Leakage-Aware User-Level ADHD Signal Classification from Social Media: When Graph Aggregation Helps, and When It Does Not

Daniel Cieślak[1,2] **Władysław Średniawa**[1]
[1]Samsung R&D Poland, Warsaw, Poland
[2]Gdańsk University of Technology, Gdańsk, Poland
{d.cieslak,w.sredniawa}@samsung.com
daniel.cieslak@pg.edu.pl

Abstract

User-level ADHD-related text classification from social media is methodologically challenging because predictions must aggregate many short posts, performance can be inflated by direct diagnostic leakage, and screening-adjacent settings require calibrated probabilities rather than discrimination alone. We introduce a leakage-aware evaluation framework organized around two controlled axes: an evidence window, defined as the maximum number of tweets available per user, and leakage control. Within this setup, we compare document-level transformers, strong non-graph embedding-pooling baselines, and heterogeneous GraphSAGE-style models combining semantic tweet embeddings, psycholinguistic features, and temporal structure. The main result is regime-dependent: the tested graph aggregation is most useful when user evidence is scarce, whereas simple embedding pooling becomes highly competitive and often slightly stronger as more evidence becomes available. Overall, the main contribution is a controlled benchmarking framework and a clearer account of when structure-aware aggregation is beneficial under fixed representation and leakage settings.

1 Introduction

There is growing interest in using NLP methods to detect mental-health-related and neurodiversity-related signals from natural language (Malgaroli et al., 2023). Social-media data are especially attractive in this context: they are longitudinal, user-centered, and rich in spontaneous language use. At the same time, they are noisy, weakly supervised, and highly vulnerable to methodological shortcuts (Han et al., 2013). ADHD is a particularly challenging target in this space (Barkley, 1997). Potential signal may appear not only in lexical content, but also in discourse organization, digressiveness, expressivity, topical variability, and stylistic instability across posts. Yet a model may appear strong simply because it detects explicit statements such as "I have ADHD", medication names, or overt diagnostic self-disclosure.

This motivates a more careful setup than the standard "train a text classifier and report AUC" recipe. In our view, a credible user-level ADHD signal classifier should satisfy three conditions. First, it should predict at the *user level* rather than at the single-post level, because individual tweets are short, noisy, and unstable. Second, it should be evaluated under explicit *leakage control*, so that results are not dominated by trivial label cues. Third, it should report not only discrimination metrics such as AUC or F1, but also probabilistic metrics such as Brier score and expected calibration error (ECE), because screening-adjacent settings depend on how trustworthy the predicted probabilities are.

In this paper, we study two orthogonal axes of evaluation. The first is an *evidence window*: the maximum number of tweets available per user, denoted by m. The second is *leakage regime*: whether texts are used in lightly normalized form or in a more strictly sanitized form that removes direct label cues. Within this framework, we compare three types of aggregation: document-level transformer baselines operating on concatenated tweets, non-graph pooling baselines built from frozen tweet embeddings, and heterogeneous GraphSAGE-style models that combine semantic tweet embeddings, psycholinguistic features, and temporal structure.

Our initial expectation was that graph structure would provide a general advantage by representing users as structured sets of related posts rather than as flat documents. The empirical picture is more informative. The tested heterogeneous GraphSAGE-style aggregation helps mainly under limited evidence: when only a small number of tweets is available, graph models with semantic embeddings and psycholinguistic features improve over simple embedding pooling, but this advantage shrinks and can reverse as more user evidence becomes avail-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 527–536

able. The key contribution of the paper is therefore not a claim of universal graph superiority, but a controlled analysis of *when* structure-aware aggregation matters.

Our contributions are threefold:

1. We introduce a leakage-aware, evidence-window-controlled evaluation protocol for user-level ADHD signal classification.
2. We compare document-level transformers, strong non-graph embedding-pooling baselines, and heterogeneous GraphSAGE-style aggregation under the same user-level setup.
3. We show that the tested graph aggregation is most useful in the low-evidence regime, while strong embedding pooling is highly competitive at moderate and larger evidence windows.

2 Related Work

Our work intersects several domains: user-level social-media NLP, graph-based text classification, as well as leakage-aware and calibration-aware evaluation. We frame our contribution primarily as a methodological study of aggregation under sparse evidence rather than as a claim that graph models uniformly outperform strong text encoders.

A first relevant line of work concerns social-media text classification under short, noisy, and weakly contextualized inputs. Prior studies on sentiment, emotion, and related social-media inference tasks show that single-post evidence is often brittle, which motivates aggregation across multiple posts or richer contextual modeling (Khemani et al., 2025; Gholami et al., 2023). Our setting differs from standard post-level classification because prediction is made at the *user* level. This makes the central problem not only representation learning for individual posts, but also reliable aggregation of many weak signals belonging to the same person.

A second line of work casts text classification as learning over graphs. In this literature, graph structure is used to encode relations such as co-occurrence, document–word incidence, or higher-order semantic dependencies, and surveys emphasize that graph construction is often a major determinant of performance and comparability (Wang et al., 2024). Recent work further shows that graph neural networks can be effective for short-text classification when local evidence is sparse and relational context becomes more informative (Sun et al., 2022). This perspective is relevant to our setting, where each user is represented not by a single long document, but by a collection of short posts that may benefit from structure-sensitive aggregation.

A closely related research direction combines pretrained language models with graph learning. Rather than replacing strong semantic encoders, these approaches typically use PLM-derived embeddings as node features and apply graph propagation on top of them (Yang and Cui, 2021; Sun et al., 2022; Lv et al., 2024). This is also the perspective adopted here. Our graph models operate over already informative tweet embeddings, and we compare them directly against strong non-graph pooling baselines built from the same underlying representations. This comparison is essential, because in modern NLP the practical question is rarely whether graphs outperform weak baselines, but whether they still add value once the text encoder is already strong.

Recent work also suggests that the benefit of explicit structure may be regime-dependent rather than universal. In particular, when language-model features are already strong, structural propagation does not necessarily improve performance in a uniform way (Xu, 2025). Our paper adopts exactly this empirical stance. We do not assume that graph aggregation should dominate across all settings; instead, we test whether its value depends on the amount of available evidence per user.

Another core axis of our work is evaluation hygiene. Prior work in graph-based learning has repeatedly shown that conclusions can be highly sensitive to benchmark construction, graph design, and evaluation protocol (Wang et al., 2024). Related work on graph pretraining likewise highlights the risk of trivial predictive pathways and the importance of avoiding leakage-like shortcuts in experimental design (Hu et al., 2020). In our setting, the most important shortcut risk is direct diagnostic leakage, such as explicit references to ADHD, treatment, medication, or overt self-identification. We therefore treat leakage control as a primary experimental variable rather than a preprocessing detail.

Finally, our evaluation is motivated not only by discrimination performance, but also by probabilistic reliability. In screening-adjacent or otherwise sensitive inference settings, calibration matters because useful decision support depends on confidence being aligned with empirical correctness. Prior work has treated metrics such as Brier score and expected calibration error as meaningful evaluation targets in addition to ranking-based measures

(Yang and Cui, 2021; Wang et al., 2023). This is especially important in our comparison of graph aggregation and non-graph pooling, since models with similar AUC may still differ materially in confidence quality and therefore in practical usability.

3 Methods and Materials

3.1 Dataset and Study Design

We study user-level ADHD signal classification on a public social-media benchmark introduced in prior ICWSM work, where each instance consists of one user, a set of tweets authored by that user, and a binary label indicating membership in the ADHD or control group (Singh et al., 2022). We treat this strictly as a *user-level* task: the model observes multiple tweets from the same user and outputs a single probability for the full profile.

After retaining only users with at least 20 available tweets, the processed cohort used in the main experiments contains 16,294 users: 8,095 in the ADHD group and 8,199 in the control group. We impose a controlled *evidence window* by keeping at most the most recent m tweets per user through tail sampling. This yields a standardized setup in which all methods are evaluated under the same maximum amount of available user evidence. We evaluate $m \in \{10, 25, 50, 100\}$ and focus interpretation on the stable regime.

We use tail sampling rather than random sampling because it better matches the intended sparse-profile setting: at prediction time, a system would typically observe the most recent available posts rather than a random subset of a user's history. Tail sampling also preserves local temporal continuity, which is important for the graph construction because adjacent-tweet and time-bin edges are meaningful only when the sampled posts remain chronologically coherent. This choice reduces sampling-induced variance, but it also means that our results should be interpreted for recent-profile evidence rather than arbitrary historical subsamples.

Our experimental design is built around two orthogonal axes. The first is the *evidence window*, which controls how many tweets are available per user at prediction time. The second is *leakage regime*, which controls whether explicit diagnostic and treatment-related cues remain in the text. In the **raw** regime, tweets are used after basic normalization. In the **hard** regime, explicit diagnostic, treatment-related, medication-related, and self-identifying cues are removed or masked. This makes it possible to distinguish genuine gains in user-level aggregation from gains driven by shortcut features.

All splits are performed at the user level, so all tweets from a given user belong entirely to either the training or the test set. We use an 80/20 train/test split and repeat all experiments with three random seeds. In the codebase, this corresponds to `test_size=0.2`. This prevents tweet-level leakage across splits and ensures that reported performance reflects generalization to unseen users rather than unseen posts from already observed users.

3.2 Materials and Representations

For graph-based models and non-graph embedding-pooling baselines, each tweet is encoded with a frozen MiniLM-based (Wang et al., 2020) sentence encoder, i.e., the same encoder used during preprocessing to produce the `*_minilm.*` embedding files in our pipeline. These embeddings are not fine-tuned during downstream user-level training; they are treated as fixed semantic representations so that the comparison focuses on aggregation rather than encoder adaptation.

In parallel, we compute tweet-level psycholinguistic and surface features. These include text length, number of words, lexical diversity, token entropy, character-type usage, punctuation statistics, repetition patterns, and selected stylistic indicators such as pronouns, negation markers, hedges, and intensifiers. These variables are used either as direct tweet-node inputs or as auxiliary graph structure, depending on the experimental condition.

3.3 User-Level Graph Construction

Our graph-based formulation represents each user as a separate heterogeneous graph, illustrated in Figure 1. Formally, for a user u with up to m tweets, we construct a graph $G_u = (V_u, E_u)$ with three node types: tweet nodes, feature nodes, and time nodes. One user corresponds to one graph, and there are no cross-user edges. This design keeps inference strictly profile-specific and avoids structural information leakage across users.

The full graph contains three main relation types. First, *tweet–tweet* edges connect adjacent tweets in sequence order and provide local chronological continuity. Second, *tweet–feature* edges connect tweets to psycholinguistic feature nodes, allowing the model to propagate information through shared stylistic or surface-level properties. Third, *tweet–time* edges connect tweets to coarse temporal bins

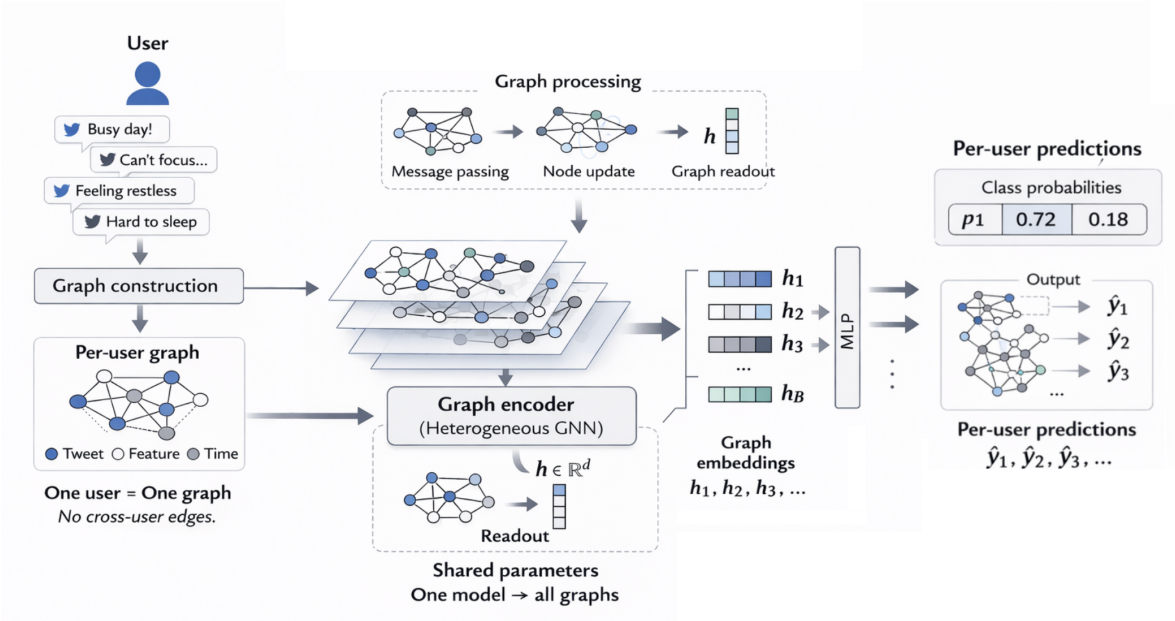

Figure 1: Overview of the proposed user-level aggregation pipeline. Each user is represented as a separate heterogeneous graph with no cross-user edges. Tweet nodes encode semantic content and optional psycholinguistic descriptors, while auxiliary feature and time nodes provide stylistic and temporal structure. A shared heterogeneous GNN maps each user graph to a graph-level embedding, which is passed to an MLP for final user-level prediction. This formulation allows direct comparison between structure-aware aggregation and non-graph pooling under controlled evidence windows and leakage regimes.

and provide a lightweight temporal scaffold.

We evaluate three tweet-node input settings. In `psych`, tweet nodes use psycholinguistic features only. In `emb`, tweet nodes use frozen sentence embeddings only. In `emb+psych`, tweet nodes use the concatenation of semantic embeddings and tweet-level psycholinguistic features. At $m = 50$, we additionally evaluate edge ablations that remove feature edges (`next+time`), remove time edges (`next+feature`), or remove psycholinguistic information from tweet representations entirely.

The graph encoder is implemented as a two-layer heterogeneous GraphSAGE (Hamilton et al., 2017) model with hidden size 128, dropout 0.2, batch size 16, and 30 training epochs. After relation-aware message passing, node representations are pooled into a graph-level embedding, which is then passed to a small MLP classifier to produce the final user-level probability.

3.4 Compared Model Families

We compare three model families under the same user-level setup.

Document-level transformer baselines. As strong text baselines, we concatenate tweets from a user into a single document and fine-tune encoder-only models for binary classification. We evaluate DistilRoBERTa (Sanh et al., 2019), RoBERTa-base (Liu et al., 2019), and ModernBERT-base (Warner et al., 2024). These models are trained for 3 epochs with batch size 8 and maximum sequence length 256. The transformer baselines are included as document-level references rather than as the central object of study. We use the same batch size and sequence length across all transformer variants to keep this comparison controlled and computationally comparable. The 256-token limit inevitably truncates longer profiles, especially at larger m, so these baselines should not be interpreted as an exhaustive evaluation of long-context transformer modeling. Instead, they represent a common practical baseline for concatenated-profile classification under a fixed context setting.

Non-graph embedding pooling. To isolate the effect of aggregation, we build user representations directly from the same frozen tweet embeddings used in the graph models. Our strongest non-graph baseline applies mean pooling over tweet embeddings followed by a small MLP classifier. In the main context-window comparison, mean pooling is the default aggregation rule. At $m = 50$, we additionally evaluate attention pooling and mean-max

pooling. The main mean-pooling model is trained for 20 epochs with batch size 64 and hidden size 256.

Graph-based aggregation. Our proposed graph-based family applies heterogeneous message passing over tweet, feature, and time nodes. The central comparison is therefore not between weak and strong text encoders, but between different *aggregation mechanisms* operating over the same underlying semantic tweet representations. This makes it possible to test the main methodological hypothesis of the paper: whether structure-aware aggregation is particularly useful when the amount of available user evidence is limited.

Scope of the aggregation comparison. Our primary comparison is not intended as a model-zoo evaluation over all possible text encoders or graph neural network architectures. Instead, we deliberately fix the tweet-level representation for the graph and non-graph pooling models to the same frozen MiniLM embeddings. This makes the comparison controlled: differences between mean pooling and graph aggregation reflect the aggregation mechanism rather than differences in the underlying language encoder. The document-level transformer baselines provide strong end-to-end references, but they answer a different question: how well a concatenated-profile encoder performs under a fixed context limit. We therefore interpret the results as evidence about aggregation under a fixed representation, not as a claim that the tested GraphSAGE model is the best possible graph architecture or that MiniLM is the best possible encoder.

3.5 Evaluation Protocol and Metrics

We report AUC, F1, Brier score (Brier, 1950), and expected calibration error (ECE) (Pakdaman Naeini et al., 2015). AUC is the main discrimination metric. F1 is computed by thresholding the positive-class probability at 0.5. Brier score and ECE quantify probabilistic reliability and calibration, which we treat as first-class evaluation targets in this screening-adjacent setting.

For selected main comparisons, we additionally report paired seed-level differences between model families in order to verify whether the observed evidence-window trend is directionally stable across random initializations. This is important because the main claim of the paper is not that one model class universally dominates, but that the usefulness of graph-based aggregation depends on the amount of user evidence available at prediction time.

4 Results

Table 1 presents the main comparison at $m = 50$. This table should be read as the core head-to-head benchmark: strong document-level transformers, strong non-graph user-level embedding pooling, and graph-based user-level aggregation.

Three points are immediate. First, all strong user-level aggregation methods outperform the document-level transformer baselines at $m = 50$. The best transformer baseline, ModernBERT-base, reaches AUC 0.8714, while the best user-level methods are near 0.887. This is a substantial gap for a setup in which the only difference is how user evidence is represented and aggregated. It suggests that for this task, user-level aggregation strategy matters at least as much as encoder choice.

Second, the graph does not dominate uniformly. In the raw regime, the full graph (`emb+psych`) reaches the best AUC and F1, but only by a narrow margin over mean pooling. In the hard regime, mean pooling is slightly stronger. This is already a useful result: graph aggregation is competitive with strong user-level pooling, but its benefit is conditional rather than universal.

Third, calibration favors the strongest non-graph baseline. At $m = 50$, mean pooling has the lowest ECE in both leakage regimes, and the best or tied-best Brier score. This matters because a method that improves AUC at the expense of poorly behaved probabilities may be less attractive in practice.

Figure 2 provides the clearest high-level view of the evidence-window effect: the tested graph aggregation is most beneficial when only a small amount of user evidence is available, whereas mean pooling becomes increasingly competitive as m increases. Table 2 presents the same pattern numerically.

4.1 Paired Comparison Against Mean Pooling

To verify that the evidence-window trend is not an artifact of aggregate means alone, we additionally computed paired per-seed differences relative to mean embedding pooling. Table 3 supports the same pattern already visible in Table 2: the `emb+psych` graph shows a small AUC advantage in the low-evidence regime ($m = 10$ and $m = 25$), this advantage becomes negligible around $m = 50$, and it reverses by $m = 100$. At the same time, mean pooling remains generally stronger in calibra-

Table 1: Main comparison at $m = 50$. Mean $\pm$ standard deviation over three seeds.

Model	AUC ↑	F1 ↑	Brier ↓	ECE ↓
Hard leakage				
ModernBERT-base	0.8688 ± 0.0089	0.7753 ± 0.0127	0.1566 ± 0.0049	0.0830 ± 0.0122
RoBERTa-base	0.8678 ± 0.0061	0.7925 ± 0.0042	0.1547 ± 0.0074	0.0765 ± 0.0236
DistilRoBERTa	0.8631 ± 0.0080	0.7881 ± 0.0022	0.1617 ± 0.0131	0.0940 ± 0.0306
Embedding pooling (mean)	$\mathbf{0.8858 \pm 0.0044}$	$\mathbf{0.8058 \pm 0.0068}$	$\mathbf{0.1371 \pm 0.0032}$	$\mathbf{0.0321 \pm 0.0020}$
Graph (emb)	0.8808 ± 0.0076	0.8016 ± 0.0135	0.1423 ± 0.0055	0.0436 ± 0.0106
Graph (emb+psych)	0.8838 ± 0.0066	0.8004 ± 0.0077	0.1445 ± 0.0053	0.0657 ± 0.0301
Raw leakage				
ModernBERT-base	0.8714 ± 0.0098	0.7779 ± 0.0151	0.1551 ± 0.0083	0.0829 ± 0.0170
RoBERTa-base	0.8685 ± 0.0049	0.7927 ± 0.0059	0.1564 ± 0.0062	0.0860 ± 0.0178
DistilRoBERTa	0.8655 ± 0.0062	0.7894 ± 0.0035	0.1555 ± 0.0057	0.0754 ± 0.0095
Embedding pooling (mean)	0.8870 ± 0.0045	0.8046 ± 0.0064	$\mathbf{0.1366 \pm 0.0033}$	$\mathbf{0.0343 \pm 0.0072}$
Graph (emb)	0.8843 ± 0.0071	0.8003 ± 0.0185	0.1412 ± 0.0062	0.0492 ± 0.0184
Graph (emb+psych)	$\mathbf{0.8876 \pm 0.0053}$	$\mathbf{0.8128 \pm 0.0103}$	0.1382 ± 0.0049	0.0475 ± 0.0196

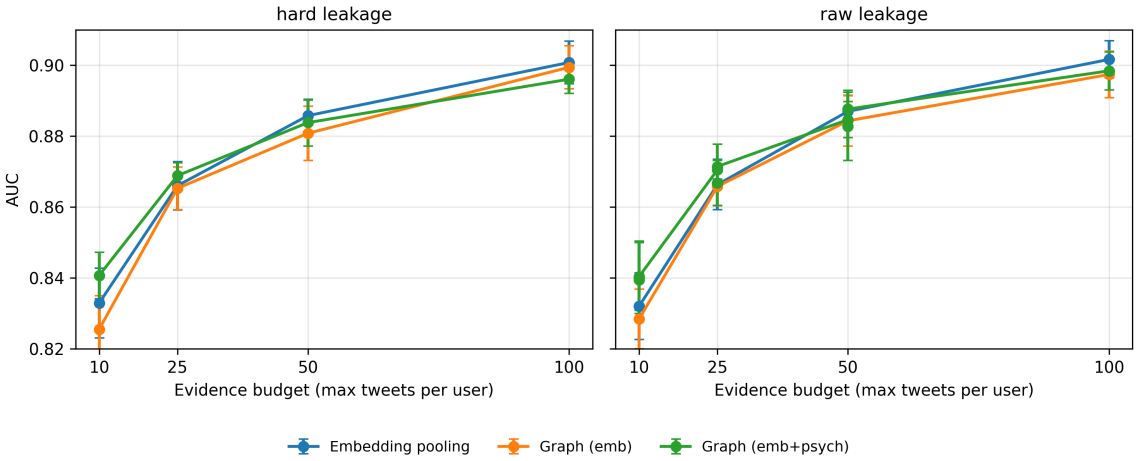

Figure 2: AUC as a function of the evidence window m for the main user-level aggregation methods. The tested graph aggregation with tweet embeddings and psycholinguistic features is most useful in the low-evidence regime ($m = 10, 25$), while simple mean pooling becomes highly competitive and often slightly stronger when more tweets are available. Error bars show standard deviation across three seeds.

tion, as the graph variants usually do not improve Brier score or ECE and often underperform on both metrics. We therefore interpret the tested graph aggregation as a regime-dependent method whose main value appears when user evidence is sparse, whereas simple embedding pooling remains the stronger default once more evidence is available.

The advantage shrinks as more evidence becomes available. At $m = 50$, the difference between the strongest graph variant and mean pooling becomes very small: in raw, the graph is marginally ahead; in hard, mean pooling is slightly better. At $m = 100$, mean pooling is best in both regimes. This indicates that structural aggregation is not uniformly beneficial. Its main value is concentrated in the low-evidence regime.

The sweep over m also clarifies the role of semantic representations. The embedding-only graph is consistently much stronger than earlier psycholinguistic-only graph results at $m = 50$ (Table 4), confirming that tweet semantics dominate the predictive signal. Psycholinguistic features provide an additional boost mainly when evidence is scarce. Methodologically, this is exactly why the evidence-window perspective matters. If we had reported only a single setting, especially a moderate or large value of m, we could easily have concluded either that the graph does not help or that the graph only marginally helps. The sweep over m shows a more useful pattern: the tested graph aggregation matters primarily when the task is genuinely underdetermined by limited evidence.

To understand where the gains come from, Table 4 compares graph variants and non-graph em-

Table 2: Evidence-window analysis for user-level aggregation methods. Mean ± standard deviation over three seeds. We focus interpretation on the stable range $m \leq 100$.

Leakage	Model	$m = 10$	$m = 25$	$m = 50$	$m = 100$
hard	Embedding pooling (mean)	0.8329 ± 0.0098	0.8660 ± 0.0068	$\mathbf{0.8858 \pm 0.0044}$	$\mathbf{0.9008 \pm 0.0060}$
	Graph (emb)	0.8255 ± 0.0095	0.8653 ± 0.0060	0.8808 ± 0.0076	0.8994 ± 0.0061
	Graph (emb+psych)	$\mathbf{0.8407 \pm 0.0066}$	$\mathbf{0.8688 \pm 0.0037}$	0.8838 ± 0.0066	0.8961 ± 0.0041
raw	Embedding pooling (mean)	0.8320 ± 0.0094	0.8663 ± 0.0071	0.8870 ± 0.0045	$\mathbf{0.9016 \pm 0.0052}$
	Graph (emb)	0.8284 ± 0.0084	0.8658 ± 0.0055	0.8843 ± 0.0071	0.8974 ± 0.0066
	Graph (emb+psych)	$\mathbf{0.8404 \pm 0.0096}$	$\mathbf{0.8714 \pm 0.0063}$	$\mathbf{0.8876 \pm 0.0053}$	0.8984 ± 0.0054

Table 3: Paired per-seed deltas relative to mean embedding pooling under the original three-seed protocol. Positive ΔAUC and ΔF1 favor the graph model, whereas negative ΔBrier and ΔECE favor the graph model.

Model	**Leakage**	m	**ΔAUC**	**ΔF1**	**ΔBrier**	**ΔECE**
Graph (`emb+psych`)	hard	10	+0.0078	+0.0186	−0.0025	+0.0073
		25	+0.0028	+0.0072	+0.0077	+0.0360
		50	−0.0020	−0.0055	+0.0074	+0.0336
		100	−0.0047	−0.0083	+0.0158	+0.0523
	raw	10	+0.0084	+0.0136	−0.0058	−0.0107
		25	+0.0051	−0.0013	+0.0020	+0.0251
		50	+0.0006	+0.0082	+0.0016	+0.0132
		100	−0.0032	−0.0105	+0.0088	+0.0310
Graph (`emb`)	hard	10	−0.0074	−0.0008	+0.0041	+0.0009
		25	−0.0008	+0.0111	+0.0037	+0.0155
		50	−0.0050	−0.0043	+0.0051	+0.0115
		100	−0.0014	−0.0083	+0.0075	+0.0308
	raw	10	−0.0036	+0.0166	+0.0023	+0.0023
		25	−0.0005	+0.0135	+0.0031	+0.0205
		50	−0.0027	−0.0044	+0.0046	+0.0149
		100	−0.0042	−0.0132	+0.0084	+0.0253

bedding pooling at $m = 50$.

The ablations support three conclusions. First, semantic embeddings drive most of the predictive power. The psycholinguistic-only graph is much weaker than all embedding-based methods in both regimes. This rules out an interpretation in which the gains mainly come from shallow style features.

Second, psycholinguistic features provide a smaller complementary gain. Moving from `emb` to `emb+psych` tends to help, but the gain is modest and not always monotonic across metrics. This is especially clear in the hard regime, where the full graph is only slightly above the embedding-only graph.

Third, edge-type ablations do not reveal a single magic relation. Removing time edges or feature edges changes performance, but not catastrophically. This suggests that the advantage of the full graph is distributed across its overall aggregation pattern rather than concentrated in one edge type.

For completeness, Table 5 isolates the document-level transformer comparison in the raw regime at $m = 50$. These are respectable baselines, but they remain below the best user-level aggregation methods, reinforcing the methodological point that aggregation design is a major determinant of performance in this task. Among the strongest models at $m = 50$, the non-graph mean-pooling baseline consistently yields the best calibration, especially in ECE.

5 Discussion

The main message of this paper is not that graphs uniformly outperform simpler methods. That claim would not be supported by the evidence. The actual result is more useful: **the tested heterogeneous GraphSAGE-style aggregation helps primarily in the low-evidence regime**. This is exactly the regime in which user-level mental-health-related classification is hardest and where structural inductive bias is most plausible.

This interpretation is also supported by paired seed-level comparisons against mean pooling under the original protocol. Those comparisons preserve the same directional pattern: `emb+psych` tends to improve AUC at small evidence windows, but the

Table 4: Graph and pooling ablations at $m = 50$. Mean $\pm$ standard deviation over three seeds.

Model	Leakage	AUC ↑	F1 ↑	Brier ↓	ECE ↓
Embedding pooling (mean)	hard	$\mathbf{0.8858 \pm 0.0044}$	$\mathbf{0.8058 \pm 0.0068}$	$\mathbf{0.1371 \pm 0.0032}$	$\mathbf{0.0321 \pm 0.0020}$
Graph (emb+psych)		0.8848 ± 0.0063	0.8042 ± 0.0079	0.1375 ± 0.0047	0.0323 ± 0.0025
Graph (emb+psych, no feature)		0.8842 ± 0.0058	0.8101 ± 0.0079	0.1464 ± 0.0046	0.0806 ± 0.0319
Graph (emb+psych, no time)		0.8833 ± 0.0039	0.7994 ± 0.0070	0.1448 ± 0.0019	0.0635 ± 0.0256
Graph (emb)		0.8814 ± 0.0087	0.7921 ± 0.0286	0.1431 ± 0.0097	0.0504 ± 0.0263
Embedding pooling (attention)		0.8812 ± 0.0072	0.7959 ± 0.0123	0.1410 ± 0.0056	0.0375 ± 0.0086
Embedding pooling (meanmax)		0.8741 ± 0.0062	0.7987 ± 0.0050	0.1457 ± 0.0073	0.0409 ± 0.0264
Graph (psych)		0.8508 ± 0.0039	0.7564 ± 0.0234	0.1620 ± 0.0069	0.0589 ± 0.0317
Embedding pooling (mean)	raw	$\mathbf{0.8870 \pm 0.0045}$	0.8046 ± 0.0064	$\mathbf{0.1366 \pm 0.0033}$	$\mathbf{0.0343 \pm 0.0072}$
Graph (emb+psych)		0.8866 ± 0.0056	0.8040 ± 0.0186	0.1390 ± 0.0053	0.0506 ± 0.0094
Graph (emb+psych, no time)		0.8851 ± 0.0052	$\mathbf{0.8090 \pm 0.0037}$	0.1383 ± 0.0034	0.0377 ± 0.0033
Graph (emb+psych, no feature)		0.8845 ± 0.0064	0.8076 ± 0.0049	0.1414 ± 0.0049	0.0579 ± 0.0041
Graph (emb)		0.8842 ± 0.0085	0.8029 ± 0.0151	0.1407 ± 0.0062	0.0468 ± 0.0147
Embedding pooling (attention)		0.8831 ± 0.0064	0.7978 ± 0.0114	0.1396 ± 0.0044	0.0316 ± 0.0083
Embedding pooling (meanmax)		0.8746 ± 0.0063	0.8005 ± 0.0055	0.1462 ± 0.0074	0.0425 ± 0.0261
Graph (psych)		0.8529 ± 0.0019	0.7468 ± 0.0517	0.1684 ± 0.0153	0.0854 ± 0.0662

Table 5: Transformer baselines for document-level classification ($m = 50$, raw regime).

Model	AUC ↑	F1 ↑	Brier ↓	ECE ↓
ModernBERT	$\mathbf{0.8714 \pm 0.0098}$	0.7779 ± 0.0151	$\mathbf{0.1551 \pm 0.0083}$	0.0829 ± 0.0170
RoBERTa	0.8685 ± 0.0049	$\mathbf{0.7927 \pm 0.0059}$	0.1564 ± 0.0062	0.0860 ± 0.0178
DistilRoBERTa	0.8655 ± 0.0062	0.7894 ± 0.0035	0.1555 ± 0.0057	$\mathbf{0.0754 \pm 0.0095}$

margin is small, diminishes as more tweets become available, and does not translate into better calibration, where mean pooling remains consistently stronger.

This finding has two methodological implications. First, future work in this area should avoid reporting only a single evidence setting, because doing so can hide the conditions under which a method is actually useful. Second, graph-based approaches should be evaluated against strong non-graph aggregation baselines, not only against psycholinguistic feature models or weak text baselines.

The paper also sharpens the interpretation of psycholinguistic features. They are helpful, but they are not the main source of predictive signal. Most of the usable signal is already present in semantic tweet embeddings. Psycholinguistic features appear to provide a complementary benefit chiefly when the evidence window is small and the model must extract more from less.

Finally, we emphasize that these results should not be interpreted as evidence for clinical ADHD diagnosis from tweets. The task studied here is user-level classification of ADHD-related signal in a benchmark dataset under a controlled experimental protocol. Even strong performance in this setting does not imply readiness for diagnostic or decision-support use.

6 Conclusion

We introduced a leakage-aware, evidence-window-controlled evaluation framework for user-level ADHD classification from social media and used it to compare document-level transformers, strong non-graph embedding pooling, and heterogeneous GraphSAGE-style aggregation.

The main contribution of this work is methodological. We show that conclusions depend strongly on both leakage regime and evidence window. Our central empirical finding is that the tested graph aggregation is useful primarily in the low-evidence regime: when only a small number of tweets is available, graph-based aggregation with tweet embeddings and psycholinguistic features is consistently strong. As more user evidence becomes available, however, simple embedding pooling becomes highly competitive and often slightly better, especially in calibration.

We believe this is the most useful takeaway for future work. The important question is not whether graph models win universally, but under which data conditions and evidence constraints they actually add value. In user-level ADHD-related language modeling, our results suggest that the answer is: mainly when evidence is sparse and aggregation is genuinely difficult.

7 Future Work

Several directions follow naturally from this study. First, the present work uses a fixed sentence-embedding backbone for graph and pooling models. A natural next step is to compare against stronger embedding extractors from larger foundation models, including modern API-based embeddings, to test whether the usefulness of graph aggregation persists as the underlying semantic representations become even stronger. Second, future work should test whether the same evidence-window pattern holds across graph architectures such as GCN, GAT, GIN, and graph-transformer variants.

Third, it would be valuable to extend the analysis beyond binary user-level classification and study robustness under stronger distribution shifts, alternative preprocessing regimes, and more realistic screening-oriented calibration settings.

Finally, the broader methodological lesson of this paper should be tested on other user-level social-media inference tasks: if graph aggregation is primarily beneficial in sparse-evidence regimes, this may be a more general phenomenon rather than one specific to ADHD-related language modeling.

8 Limitations

This study has several limitations. First, all experiments are conducted on a single public benchmark, Twitter-STMHD. The evidence-window trend is therefore an empirical finding within this dataset and should not be treated as a universal property of ADHD-related language modeling or of graph aggregation for mental-health inference. Replication on additional user-level social-media datasets and other screening-adjacent tasks is necessary before making broader claims.

Second, the graph-side experiments use one heterogeneous GraphSAGE-style encoder. We chose GraphSAGE because it naturally supports inductive graph-level prediction over separate per-user graphs and provides a simple message-passing baseline for sparse tweet graphs. However, the current experiments do not establish whether the same trend would hold for GCN, GAT, GIN, or transformer-style graph encoders. We therefore avoid interpreting the results as architecture-invariant evidence about all GNNs.

Third, the graph and pooling models use a fixed MiniLM-based sentence-embedding backbone. This was a deliberate control choice that isolates aggregation effects, but it also means that the results do not answer whether stronger embedding extractors, larger language models, or end-to-end tweet encoders would change the relative value of graph aggregation.

Fourth, the labels reflect assignments in the source benchmark rather than independently verified clinical diagnoses. Social-media users are not representative of the broader population, and language-based models may capture demographic, cultural, or platform-specific correlates rather than ADHD-specific signal. Even after leakage control, residual shortcuts may remain.

Because the task concerns sensitive mental-health-related inference, the proposed models should be treated strictly as research tools for studying linguistic signal and aggregation methodology. The intended use of this work is methodological benchmarking, not individual-level screening or intervention. We do not recommend these models for clinical, administrative, educational, or employment decisions about individuals.

References

Russell A. Barkley. 1997. Behavioral inhibition, sustained attention, and executive functions: Constructing a unifying theory of ADHD. *Psychological Bulletin*, 121(1):65–94.

Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Fatemeh Gholami, Zahed Rahmati, Alireza Mofidi, and Mostafa Abbaszadeh. 2023. On enhancement of text classification and analysis of text emotions using graph machine learning and ensemble learning methods on non-english datasets. *Algorithms*, 16(10):470.

William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30.

Bo Han, Paul Cook, and Timothy Baldwin. 2013. Lexical normalization for social media text. *ACM Transactions on Intelligent Systems and Technology*, 4(1):1–27.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, Kai-Wei Chang, and Yizhou Sun. 2020. GPT-GNN: Generative pre-training of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1857–1867.

Bharti Khemani, Shruti Patil, Sachin Malave, and Jaya Gupta. 2025. Improved graph convolutional network for emotion analysis in social media text. *MethodsX*, 14:103325.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *Preprint*, arXiv:1907.11692.

Shaoqing Lv, Jungang Dong, Chichi Wang, Xuanhong Wang, and Zhenan Bao. 2024. RB-GAT: A text classification model based on RoBERTa-BiGRU with graph attention network.

Matteo Malgaroli, Thomas D. Hull, James M. Zech, and Tim Althoff. 2023. Natural language processing for mental health interventions: A systematic review and research framework. *Translational Psychiatry*, 13(1).

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Asmit Kumar Singh, Udit Arora, Somyadeep Shrivastava, Aryaveer Singh, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. Twitter-STMHD: An extensive user-level database of multiple mental health disorders. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1182–1191.

Zhongtian Sun, Anoushka Harit, Alexandra I. Cristea, Jialin Yu, Lei Shi, and Noura Al Moubayed. 2022. Contrastive learning with heterogeneous graph attention networks on short text classification. In *Proceedings of the International Joint Conference on Neural Networks*.

Haotian Wang, Zhen Zhang, Mengting Hu, Qichao Wang, Liang Chen, Yatao Bian, and Bingzhe Wu. 2023. RECAL: Sample-relation guided confidence calibration over tabular data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7167–7177.

Kunze Wang, Yihao Ding, and Soyeon Caren Han. 2024. Graph neural networks for text classification: A survey. *Artificial Intelligence Review*, 57.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Haotian Xu. 2025. When structure doesn't help: LLMs do not read text-attributed graphs as effectively as we expected. *Preprint*, arXiv:2511.16767.

Yiping Yang and Xiaohui Cui. 2021. BERT-enhanced text graph neural network for classification. *Entropy*, 23(11):1536.