	1	0.94	0.93	0.92	0.94
5	2	0.93	0.92	0.91	0.93
5	3	0.92	0.91	0.90	0.92
5	4	0.92	0.91	0.90	0.92
5	5	0.91	0.90	0.89	0.91