Understanding the Sociocultural Dimensions of Mental Health Discourse in Arabic-Language X Communities

Amal Alqahtani [1], Rana Salama [2], and Mona Diab[3]
[1]King Saud University, Riyadh, KSA
[2]Faculty of Computers and Artificial Intelligence, Cairo University, Egypt
[3]Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
amaalqahtani@ksu.edu.sa, r.aref@fci-cu.edu.eg, mdiab@andrew.cmu.edu

Abstract

Computational mental health research has predominantly centered on English-speaking populations, leaving Arabic-language discourse comparatively under-examined. We present an exploratory computational study of **8,147** tweets from **607** users classified by a GPT-4.1 personal-disclosure pipeline as likely lived-experience authors in three condition-specific Arabic-language X (formerly Twitter) Communities. We focus on discourse related to borderline personality disorder (BPD), bipolar disorder, and ADHD, and characterize community-associated linguistic patterns using a multi-domain cultural keyword framework. The results suggest that in this corpus, Bipolar tweets contain more religious and medical vocabulary, BPD tweets contain more relational, identity, and emotional-distress vocabulary, and ADHD tweets more often focus on practical symptoms and medication management. We treat these patterns as hypothesis-generating rather than confirmatory because the corpus is imbalanced across conditions, some subcorpora are temporally concentrated, and the keyword framework is an initial operationalization rather than a validated measurement instrument. The paper contributes a reusable LLM-assisted personal-disclosure pipeline and an exploratory cultural keyword framework for Arabic mental health discourse.

1 Introduction

In the Arab world, mental health discourse is strongly shaped by sociocultural norms. In particular, stigma associated with family honor and traditional religious interpretive frameworks continues to impede clinical help-seeking (Dardas and Simmons, 2015; Zolezzi et al., 2018). Despite these barriers, condition-specific Arabic-language social media communities have emerged as important spaces for peer support and the exchange of lived experiences. Nevertheless, these communities remain largely underexplored in computational mental health research. Existing computational mental health work has focused overwhelmingly on English, largely framing the problem as supervised classification of at-risk individuals (Coppersmith et al., 2014; De Choudhury et al., 2013). Arabic NLP has been advanced by transformer-based models such as AraBERT (Antoun et al., 2020) and MARBERT (Abdul-Mageed et al., 2021), which establish strong baselines across Arabic NLP benchmarks, yet Arabic mental health NLP has paid comparatively less attention to culturally situated discourse characterization.

We adopt a *characterization-oriented* approach grounded in Computational Social Science (Lazer et al., 2009). By moving beyond the diagnostic paradigm, we prioritize a descriptive analysis of community-associated discourse, investigating how users articulate mental health experiences while avoiding clinical inferences about diagnosis or patient status. We analyze **8,147** tweets from **607** users across three condition-specific X communities (BPD, Bipolar, ADHD) using a GPT-4.1 personal-disclosure pipeline validated against human annotators.

Our main contributions include:

1. **Dataset:** We introduce a multi-condition Arabic mental health corpus comprising 9,582 preprocessed posts, reduced to 8,147 posts after personal-disclosure filtering.[1]
2. **Annotation Pipeline:** We develop a GPT-4.1-based tweet-level classification pipeline augmented with a reason-tag taxonomy and confidence scoring, and validate its outputs against a human-annotated gold standard.
3. **Discourse Analysis:** We conduct a comparative discourse analysis using circadian activity profiling, weighted log-odds, non-negative matrix factorization (NMF) topic modeling, and

[1]Available at: https://github.com/amalqahtani/arabic-x-mental-health-discourse.

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 286–304

a six-domain cultural keyword framework.

4. **Empirical Findings:** We identify preliminary community-associated discourse patterns, including religious and medical vocabulary in Bipolar communities, identity and distress oriented language in BPD communities, and practical discussions of symptoms and medication management in ADHD communities. Given corpus imbalance and related methodological limitations, all analyses are interpreted as hypothesis-generating rather than confirmatory, and no between-community significance testing is performed.

2 Related Work

Computational mental health on social media. De Choudhury et al. (2013) showed that behavioral and linguistic signals in Twitter data predict depression onset. Coppersmith et al. (2014) established a scalable self-reported diagnosis methodology and demonstrated condition-level linguistic differences across post-traumatic stress disorder (PTSD), depression, bipolar disorder, and social anxiety disorder (SAD). Coppersmith et al. (2015) extended this to ten conditions. More recently, Yang et al. (2023) explored LLM-generated explanations for mental health severity assessment, highlighting persistent challenges in grounding model outputs within clinically interpretable frameworks. Our work adapts the condition comparative paradigm to Arabic, using LLMs for personal disclosure annotation rather than classification, and reserving analysis for interpretable statistical methods.

Arabic mental health, stigma, and cultural context. Mental health discourse in Arab societies is deeply shaped by social, religious, and cultural frameworks that influence how psychological distress is understood and discussed. Dardas and Simmons (2015) discussed that mental illness stigma in Arab societies is closely intertwined with family honor norms and religious interpretive frameworks, while Zolezzi et al. (2018) confirmed these patterns in a systematic review spanning 33 studies. The theory of explanatory models (Kleinman, 1980), which encompasses biomedical, spiritual, and relational interpretations of illness, provides the conceptual foundation for our analysis. In Arab contexts, psychological distress may be interpreted through biomedical, religious, and supernatural frameworks, including explanations such as the evil eye (*hasad*) or jinn possession (*mass*), with religious or traditional healing sometimes considered alongside medical treatment (Eid et al., 2025). Emerging NLP evidence further reflects this explanatory pluralism. Zaghouani et al. (2026) found that religious and therapeutic vocabulary appear with comparable frequency in Arabic stress discourse, while Ayash et al. (2025) proposed that patient questions are frequently grounded in relational and faith-based reasoning that extends beyond conventional clinical taxonomies. Motivated by these findings, our work explicitly annotates sociocultural framing in Arabic mental health discourse, examining the co-occurrence of Social, Cultural, Religious, Medical, and Stigma dimensions across online mental health communities.

LLM-Assisted Annotation for NLP. Although traditional NLP classification pipelines rely on human-annotated ground truth, recent studies suggest that large language models (LLMs) can serve as effective annotators for complex and subjective tasks, with performance depending on the task domain, language, and prompting strategy (Gilardi et al., 2023; Ding et al., 2023). In this work, we employ GPT-4.1 as the primary annotator for personal-disclosure identification (Section 3.3) and assess reliability through human validation rather than assuming human-level equivalence. Collectively, these three bodies of work motivate our approach. We extend the condition-comparative framework of Coppersmith et al. (2014) to Arabic, use LLMs for personal-disclosure annotation rather than clinical classification, operationalize Kleinman's 1980 explanatory models computationally, and employ weighted log-odds, NMF topic modeling, and a cultural keyword framework to characterize sociocultural discourse.

3 Methodology

3.1 Corpus Collection

We collected Arabic-language posts from condition-specific X Communities on X (formerly Twitter) using publicly accessible platform data. These X Communities are moderated spaces in which users join around shared interests and agree to community-specific participation rules prior to posting [2]. Moderation practices vary across communities, ranging from professionally supervised spaces led by licensed psychologists to peer-supported groups. In Arabic-speaking contexts, where mental health conditions remain highly stigmatized

[2] https://help.x.com/en/using-x/communities

(Dardas and Simmons, 2015; Zolezzi et al., 2018), participation in condition-specific communities reflects meaningful engagement with mental health discourse. We therefore do not treat community membership as a diagnostic indicator; prior work has demonstrated that affiliation-based proxy signals perform poorly against clinical ground truth (Ernala et al., 2019). Instead, we use community structure as a pragmatic sampling frame for condition-relevant discourse (AbouWarda et al., 2024). The resulting corpus consists of **10,091 tweets** collected from three condition-specific X Communities: BPD, Bipolar, and ADHD. Throughout this paper, these labels are capitalized when referring to the corresponding X Communities as data sources, whereas the associated clinical conditions (*borderline personality disorder*, *bipolar disorder*, and *attention-deficit/hyperactivity disorder*) are written in lowercase. Data were collected between March 31, 2022, and February 12, 2026. Each condition was represented by one Arabic-language X Community. Communities were selected based on four criteria: (1) an explicit focus on a specific mental health condition, (2) Arabic as the primary language of discourse, (3) active moderation, and (4) sustained member engagement. Additional community details are provided in (Appendix I).

3.2 Pre-processing

From the initial set of 10,091 raw tweets, we removed URL-only posts, non-Arabic and non-English content, duplicate entries, and single-token tweets, resulting in a preprocessed corpus of **9,582** tweets produced by **1,286** unique users. The corpus is predominantly Arabic (98.4%, n=9,428), with a small English component (1.2%, n=116) and code-mixed undefined tweets (0.4%, n=38). Figure 1 presents an overview of the complete data collection, annotation, and filtering pipeline.

3.3 LLM-Assisted Annotation and Human Validation Framework

To identify users whose posts contain evidence of personal mental health disclosure, we applied a tweet-level personal-disclosure classification pipeline to all 9,582 preprocessed tweets from 1,286 unique users. Classification was performed using GPT-4.1[3] as the primary annotator, with Qwen3-235B-A22B [4] running the same prompt in parallel as a conservative screening model, both at temperature=0.0, max_tokens=250. Each tweet was classified as either POSITIVE (containing evidence of personal mental health disclosure) or NEGATIVE (containing no such evidence) with the user bio incorporated as supporting context. The prompt encodes explicit classification rules, a bio override rule, a conservative NEGATIVE default, and a 13 tag reason taxonomy (Appendix A). Disagreements between the two models flag ambiguous cases; inter-model agreement is reported in Appendix D. User-level aggregation was derived from tweet-level classifications using a deterministic priority-ordered aggregation procedure (Appendix A.7). The resulting operational grouping was used exclusively for corpus filtering and downstream discourse analysis, and does not constitute a clinical or diagnostic categorization.

Likely personal-disclosure authors are operationally defined as users for whom at least one tweet contains a personal-disclosure signal or whose bio includes self-identification language; this designation does not imply or establish a clinical diagnosis. **Other** users are those whose tweets contain no detected personal-disclosure signals and whose bios contain no self-identification indicators, including professionals, caregivers, and general community participants. Full aggregation rules are provided in Appendix A.7. Pipeline reliability is assessed through a human validation study (Section 3.4 and Appendix D). Under the GPT-4.1 primary annotation framework, **607** of 1,286 users (**47.2%**) were labeled as likely personal-disclosure authors, while **679** users (**52.8%**) showed no detectable personal-disclosure signals. Among the 607 users identified by GPT-4.1, **528** were also identified by Qwen3, whereas **79** represented GPT-positive/Qwen-negative disagreement cases. Users without detected personal-disclosure signals under GPT-4.1 were excluded from downstream analysis, resulting in a final dataset of **8,147** tweets from **607** users. In total, **1,435** tweets (**15.0%**) were excluded together with their associated authors.

3.4 Human Validation

Two native Arabic-speaking annotators with prior experience in Arabic NLP independently annotated a stratified sample of 200 tweets using the annotation guidelines described in Appendix D. At the tweet-level, GPT-4.1 labeled 47.4% of tweets as positive and Qwen3 labeled 42.9%, reaching consensus on 90.8% ($\kappa = 0.84$; Appendix D). As the results below show, human validation aligns more strongly with GPT-4.1, whose labels define the final

[3]Model:`GPT-4.1-2025-04-14`

[4]Model:`Qwen3-235B-A22B-Instruct-2507`

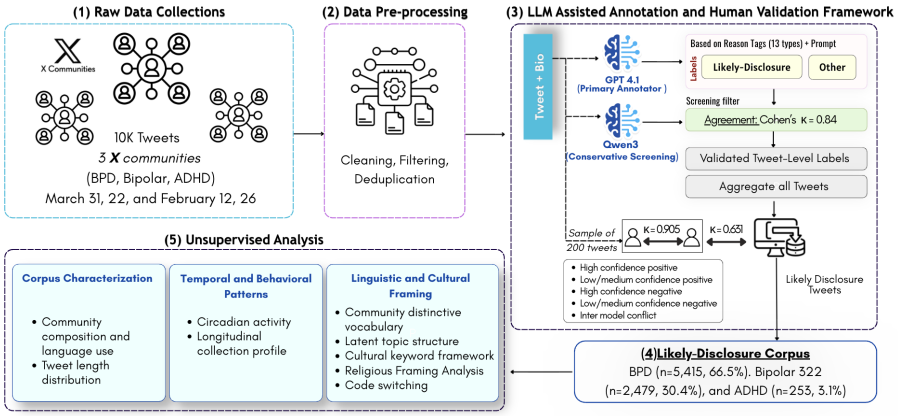

Figure 1: Overview of the computational pipeline. Exact collection period: March 31, 2022–February 12, 2026.

corpus.

Inter human agreement. The two annotators agreed on 192 of 200 tweets (raw agreement 96.00%), yielding $\kappa = 0.905$ which corresponds to almost perfect agreement (Landis and Koch, 1977). All eight disagreements were directionally consistent, reflecting differences in thresholding for ambiguous positive cases rather than fundamentally conflicting interpretations. This high agreement ceiling suggests that the annotation task is well-defined and that the labeling guidelines are internally consistent.

GPT-4.1 against human gold. Using the 192 mutually agreed tweets as the reference set, GPT-4.1 achieved $\kappa = 0.631$ (substantial agreement), with precision of 0.92, recall of 0.85, and $F_1 = 0.88$ on the positive personal-disclosure class. This places GPT-4.1 approximately 0.27 κ points below the inter-annotator agreement ceiling, supporting its use as the primary annotation model in this study.

Qwen3-235B-A22B against human gold. After excluding parse failures, Qwen3 achieved $\kappa = 0.329$ (fair agreement) against the human reference set, primarily due to lower recall (0.61) on the positive class. These results suggest that the high inter-model agreement between GPT-4.1 and Qwen3 ($\kappa = 0.84$) is largely driven by agreement on clear and predominantly negative cases rather than by near-human annotation reliability. Accordingly, Qwen3 was used as a conservative screening model whose disagreements with GPT-4.1 were treated as indicators of potentially ambiguous tweets, rather than as independent validation.

Stratum-level results. GPT-4.1 achieved its highest agreement with human annotations on clear-label strata, with performance declining to 46% agreement on low and medium confidence negative strata, where the conservative NEGATIVE default appears to suppress some genuine disclosures. Qwen3 agreed with human annotations on only 18% of tweets within the inter-model conflict stratum, further indicating that such cases require human adjudication. Agreement also varied across communities, with the highest agreement observed for ADHD ($\kappa = 0.73$), followed by BPD ($\kappa = 0.66$), and the lowest for Bipolar ($\kappa \approx 0.49$). The lower agreement for Bipolar is consistent with the more indirect and metaphorical disclosure style observed in that community. Full per-stratum results and complete agreement tables are provided in Appendix D (Tables 4 and 5).

4 Exploratory Analyses of the Self-Disclosure-Filtered Corpus

This section presents exploratory analyses conducted on the final corpus of **8,147** tweets produced by **607** users classified as likely personal-disclosure authors. The downstream analyses combine statistical and dictionary-based methods; however, topic interpretations and keyword-domain assignments involve researcher judgment and should therefore be regarded as exploratory analytical constructs rather than validated annotations. For clarity, we organize the analyses into three thematic groups: (i) Corpus Characterization, (ii) Temporal and Behavioral Patterns, and (iii) Linguistic and Cultural Framing. These analyses include weighted log-odds distinctive vocabulary analysis, NMF topic modeling, cultural keyword profiling, religious framing analysis, and English code-switching analysis. **Note on ADHD subgroup size and statistical power.** The ADHD subcorpus (n=**253**, **3.1**% of the total corpus) is substantially smaller than the BPD (n=**5,415**)

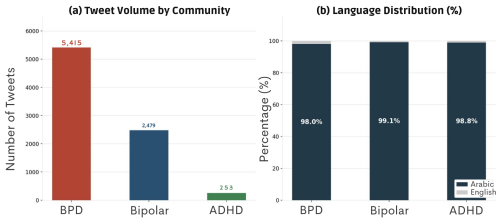

Figure 2: Tweet volume (left) and language distribution (right) per community. BPD dominates; all three communities are overwhelmingly Arabic-language.

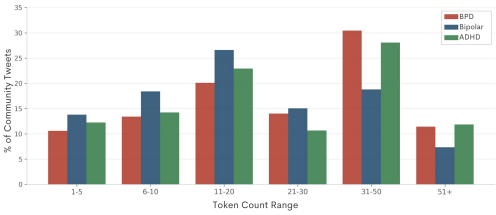

Figure 3: Normalized tweet length distributions. BPD and ADHD peak at 31–50 tokens; Bipolar at 11–20 tokens.

and Bipolar (n=**2,479**) subcorpora. As a result, ADHD-specific findings should be interpreted as preliminary and subject to greater uncertainty due to reduced statistical power. We retain ADHD in the main analysis for descriptive completeness, while treating ADHD-specific results as low-confidence.

4.1 Corpus Characterization

Community composition and language use. Following the exclusion of users without detected personal-disclosure signals under the GPT-4.1 annotation framework, the filtered corpus comprises BPD (n = **5,415**; **66.5**%), Bipolar (n = **2,479**; **30.4**%), and ADHD (n = **253**; **3.1**%). Discourse across all three communities is overwhelmingly Arabic-language (98.1–99.2%; Figure 2), distinguishing the corpus from predominantly English-centric social media mental health datasets and highlighting its value for under-resourced Arabic NLP research.

Tweet length distribution. All three communities exhibit right-skewed token-length distributions (Figure 3). Median tweet lengths are 25 tokens for BPD, 16 for Bipolar, and 21 for ADHD. BPD and ADHD discourse peaks within the 31 to 50 token range, whereas Bipolar discourse peaks earlier, within the 11 to 20 token range, consistent with shorter and more conversational interaction patterns. The pronounced long-tail distribution observed in BPD further suggests a higher prevalence of extended self-expression and help-seeking narratives.

4.2 Temporal and Behavioral Patterns

Circadian activity. Figure 4(a) displays hourly tweet volume normalized within each community to account for corpus size differences. Because the corpus is not geolocated, we report hours in UTC and provide Gulf Standard Time (GST) conversions only as contextual approximations. BPD and Bipolar share a broadly similar circadian profile: activity troughs in the early morning UTC window (02:00 to 06:00 UTC; approximately 05:00 to 09:00 GST) and peaks in the afternoon/evening UTC window (BPD: 18:00 UTC; Bipolar: 21:00 UTC). ADHD differs from this pattern, exhibiting a midday UTC peak (11:00 UTC) with only 20.2% of tweets falling in the 17:00 to 21:00 UTC window compared to 27.8% for BPD and 33.1% for Bipolar. All three communities share an early morning UTC trough (02:00 to 06:00 UTC). BPD exhibits a gradual rise toward evening with moderate daytime activity; ADHD shows a midday-concentrated profile with greater hour-to-hour variability. Because no user-location validation is available, these results should be interpreted as platform-time activity patterns rather than direct evidence of Gulf-population temporal norms. Day-of-week peaks differ by community (Figure 4(b)): BPD on Monday, Bipolar on Sunday, ADHD on Friday.

Longitudinal collection profile. Appendix Figure 9 shows monthly tweet volume across the 2022 to 2026 collection window. The Bipolar community data span the full period; however, meaningful volume only emerges from 2024 onward (12 tweets before 2024 vs. 2,467 from 2024 to 2026), reflecting the expanding reach of the data collection pipeline rather than a gradual organic growth process. The BPD community is heavily concentrated in the 2025 collection window (February to October 2025), accounting for 83.5% of its total volume in that period, with an additional 7.5% from early 2026. ADHD data cover a narrower window (April 2024 to December 2025), with the majority of tweets from 2025 (n=211, 83.4%), and are sparser overall, consistent with the smaller community sizes. The BPD temporal concentration introduces a confound: all BPD discourse patterns reported in this paper, vocabulary, cultural keyword rates, circadian patterns, and code-switching rates, are derived almost entirely from a single nine-month window and may reflect period-specific discourse rather than stable community characteristics. No date-stratified robustness check was performed; whether BPD findings replicate across different collection windows

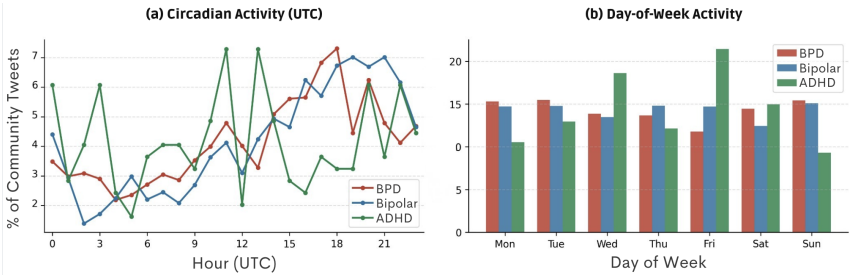

Figure 4: (a) Circadian (UTC) and (b) day-of-week tweet activity. BPD and Bipolar peak in the evening; ADHD at midday.

is unknown.

4.3 Linguistic and Cultural Framing

Community distinctive vocabulary via weighted log odds. To identify vocabulary statistically characteristic of each community, we apply the weighted log odds ratio with an informative Dirichlet prior (Monroe et al., 2008), a variance-normalized formulation that prevents rare words from dominating rankings through chance (full specification in Appendix E.1). Text is preprocessed by removing `[USER]` and `[URL]` placeholders, stripping non-Arabic characters, and filtering a standard Arabic stopword list. Words appearing fewer than five times in the target community are excluded.

Figure 7 shows the top-10 distinctive terms per community ranked by standardized weighted log-odds (z) score, where higher positive z values indicate words that are disproportionately associated with a given community relative to the remaining corpus. **The BPD community** is most strongly characterized by diagnostic terminology, including الحدية (*al-ḥadiyya*, "borderline"; z=4.38) and الحدي (*al-ḥaddi*, "borderline"; z=3.76), alongside identity vocabulary such as الشخصية (*al-shakhṣiyya*, "personality"; z=4.10), relational terms including العلاقات (*al-ʿalāqāt*, "relationships"; z=3.70), and affective vocabulary such as المشاعر (*al-mashāʿir*, "feelings"; z=3.68) and الحب (*al-ḥubb*, "love"; z=3.44). This pattern lexically overlaps with clinical constructs often discussed in relation to BPD, including interpersonal relationships, identity, and emotional regulation; however, lexical evidence alone cannot establish clinical features in users. As 83.5% of BPD tweets originate from a single nine-month collection window (see Section 4.2), the temporal stability of this profile cannot be verified. The Bipolar community yields the highest absolute z scores across all communities, led by القطب (*al-quṭb*, "pole"; z=10.30) and ثنائي (*thunāʾī*, "bi"; z=9.94), which together form the expression *thunāʾī al-quṭb* ("bipolar"). Additional high-scoring terms include الله (*Allāh*, "God/Allah"; z=9.79), الاكتئاب (*al-iktiʾāb*, "depression"; z=7.83), الهوس (*al-hawas*, "mania"; z=7.72), هوس (*hawas*, "mania"; z=7.32), نوبه (*nawba*, "episode"; z=6.14), and نوبة (*nawba*, "episode"; z=6.00). Notably, الله (*Allāh*, "God/Allah") ranks third among the distinctive terms, aligning with the elevated religious keyword frequencies reported below. Collectively, the high z scores suggest that discourse within the Bipolar community is characterized by a combination of condition-related, episodic, and religious vocabulary. **The ADHD community** shows a tighter z score range (1.88 to 3.10), consistent with its smaller corpus reducing statistical power. The distinctive vocabulary is nonetheless semantically coherent: كونسيرتا (*Kūnsīrtā*, "Concerta"; methylphenidate; z=3.10), الذكاء (*al-dhakāʾ*, "intelligence"; z=2.92), الحركة (*al-ḥaraka*, "movement/hyperactivity"; z=2.90), فرط (*farṭ*, "excess/hyper-"; z=2.87), and الانتباه (*al-intibāh*, "attention"; z=2.73).

Latent topic structure via NMF. To characterize corpus-level lexical themes, we fit a non-negative matrix factorization (NMF; Lee and Seung, 1999) topic model to TF–IDF representations of the self-disclosure-filtered tweets. The NMF pipeline removed very short reply fragments (<30 characters), stripped placeholders, normalized Arabic orthography, removed Arabic stopwords and community label terms, and used unigram TF–IDF features (min_df=8, max_df=0.70, max_features=5,000). This left 7,192 tweets for the topic model, so the NMF analysis complements the full-corpus lexical and cultural-keyword analyses rather than replacing them. We selected the number of topics by searching $k = 5$–14 using C_v coherence; $k = 12$ yielded the highest score ($C_v = 0.5013$), although the improvement over $k = 5$ ($C_v = 0.4987$) was small. The resulting topics and top terms are listed in Appendix E (Table 6). The NMF results provide a view of community-level lexical structure.

BPD shows the largest concentration in a feelings and relational-pain topic (Topic 1; 36.58% of BPD tweets, versus 16.14% of ADHD and 14.01% of Bipolar), with additional BPD-weighted topics involving social support/advice (Topic 7; 6.63%) and person/condition attribution (Topic 8; 7.77%). Bipolar is comparatively elevated on an episode-vocabulary topic centered on mania and depression episodes (Topic 4; 14.48%, versus 3.49% for BPD and 3.59% for ADHD), a treatment and medication topic (Topic 10; 22.33%), and a gratitude/supplication topic (Topic 6; 7.80%). ADHD, while still low-powered, is concentrated in first-person autobiographical framing (Topic 0; 19.73%) and treatment/medication management (Topic 10; 20.18%).

Cultural keyword framework. We apply a dictionary-based approach using six Arabic keyword lists spanning religious, medical, family/social, emotional distress, identity, and stigma domains (Table 8, Appendix H), derived through iterative corpus review anchored to Kleinman's 1980 explanatory models framework. The lists are an initial operationalization, not a validated instrument; rates reflect keyword prevalence, not direct measurements of latent constructs. All rates below are *raw occurrence counts* per 100 tweets; the Religious Framing paragraph uses a more restricted list with binary tweet-level hit rates, and the two metrics are not directly comparable. Figure 8 shows keyword rates across all six domains; three findings are noteworthy. First, the Bipolar community displays the highest religious keyword rate (41.3 raw occurrences per 100 tweets), nearly 2.5 times that of BPD (16.7) and 2.1 times that of ADHD (19.8). This pattern is compatible with prior accounts of faith-based coping and meaning-making in Arabic mental health contexts (Eid et al., 2025; Zaghouani et al., 2026), though keyword counts alone cannot establish whether religiosity functions as coping, causation, routine pragmatic expression, or broader cultural discourse. Importantly, this finding should be read alongside the pipeline's documented under-sensitivity to indirect and metaphorical disclosure in the Bipolar community (the community with the lowest GPT-4.1 agreement against human gold, $\kappa \approx 0.49$; Section 3.4). The reported rate of 41.3 may therefore underestimate religiously inflected disclosure, but the exact magnitude of any bias is unknown. Second, the Bipolar community also leads in medical keyword use (28.6 per 100 tweets), exceeding ADHD (24.1) and substantially exceeding BPD (11.7). To assess whether co-occurrence of religious and medical vocabulary reflects individual-level pluralism rather than two separate user populations, we computed the tweet-level intersection: **10.3%** of Bipolar tweets (256 of 2,479) contain at least one keyword from both the religious and medical domains simultaneously, compared to 3.0% for BPD (164 of 5,415) and 6.7% for ADHD (17 of 253; this cell is too sparse for meaningful interpretation and is reported for completeness only). The Bipolar–BPD difference is statistically reliable ($\chi^2(1) = 178.4$, $p < 10^{-40}$), though we note this test addresses only the reliability of the contrast, not whether the keywords capture the constructs they are intended to measure. This tweet-level co-occurrence provides stronger, though still exploratory, evidence compatible with explanatory model pluralism (Kleinman, 1980) at the level of individual posts. Third, BPD exhibits the highest identity keyword rate (35.6 per 100 tweets) and emotional distress keyword rate (29.6), suggesting stronger lexical emphasis on selfhood and distress in this corpus. The ADHD community shows the lowest emotional distress keyword rate (12.3), suggesting a comparatively more practical or symptom-management-oriented discourse. All ADHD rates should be interpreted with lower confidence given the small subcorpus (see the ADHD statistical power note, Section 4). Family/Social and Stigma keyword rates are shown in Figure 8 for completeness; cross-community differences on these domains are smaller and are not among the three strongest signals in this corpus.

Religious Framing. To examine religious language in greater depth, we applied a restricted religious keyword dictionary and organized matches into a four-tier exploratory taxonomy grounded in framing theory (Goffman, 1974) and interpreted through Kleinman's explanatory-model framework (Kleinman, 1980). Unlike the broader religious inventory reported in Appendix H, which includes all religiously associated terms (including highly polysemous or pragmatically ambient expressions), the restricted analysis retains only terms that function as relatively unambiguous religious markers in mental health contexts, such as explicit supplication, Quranic references, and supernatural-causation vocabulary. This analysis uses a binary tweet-level metric in which each tweet contributes at most one match per tier. The taxonomy itself is researcher-defined and should therefore be interpreted as an exploratory analytical framework rather than a validated annotation scheme. Across the corpus, 15.3%

of tweets contain at least one keyword from the restricted analysis, with the highest binary hit rate observed in the Bipolar community (24.9 tweets per 100), followed by ADHD (13.8) and BPD (11.0). The majority of religious tweets fall within *(1) Ambient Expression* (n = 1,052; 84.2% of religious tweets), comprising culturally conventional expressions such as الحمد لله (*al-ḥamdu lillāh*, "praise be to God") and إن شاء الله (*in shāʾ Allāh*, "God willing"), which frequently function as pragmatic discourse markers rather than illness-specific theological claims. A smaller category, *(2) Coping & Practice* (n = 104; 9.1%), includes references to prayer and Quranic recitation, such as الصلاة (*al-ṣalāh*, "prayer") and القرآن (*al-Qurʾān*, "the Quran"), which may reflect faith-based coping practices or routine religious observance. More explicit moral and supernatural framing appears in *(3) Guilt and Supernatural* (n = 61; 4.9%), which includes terms such as ذنب (*dhanb*, "sin"), عقاب (*ʿiqāb*, "punishment"), and الشيطان (*al-shayṭān*, "Satan"). At the same time, 29 tweets (0.4%) contain apparent counter-narratives, including expressions such as مو ذنبك (*mū dhanbak*, "it is not your fault"), which explicitly reject blame-based attributions. The least frequent category, *(4) Illness Causation Attribution* (n = 32; 2.6%), includes tweets that appear to frame mental health conditions in terms of divine trial, fate, or spiritual causation, including references such as ابتلاء (*ibtilāʾ*, "divine trial"), القدر (*al-qadar*, "fate/destiny"), and سببه روحي (*sababuhu rūḥī*, "its cause is spiritual"). Taken together, the co-occurrence of medical and religious vocabulary within Bipolar discourse is compatible with explanatory-model pluralism; however, the present evidence remains dictionary-level and cannot establish users' underlying causal beliefs or treatment preferences. Overall, religious language in the corpus appears functionally heterogeneous: most instances consist of culturally ambient expressions, a smaller subset may reflect coping practices, and only a limited proportion encode explicit causal interpretations of mental health conditions (see Appendix G).

Code-switching. To measure English code-switching, we removed `[USER]` and `[URL]` placeholder tokens from each tweet, then identified tweets containing at least one content-bearing English word, defined as a Latin-script token of two or more characters after excluding a standard English stopword list. Under this operationalization, **343** of **5,415** BPD tweets (**6.3**%), **80** of **2,479** Bipolar tweets (**3.2**%), and **72** of **253** ADHD tweets (**28.5**%, 95% CI: 22.9 to 34.1%, binomial) contain content-bearing English. The ADHD rate (28.5%) is strikingly higher than BPD (6.3%) and Bipolar (3.2%). BPD English is dominated by diagnostic and therapeutic terminology: *BPD* (n=79), *DBT* (32), *splitting* (20). Bipolar English is sparse, centering on condition labels (*depression*, *BD*). ADHD English is strongly anchored to the condition label itself (*ADHD*, n=54) alongside neurodiversity-specific vocabulary (*mindfulness*, *Russell Barkley*). One plausible explanation is that the English acronym *ADHD* functions as a compact, globally recognizable shorthand in online discourse, despite the availability of Arabic terminology for the condition (Alkhateeb and Alhadidi, 2019; Alqahtani et al., 2025); this interpretation aligns with prior work on Arabic–English code-switching and bilingual lexical choice (Alamri, 2022; Myers-Scotton, 1997), though user-level sociolinguistic factors cannot be ruled out.

5 Conclusion

We presented an exploratory computational characterization of Arabic mental health discourse across multiple condition-specific X Communities. Using a GPT-4.1-assisted personal-disclosure pipeline, we constructed a self-disclosure-filtered corpus of **8,147** tweets from **607** users and analyzed corpus composition, temporal activity, lexical distinctiveness, topic structure, code-switching, and cultural keyword prevalence. The results suggest community-associated patterns that extend beyond diagnostic vocabulary. In this corpus, Bipolar tweets show co-occurring religious and medical vocabulary, a pattern compatible with explanatory-model pluralism but insufficient to establish users' causal beliefs. Notably, **10.3**% of Bipolar tweets contain keywords from both domains, providing a tweet-level signal more consistent with individual explanatory pluralism than aggregate community-level rates alone. BPD tweets foreground relational, identity, and emotional-distress vocabulary, whereas ADHD tweets more often center on practical symptom and medication management. These findings should therefore be interpreted as hypotheses for future work rather than confirmed condition-level properties. Future work should expand the ADHD corpus, validate the keyword and religious-framing schemes through inter-rater annotation, and extend this interpretable approach to additional conditions and Arabic dialect regions.

Limitations

Pipeline. The prompt was developed on a Saudi-centric dataset, so adaptation may be needed for other Arabic dialects and regions. Inter-model agreement ($\kappa = 0.84$) overstates reliability relative to human-grounded estimates ($\kappa_{\text{GPT-human}} = 0.631$; $\kappa_{\text{Qwen-human}} = 0.329$). The conservative NEGATIVE default, bio override rule, and any-positive user aggregation may affect recall and precision, especially for indirect disclosures and highly active users.

Corpus. BPD data are temporally concentrated, with 83.5% of tweets from a single nine-month window, and the ADHD subcorpus is small (n=253). The corpus is also not geolocated, and no dedicated bot-detection procedure was applied; temporal and community-level findings should therefore be interpreted cautiously.

Analyses. The cultural keyword framework and religious-framing taxonomy are exploratory rather than validated instruments. Keyword rates and the χ^2 co-occurrence test support descriptive contrasts, but not construct validity. Future work should add user-level validation, dialect-sensitive keyword validation, robustness checks, and sensitivity analyses excluding GPT-positive/Qwen-negative users.

Ethical Considerations

This study analyzes publicly visible posts from X Communities and does not involve direct contact with users, recruitment, or intervention. Because the data concern sensitive mental-health discourse, we treat the study as discourse-level analysis rather than individual-level assessment. User identifiers were removed from the analytic dataset prior to analysis, raw post text is not redistributed, and results are reported only in aggregate.

Consistent with X content-redistribution restrictions, raw post text, usernames, bios, profile metadata, user IDs, and community labels are not redistributed. The public release provides only X Post IDs, which may be rehydrated by authorized users through the X API subject to X's applicable Developer Agreement, Developer Policy, access limits, and any required approvals. Because Post IDs can be used to retrieve original posts when they remain available, the released data should not be considered fully anonymized.

The `likely_disclosure` label is an operational descriptor based on self-reported experiential language and does not constitute a clinical diagnosis; no clinical inferences should be drawn from pipeline outputs. The LLM-assisted annotation pipeline may under-detect indirect, figurative, or culturally specific forms of disclosure, particularly where religious or metaphorical language is prevalent. The cultural keyword framework and religious-framing taxonomy are exploratory instruments that have not been validated across Arabic dialect regions. Findings should not be used to characterize, profile, screen, or intervene on individual users, nor should aggregated discourse patterns be generalized to Arabic-speaking populations or used to estimate the prevalence or nature of mental-health conditions in Arab societies.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Horeya AbouWarda, Mateusz Dolata, and Gerhard Schwabe. 2024. How does an online mental health community on Twitter empower diverse population levels and groups? a qualitative analysis of #BipolarClub. *Journal of Medical Internet Research*, 26:e55965.

Norah Mohammed Alamri. 2022. Arabic-English code-switching among KKU students on social media Twitter. M.A. thesis, English Department, Faculty of Languages and Translation, King Khalid University; hosted by Arab World English Journal, ID No. 284.

Jamal M Alkhateeb and Muna S Alhadidi. 2019. ADHD research in Arab countries: A systematic review of literature. *Journal of attention disorders*, 23(13):1531–1545.

Mohammed M. J. Alqahtani, Nouf Mohammed Al Saud, Nawal Mohammed Alsharef, Ahmad N. AlHadi, Saleh Mohammed Alsalhi, Elham H. Al-Hifthy, Yasser Ad-Dab'bagh, Nader Alrahili, Fawwaz Abdulrazaq Alenazi, Barakat M. Alotaibi, Sultan Mahmoud Alsaeed, Boshra A. Arnout, Latifah ALQasem, Abdulkarim Alhossein, Yasser Jubran Alqahtani, Samirah A. AlGhamdi, Jeremy Varnham, Saeed Abdulwahab Asiri, and Maysaa W. Buraik. 2025. Standardization of the Arabic version of the adult ADHD self-report screening scale for DSM-5 (ASRS-5) among adults in Saudi Arabia: Variability of ADHD screening according to sociodemographic variables. *Journal of Attention Disorders*, 29(6):445–457.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic lan-

guage understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resources Association.

Lama Ayash, Ashwag Alasmari, and Hassan Alhuzali. 2025. ContextMentalQA: Modeling cultural, social, and religious context in Arabic mental health question answering. TechRxiv Preprint.

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, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 1–10.

Latefa A. Dardas and Leigh Ann Simmons. 2015. The stigma of mental illness in Arab families: A concept analysis. *Journal of Psychiatric and Mental Health Nursing*, 22(9):668–679.

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.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Mario Eid, Venise Abi Kheir, Maya Bizri, Amine Larnaout, and Samer El Hayek. 2025. Somatic symptom and related disorders in the Arab world: a narrative review of clinical features and care implications. *Frontiers in Psychiatry*, 16:1692267.

Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–16.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Erving Goffman. 1959. *The Presentation of Self in Everyday Life*. Doubleday, Garden City, New York.

Erving Goffman. 1974. *Frame Analysis: An Essay on the Organization of Experience*. Harper & Row, New York.

Arthur Kleinman. 1980. *Patients and healers in the context of culture: An exploration of the borderland between anthropology, medicine, and psychiatry*, volume 3. Univ of California Press.

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

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. Computational social science. *Science*, 323(5915):721–723.

Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Carol Myers-Scotton. 1997. *Duelling Languages: Grammatical Structure in Codeswitching*. Oxford University Press, Oxford. Reprint/edition; originally published in 1993.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077, Singapore. Association for Computational Linguistics.

Wajdi Zaghouani, Eman Sedqy Shlkamy, and Mabrouka Bessghaier. 2026. From posts to pressure: An Arabic dataset about stress and mental-health monitoring. In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script*, pages 422–432, Rabat, Morocco. Association for Computational Linguistics.

Monica Zolezzi, Maha Alamri, Shahd Shaar, and Daniel Rainkie. 2018. Stigma associated with mental illness and its treatment in the Arab culture: A systematic review. *International Journal of Social Psychiatry*, 64(6):597–609.

A Personal Disclosure Classification Prompt

This appendix presents the complete prompt used for LLM assisted personal disclosure classification of tweets from Arabic mental health communities. The prompt was run with GPT-4.1 as the primary annotator and Qwen3-235B-A22B as a

Component	Function
System Role	Establishes domain expertise and task framing
Input Format	JSON object with community, bio, and tweet
Label Definitions	Positive/Negative criteria
Bio Override Rule	Bio driven Positive override and professional bio handling
Reason Tag Taxonomy	13 tags covering disclosure signals, non-disclosure signals, bio signals, and edge cases
Confidence Levels	Three level confidence scale
Default Rule	Conservative Negative default
Few Shot Examples	6 calibration exemplars (fully synthetic)

Table 1: Components of the personal disclosure classification prompt.

screening model in parallel: GPT-4.1 (`GPT-4.1`, temperature=0.0, max_tokens=250) and Qwen3-235B-A22B (`Qwen3-235B-A22B instruct 2507`, temperature=0.0, max_tokens=250). Each model independently classifies each tweet as POSITIVE (the tweet contains personal-disclosure signals: the author appears to be personally living with or experiencing a mental health condition) or NEGATIVE (no personal-disclosure signals), together with a confidence level and a set of reason tags. Classification operates at the tweet-level; the user bio is provided as supporting context. Table 1 summarizes the components.

A.1 System Role and Task Definition

SYSTEM ROLE
You are an expert classifier working with Arabic-language social media data collected from focused mental health communities on X (Twitter). These communities focus on ADHD, Bipolar Disorder, and BPD (Borderline Personality Disorder).

TASK
Your task is to classify a single tweet as either:
- `positive`: the tweet contains personal-disclosure signals: the author appears to be personally living with or experiencing a mental health condition
- `negative`: the tweet contains no personal-disclosure signals: the content is educational, professional, neutral, or irrelevant

You are classifying the tweet only. User level decisions are handled separately.

A.2 Input Format and Bio Override Rule

INPUT FORMAT

```
{
  "community": "ADHD | BPD | bipolar",
  "user_bio": "...",
  "tweet_text": "..."
}
```

Use BOTH the bio and the tweet together to make your decision.

BIO OVERRIDE RULE
If the bio clearly identifies the user as living with a condition, explicit diagnosis, living with language, or condition as personal identity (e.g., أعيش تم تشخيصي بـ $ADHD$ مع ثنائي القطب #BPD, "bipolar girl"), then:
- Set `tweet_label` to `positive`
- Add `BIO_SELF_IDENTIFICATION` to reason_tags
- This applies even if the tweet itself contains no disclosure signal

If the bio indicates a professional or institutional account, this supports `negative`, but does NOT override a clearly personal tweet.
If the bio is empty, rely on the tweet alone and add `EDGE_EMPTY_BIO` to reason_tags.

A.3 Classification Rules

A.3.1 Classify as Positive

Classify as **positive** if ANY of the following are true:
- First person account of experiencing symptoms (e.g., ما أقدر أنام نفسيتي مدمرة أعاني من تشتت)
- Disclosing a personal diagnosis (e.g., تم تشخيصي بـ $ADHD$ عندي ثنائي القطب)
- Sharing personal emotional distress or struggles related to a mental health condition
- Seeking peer support or venting about daily life with a condition
- Asking whether a personally experienced symptom belongs to a condition (first person question)
- Expressing experience from the inside, not explaining or educating others about a condition

A.3.2 Classify as Negative

Classify as **negative** if ANY of the following are true:
- Educational or psychoeducational content explaining symptoms or treatments in third person
- Written as a professional advising or answering someone else's question
- Promoting a therapy session, app, webinar, course, book, or product
- Discussing research findings, clinical definitions, or diagnostic criteria
- Addressing community members as a separate audience (e.g., هؤلاء الأشخاص يحتاجون...)
- Spam, off topic, or completely irrelevant content

Default Rule: When in doubt, label **negative**.

A.4 Output Format, Reason Tag Taxonomy, and Confidence Levels

OUTPUT FORMAT
Return ONLY a valid JSON object with no extra text, no markdown fences:

```
{
  "tweet_label": "positive | negative",
  "confidence": "high | medium | low",
  "reason_tags": ["TAG_1", "TAG_2"]
}
```

REASON TAG TAXONOMY

Disclosure signal tags (support `positive`):
- `TWEET_SYMPTOM_DISCLOSURE`: Tweet describes experiencing symptoms in first person
- `TWEET_PERSONAL_DIAGNOSIS_DISCLOSURE`: Tweet explicitly states the user was diagnosed
- `TWEET_PEER_SUPPORT_SEEKING`: Tweet seeks support or validation from others with the condition
- `TWEET_EMOTIONAL_VENTING`: Tweet expresses raw personal emotion or distress without educational intent
- `TWEET_FIRST_PERSON_SYMPTOM_QUESTION`: Tweet asks whether a personally experienced symptom belongs to a condition

Non-disclosure signal tags (support `negative`):
- `TWEET_EDUCATIONAL_CONTENT`: Tweet explains symptoms, conditions, or treatment in informational third person style
- `TWEET_PROFESSIONAL_ADVICE`: Tweet offers clinical guidance or answers someone else's question professionally
- `TWEET_SERVICE_PROMOTION`: Tweet promotes a therapy session, app, webinar, course, or mental health product
- `TWEET_RESEARCH_OR_CLINICAL`: Tweet discusses diagnostic criteria, research findings, or clinical definitions
- `TWEET_THIRD_PERSON_FRAMING`: Tweet addresses community members as a separate audience
- `TWEET_SPAM_OR_IRRELEVANT`: Tweet is off topic, spam, or unrelated to mental health

Bio signal tags (always added when detected, regardless of tweet label):
- `BIO_SELF_IDENTIFICATION`: Bio clearly identifies the user as personally living with or diagnosed with a condition

Edge case tags:
- `EDGE_EMPTY_BIO`: Bio is absent; classification relies entirely on tweet content
- `EDGE_AMBIGUOUS_FIRST_PERSON`: Tweet could be personal or professional; classified based on best available signal
- `EDGE_PROFESSIONAL_BIO_PERSONAL_TWEET`: Bio suggests a professional but tweet content is clearly personal/experiential

Dimension	Coverage (6 examples)
Label	Positive (3), Negative (3)
Confidence	High (4), Medium (2)
Bio signal tags	`BIO_SELF_IDENTIFICATION` (3)
Edge case tags	`EDGE_EMPTY_BIO` (1), `EDGE_AMBIGUOUS_FIRST_PERSON` (1), `EDGE_PROFESSIONAL_BIO_PERSONAL_TWEET` (1)

Table 2: Coverage of the six synthetic few-shot examples used in the prompt.

CONFIDENCE LEVELS
- `high`: Strong unambiguous signal (e.g., explicit diagnosis disclosure, clear third person educational content)
- `medium`: Signal present but indirect or requires inference
- `low`: Very weak or contradictory signals; classification is a best guess

A.5 Important Notes

- This dataset is primarily in Arabic (Modern Standard and Gulf/Saudi dialect). Be sensitive to dialectal expressions of distress (e.g., نفسيتي مدمرة مو زين قرفانة من كل شي).
- Do NOT base the classification solely on the community tag: professionals, researchers, and caregivers are present in all three communities.
- If the bio clearly identifies the user as living with a condition, always return `tweet_label` `positive` and add `BIO_SELF_IDENTIFICATION`: even if the tweet itself is educational or neutral.
- A professional bio does NOT override a clearly personal tweet: classify such cases as `positive` and add `EDGE_PROFESSIONAL_BIO_PERSONAL_TWEET`.
- This classification is for research purposes. Handle all data with care and do not make clinical inferences beyond the binary label requested.

A.6 Few Shot Examples

All examples are **fully synthetic**; no real user data is reproduced. We present six exemplars covering representative label, confidence, and tag combinations. Table 2 summarizes coverage.

Example 1: Positive, high confidence (emotional venting + self-identifying bio).

Input:
`community: "BPD"`
`user_bio:` إنسانة تتعلم كيف تعيش مع اضطراب الشخصية الحدية يومًا بيوم | #BPD
`tweet_text:` أصعب شي في الحدية إنك تحب بشكل كامل وفجأة تحس إن كل شي انهار بدون سبب واضح

Output: `{"tweet_label": "positive", "confidence": "high", "reason_tags": ["TWEET_EMOTIONAL_VENTING", "BIO_SELF_IDENTIFICATION"]}`

Example 2: Negative, high confidence (educational content + third person framing).

Input:
community: "ADHD"
user_bio: أخصائي نفسي إكلينيكي , ماجستير إرشاد نفسي , مرخص من هيئة التخصصات الصحية
tweet_text: الفرق بين فرط الحركة عند الأطفال والبالغين: الأطفال يُظهرون أعراضًا حركية واضحة، بينما يعاني البالغون من أعراض داخلية كالقلق الذهني وصعوبة التنظيم.

Output: {"tweet_label": "negative", "confidence": "high", "reason_tags": ["TWEET_EDUCATIONAL_CONTENT", "TWEET_THIRD_PERSON_FRAMING"]}

Example 3: Positive, high confidence (symptom disclosure + first person question, empty bio).

Input:
community: "bipolar"
user_bio: ""
tweet_text: أحس هالأيام بطاقة زايدة عن اللزوم، ما أنام، وعندي رغبة أشتري أشياء ما أحتاجها ! هل هذا طبيعي ولا ممكن يكون هوس؟

Output: {"tweet_label": "positive", "confidence": "high", "reason_tags": ["TWEET_SYMPTOM_DISCLOSURE", "TWEET_FIRST_PERSON_SYMPTOM_QUESTION", "EDGE_EMPTY_BIO"]}

Example 4: Negative, medium confidence (professional bio, ambiguous tweet).

Input:
community: "BPD"
user_bio: معالج نفسي معتمد , متخصص في اضطرابات الشخصية , باحث في العلاج الجدلي السلوكي DBT
tweet_text: أحيانًا الشفاء لا يبدو كالشفاء ! بل يبدو كلحظة هدوء صغيرة وسط العاصفة. أتمنى لكم تلك اللحظة

Output: {"tweet_label": "negative", "confidence": "medium", "reason_tags": ["TWEET_THIRD_PERSON_FRAMING", "EDGE_AMBIGUOUS_FIRST_PERSON"]}

Example 5: Positive, high confidence (educational tweet, bio self-identification override).

Input:
community: "ADHD"
user_bio: مبرمج ومهتم بالتقنية , تم تشخيصي بـ $ADHD$ منذ سنتين
tweet_text: الفرق بين فرط الحركة عند الأطفال والبالغين من وجهة نظر علمية

Output: {"tweet_label": "positive", "confidence": "high", "reason_tags": ["TWEET_EDUCATIONAL_CONTENT", "BIO_SELF_IDENTIFICATION"]}

Example 6: Positive, medium confidence (emotional venting + symptom disclosure, professional bio override).

Input:
community: "bipolar"
user_bio: طالبة دكتوراه علم نفس , أعيش مع ثنائي القطب وأحاول أفهمه من الداخل والخارج
tweet_text: لما تكون في نوبة اكتئاب وتعرف نظريًا كل الأدوات العلاجية بس ما تقدر تطبق ولو واحدة ! هذا تناقض ما يفهمه غير اللي عاشه

Output: {"tweet_label": "positive", "confidence": "medium", "reason_tags": ["TWEET_EMOTIONAL_VENTING", "TWEET_SYMPTOM_DISCLOSURE", "BIO_SELF_IDENTIFICATION", "EDGE_PROFESSIONAL_BIO_PERSONAL_TWEET"]}

A.7 User Level Aggregation

Tweet level classifications are aggregated into a single user-level label using a deterministic priority ordered procedure. Each user is assigned one of two labels: LIKELY_DISCLOSURE or OTHER. The rules are applied in order; the first matching rule determines the outcome.

1. **Bio override (AGG_BIO_DISCLOSURE_OVERRIDE).** If *any* tweet for the user carries BIO_SELF_IDENTIFICATION in its reason_tags, the user is immediately labeled LIKELY_DISCLOSURE, regardless of tweet-level labels. This rule fires because the bio override in the tweet-level prompt propagates the same tag to every tweet for that user; detecting it once is sufficient.
2. **Any positive tweet wins (AGG_ANY_POSITIVE_TWEET / AGG_CONFLICT_POSITIVE_WINS).** If any tweet-level label is positive, the user is labeled LIKELY_DISCLOSURE. The aggregation reason distinguishes two sub cases:
   - AGG_ANY_POSITIVE_TWEET: all tweets are positive (unanimous)
   - AGG_CONFLICT_POSITIVE_WINS: at least one tweet is positive but at least one is negative (conflict resolved in favor of positive)

   The triggering tweet IDs are recorded for traceability.
3. **All negative, no bio signal (AGG_ALL_NEGATIVE_NO_BIO_SIGNAL).** If no tweet is positive and no bio signal was detected, the user is labeled OTHER.

The design of Rules 1 and 2 prioritizes recall for personal-disclosure evidence; this choice may increase false positives and should be considered when interpreting the corpus. Rule 3 provides the default for users whose discourse is entirely non-

personal-disclosure.

B Prompt Design Rationale

B.1 Theoretical Framework

The prompt's core design decisions are grounded in three complementary frameworks. **Goffman (1959) presentation of self.** The bio is the user's "front stage" presentation; tweets represent "back stage" behavior. This justifies reading both together and allowing a self-identifying bio to override a neutral tweet, while a clearly personal tweet overrides a professional bio signal.

The explanatory models in Kleinman (1980). The distinction between personal experience of illness and third party or professional discourse about illness aligns with Kleinman's separation of illness experience from disease frameworks, motivating the Positive/Negative boundary and the tweet-level classification rules.

Code-switching theory (Myers-Scotton, 1997). The important notes section's explicit sensitivity to Gulf Arabic and Saudi dialect expressions of distress ensures that dialectally expressed disclosures are recognized regardless of linguistic form.

The two label scheme reflects a single analytically motivated boundary: whether the tweet contains personal-disclosure signals suggesting the author is personally living with or experiencing a mental health condition, versus any other stance (professional, educational, or irrelevant). The structured reason tag taxonomy, covering disclosure signal tags, non-disclosure signal tags, bio signal tags, and edge case tags, provides a multi dimensional audit trail that makes the basis of each classification transparent and supports systematic error analysis.

B.2 Conservative Default and Bio Override Rules

Two design choices are especially consequential. First, the **Negative default** ensures that ambiguous tweets, content that discusses mental health generally, provides educational information, or addresses community members in the third person, are not mistakenly counted as personal-disclosure discourse. This is analytically conservative: some genuine disclosure tweets may be lost, but the resulting corpus is less contaminated by non-disclosure content. Second, the **bio override** rule, a self-identifying bio triggers a Positive label regardless of tweet content, captures users who may post educational or neutral content on a given day while living with a condition. Conversely, a professional bio does not override a clearly personal tweet, preventing the systematic exclusion of clinicians or researchers who also share their own lived experience.

Statistic	**Value**
Total tweets (post preprocessing)	9,582
Total tweets (post removal)	**8,147**
Unique users (post preprocessing)	1,286
Unique users (post removal)	**607**
Communities	3 communities across 3 conditions (see Table 9)
Bio status ($\leq$ 2 words = MINIMAL)	
FULL ($\geq$ 3 words)	790 (61.4%)
MINIMAL ($\leq$ 2 words)	152 (11.8%)
EMPTY (null/blank)	344 (26.7%)
User label distribution (pipeline output)	
LIKELY_DISCLOSURE	**607** (47.2%)
OTHER	**679** (52.8%)

Table 3: Dataset statistics. EMPTY = null/blank bio; MINIMAL = $\leq$ 2 words; FULL = $\geq$ 3 words.

C Dataset Statistics

D Annotation Quality and Validation

D.1 Inter Model Agreement Protocol

GPT-4.1 and Qwen3-235B-A22B were run on the same prompt across all 9,582 preprocessed tweets. Both models received identical inputs (community, user_bio, tweet_text) and produced independent POSITIVE/NEGATIVE labels. Cohen's κ was computed across all tweet pairs where neither model returned a parse failure. GPT-4.1 returned 20 partial parses (0.2%) and zero full failures; Qwen3 returned 111 parse failures (1.2%). Raw agreement on valid pairs ($n = 9{,}471$): 90.8%; Cohen's $\kappa = 0.84$.

Interpretation. This inter-model κ should not be read as a standalone validity claim. As shown in the human validation study below, it overstates pipeline reliability because GPT-4.1 and Qwen3 share a conservative bias on easy cases while diverging sharply on ambiguous ones. Qwen3's primary function is to flag disagreements for review, not to serve as an independent validator.

D.2 Human Validation Study

D.2.1 Sample and annotators

Two native Arabic-speaking annotators independently labeled a stratified sample of 200 tweets, all of which had valid paired labels. The sample was stratified across five difficulty tiers: high confidence positive (*C_pos_high*, n=50), low/medium confidence positive (*B_pos_low_med*, n=30), high confidence negative (*E_neg_high*, n=30), low/medium confidence negative (*D_neg_low_med*, n=30), and inter-model conflict (*A_conflict*, n=60). Each tweet was labeled POSITIVE or NEGATIVE using the same guidelines as the LLM prompt.

D.2.2 Inter human agreement

The annotators agreed on 192 of 200 tweets (96.00%), with $\kappa = 0.905$ (almost perfect). All 8 disagreements were directionally consistent: one annotator applied a more liberal threshold on ambiguous positive cases. The 192 agreed tweets constitute the human gold standard used to evaluate the LLMs.

D.2.3 LLM performance against human gold

Table 4 summarizes GPT-4.1 and Qwen3 performance against the 192 agreed human gold labels. GPT-4.1 reaches substantial agreement ($\kappa = 0.631$; $F_1^+ = 0.88$); Qwen3 reaches only fair agreement ($\kappa = 0.329$; $F_1^+ = 0.72$), confirming the inter-model $\kappa = 0.84$ overstates pipeline reliability.

Comparison	κ	Agree.	F_1^+	Band
Inter human (ceiling)	0.905	96.00%	NA	Almost perfect
GPT-4.1 vs. human gold	0.631	83.85%	0.88	Substantial
Qwen3 vs. human gold	0.329	66.30%	0.72	Fair
GPT-4.1 vs. Qwen3	0.84	90.80%	NA	Almost perfect

Table 4: Human validation results (n=200). F_1^+: positive-class F_1. GPT-4.1 evaluated on 192 gold labels; Qwen3 on 184 (after parse failures). Inter-model row shown for reference only.

Stratum	GPT agree.	Qwen3 agree.
C_pos_high (n=50)	96%	96%
B_pos_low_med (n=30)	90%	90%
E_neg_high (n=28)	93%	93%
D_neg_low_med (n=26)	46%	46%
A_conflict	83% (n=58)	18% (n=50)

Table 5: Per-stratum agreement with human gold. n values reflect the subset of each stratum with mutual human agreement (192 total), after excluding the 8 human-disagreement tweets from the full 200-tweet sample. D_neg_low_med is the critical failure stratum; Qwen3 collapses on conflict cases (18%).

D.2.4 Stratum level breakdown

Table 5 shows per-stratum agreement. GPT-4.1 performs well on clear-label strata (C_pos_high: 96%; E_neg_high: 93%; B_pos_low_med: 90%) but drops to 46% on low/medium confidence negatives (D_neg_low_med), where the conservative NEGATIVE default suppresses genuine disclosures. Qwen3 collapses on inter-model conflict cases (18%), confirming it cannot serve as an independent annotator on ambiguous tweets.

D.2.5 Community breakdown (GPT-4.1 vs. human gold)

Agreement is highest for ADHD ($\kappa = 0.73$), followed by BPD ($\kappa = 0.66$), and lowest for Bipolar ($\kappa \approx 0.49$). Bipolar accounts for 10 of 21 GPT false negatives on the human gold set, consistent with that community's indirect, religious, and metaphorical disclosure style being most susceptible to the conservative NEGATIVE default.

E Supplementary Figures and Formulas

E.1 Log-Odds Formulation

For a target group i and comparison group j, the log-odds of word w under the weighted log-odds ratio with an informative Dirichlet prior (Monroe

et al., 2008) is:

$$\delta_w^{(i-j)} = \log \frac{y_w^i + \alpha_w}{n^i + \alpha_0 - y_w^i - \alpha_w} - \log \frac{y_w^j + \alpha_w}{n^j + \alpha_0 - y_w^j - \alpha_w} \quad (1)$$

where y_w^i is the count of word w in group i, n^i is the total word count of group i, and α_w is the prior count drawn from the pooled background corpus ($\alpha_0 = \sum_w \alpha_w$). The z score normalizes by variance:

$$\zeta_w^{(i-j)} = \frac{\delta_w^{(i-j)}}{\sqrt{\sigma_w^2}}, \quad \sigma_w^2 = \frac{1}{y_w^i + \alpha_w} + \frac{1}{y_w^j + \alpha_w} \quad (2)$$

The background prior α_w is set to the word's frequency in the full corpus; unseen words receive additive smoothing of 0.01.

E.2 NMF Topic Model

F Limitations and Dialect Extensibility

The classification prompt was developed and validated on a Saudi-centric dataset. The important notes section explicitly flags sensitivity to Gulf Arabic and Saudi dialect expressions of distress, but the few shot examples and the implicit discourse norms encoded in the classification rules skew Gulf Arabic. Table 7 identifies areas requiring adaptation for other Arabic dialect groups.

G Religious Framing Analysis

The religious analysis defines a four-tier taxonomy of language use: *Ambient Expression*, *Coping & Practice*, *Guilt & Supernatural*, and *Illness Causation Attribution*.

Tier Distribution and Within-Tier Sentiment Figure 5 shows that *Ambient Expression* dominates across communities, accounting for 8.8% (BPD) to 22.3% (Bipolar) of total tweet volume, with the Bipolar rate more than 2.5× BPD and roughly double ADHD (10.3%), reinforcing earlier evidence of elevated religious framing; this estimate is likely conservative due to under-detection in Bipolar discourse. The remaining tiers each account for under 3% of tweets but are analytically important: *Coping & Practice* peaks in ADHD (2.8%), reflecting symptom-practice interplay, while *Guilt & Supernatural* and *Illness Causation Attribution* are most concentrated in Bipolar (0.7% each), consistent with explanatory-model pluralism. Sentiment patterns differ by tier: *Ambient Expression* is near-balanced (51% positive, 49% negative), *Coping & Practice* skews negative (57%), *Illness Causation Attribution* shows the strongest positive skew (56%), and *Guilt & Supernatural* remains near-balanced, indicating that morally and supernaturally framed language does not straightforwardly align with negative affect.

Top Religious Keywords per Community Figure 6 compares the most frequent religious keywords across BPD, Bipolar, and ADHD communities. Across all three, الله ("God/Allah") is the most frequent term, followed by والله ("by God") and يارب ("O Lord"), indicating a shared reliance on core religious expressions in mental health discourse. The Bipolar and BPD communities show similar patterns, with frequent use of الحمدلله and إن شاء الله. The Bipolar community additionally exhibits higher use of more formal supplicatory expressions such as بسم الله and آمين اللهم. The BPD community is distinct in its use of ما شاء الله, often occurring in peer-support contexts. In contrast, the ADHD community shows lower overall frequencies but includes more references to prayer-related terms such as أصلي and الصلاة, suggesting greater emphasis on ritual practice and routine. Overall, while all communities share common religious expressions, their usage differs in function across coping, peer interaction, and practice-oriented framing.

H Cultural Keyword Framework

Table 8 lists the six cultural keyword domains and representative Arabic terms used in the dictionary-based analysis (Section 4.3). Each domain was derived through iterative corpus vocabulary review anchored to Kleinman's 1980 explanatory models framework. The lists are an initial operationalization and have not yet been validated through an independent inter-rater annotation study.

I X Community Overview

Table 9 summarizes the three X Communities used for corpus construction, including total membership size and moderation type at the time of data collection.

ID	Human label	Top terms	n	BPD %	Bipolar %	ADHD %
0	First-person narrative	اني، كنت، اقدر، اعرف، وانا، اقول، حياتي، لاني، ابدا، اكون، افكر، وما	522	6.86	6.86	19.73
1	Feelings and relational pain	بل، المشاعر، الشخص، شعور، الحدي، احيانا، ليس، الشعور، دون، الالم، الحب، الخوف	2105	36.58	14.01	16.14
2	Questions and peer validation	هل، ام، سوال، طبيعي، علاقه، الحديين، عليكم، ليه، بالحديه، وهل، السوال، تحسون	361	4.44	6.30	5.38
3	Clinical encounters and causes	لي، بالنسبه، سبب، وانا، صار، الدكتور، يسبب، يوم، قالت، شهور، فيني، تقريبا	593	8.03	8.84	7.17
4	Bipolar episode vocabulary	نوبه، الهوس، الاكتياب، هوس، اكتياب، وقت، نوبات، النوبه، مختلطه، طبيعي، تجيني، قلبي	485	3.49	14.48	3.59
5	Self-struggle/help seeking	نفسي، يوم، احاول، اكره، لاني، اكون، الناس، اخصايي، تعبت، ليش، احتاج، اسوي	604	8.78	7.33	10.31
6	Gratitude and supplication	شكرا، يارب، اللهم، امين، خير، فعلا، عليك، كلامك، شاء، يسعدك، ربي، العافيه	306	2.79	7.80	2.24
7	Advice and social support	نفسك، عليك، الا، محد، راح، اهم، معك، نصيحه، وانت، حاول، حياتك، منك	407	6.63	3.62	4.04
8	Person/condition attribution	شخص، حدي، طبيعي، الشخص، راح، عنده، يعاني، معي، لانه، مصاب، واحد، نفسه	471	7.77	3.95	4.93
9	Uncertainty and diffuse distress	احس، مدري، ناس، حولي، تعبت، الشي، دايما، ذا، ليش، ليه، مشاعري، الناس	306	4.58	3.57	3.59
10	Treatment and medication	الادويه، العلاج، السلوكي، علاج، بعض، الجدلي، النفسي، ادويه، الحمدلله، اخذ، النفسيه، الدكتور	927	8.41	22.33	20.18

Table 6: NMF topic labels, top terms, topic sizes, and within-community percentages. The model was selected by C_v coherence over $k = 5$–14; $k = 12$ produced the highest coherence ($C_v = 0.5013$). Labels are researcher-assigned and exploratory.

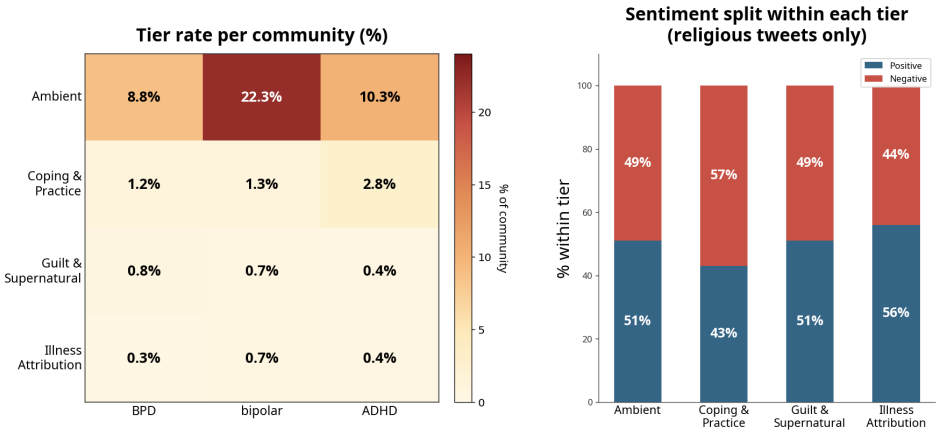

Figure 5: Religious framing tier rates per community (left) and sentiment split within each tier (right). Cell values = % of community tweets containing a tier keyword; sentiment computed over religious-language tweets only.

Dialect	Required Adaptations
Egyptian	Indirect or figurative disclosure patterns (e.g., زهقت من نفسي); colloquial symptom vocabulary
Levantine	Reflexive expressions for personal suffering (حالتي/حالي); different medication brand names
N. African	French Arabic code-switching in clinical and symptom vocabulary
Cross dialect	
Indirect disclosure	Users who describe experience metaphorically or religiously without explicit diagnosis language may be under detected by the explicit disclosure criterion
Condition labels	ADHD has multiple Arabic renderings and frequent English-acronym use in online discourse; other conditions may have competing translations

Table 7: Dialect specific and cross dialect adaptations needed beyond Gulf Arabic / MSA.

Domain	Example Keywords (Arabic)
Religious	الله، ربي، يارب، صبر، الشفاء، شاء
Medical	دواء، دكتور، تشخيص، مستشفى، جلسة
Family/Social	أهل، ماما، بابا، زوج، أصدقاء، مجتمع
Emot. Distress	حزن، خوف، قلق، تعب، وحيد، اكتئاب
Identity	أنا، نفسي، شخصيتي، ذاتي، إحساسي
Stigma	مجنون، خبل، عيب، خجل، انكار

Table 8: Cultural keyword domains and representative Arabic terms.

Condition	# Members	Moderation
BPD	5,876	Professional
ADHD	2,321	Varies
Bipolar	1,320	Peer led
Total	**9,517**	

Table 9: X Communities used for corpus construction. *Professional*: licensed psychologists; *Varies*: mixed moderation; *Peer-led*: lived-experience individuals or relatives.

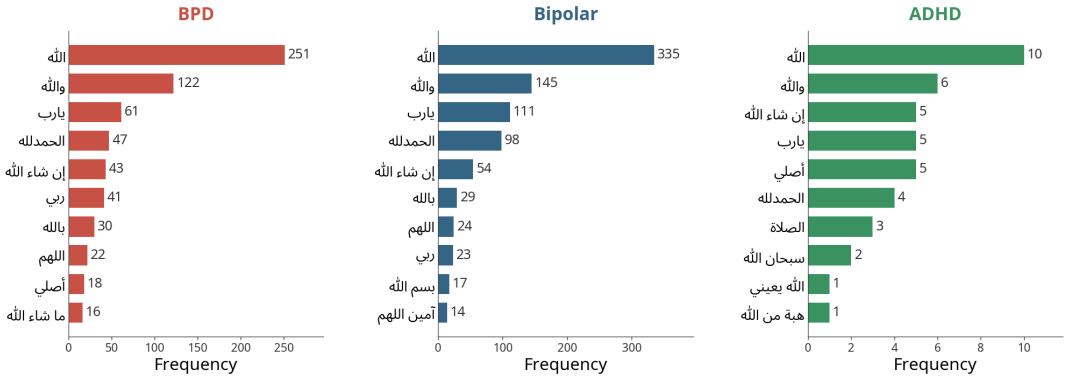

Figure 6: Top-10 religious keywords per community (ranked by raw count). Note the scale difference: BPD and Bipolar counts reach hundreds; ADHD reaches a maximum of 10 (n=253 tweets).

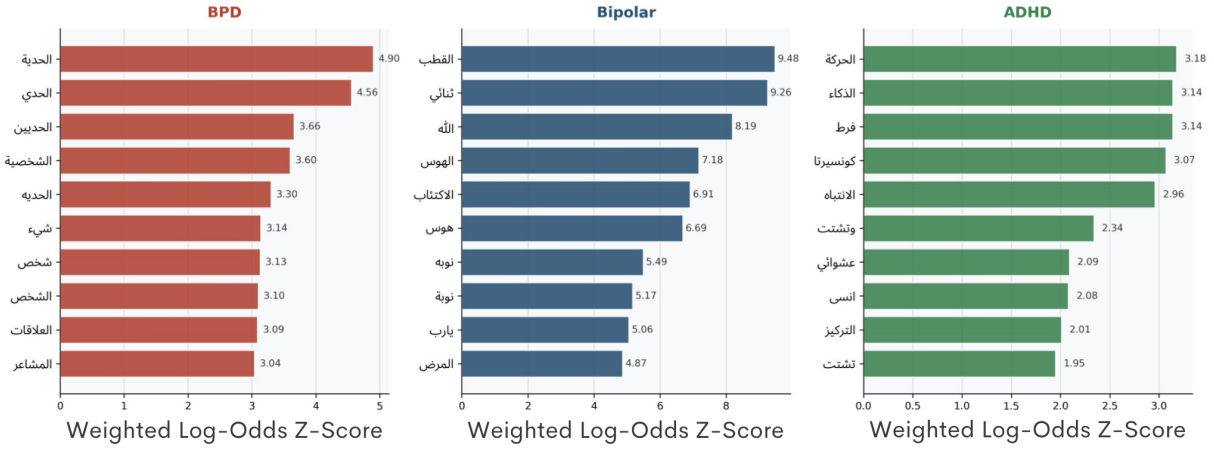

Figure 7: Top-10 community-distinctive words by weighted log-odds z-score. BPD: relational/diagnostic; Bipolar: episode/religious; ADHD: symptom/medication.

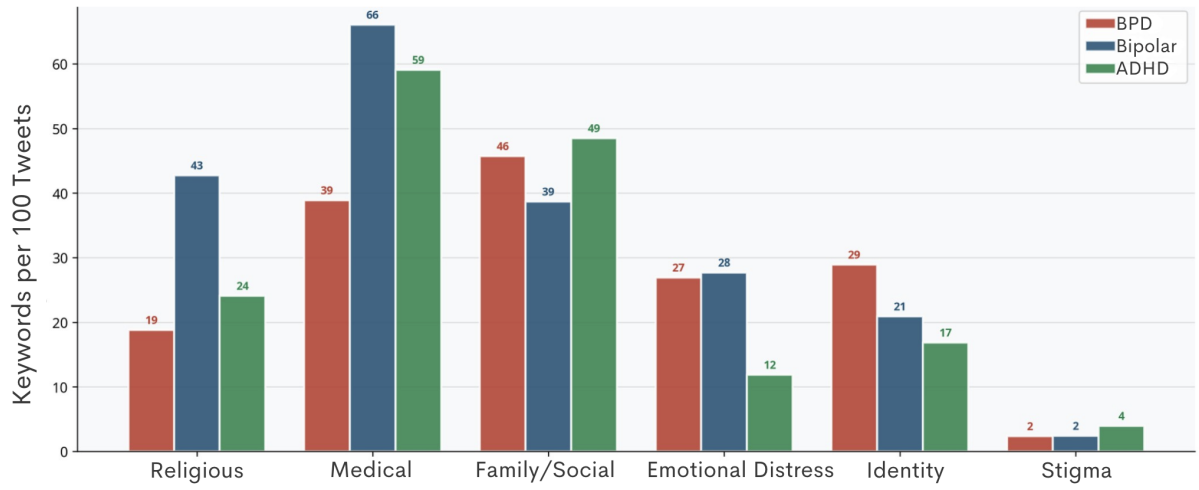

Figure 8: Cultural domain keyword rates (per 100 tweets). Bipolar leads on Religious and Medical; BPD on Identity and Emotional Distress.

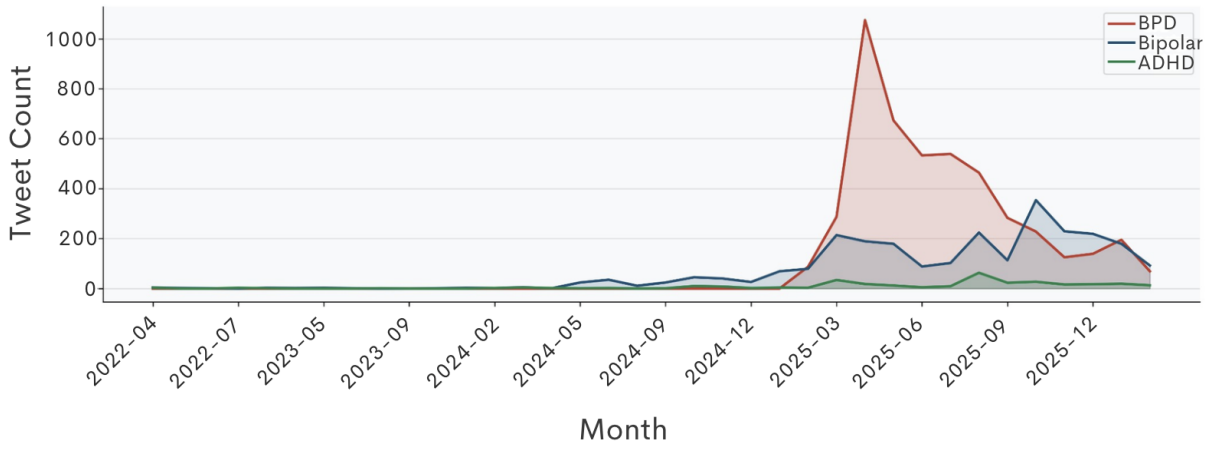

Figure 9: Monthly tweet volume (2022–2026). BPD concentrated in 2025; Bipolar spans the full period but with meaningful volume from 2024 onward.