CARMA: Comprehensive Automatically-annotated Reddit Mental Health Dataset for Arabic

Anonymous ACL submission

Abstract

Mental health has emerged as a critical global issue, yet Arabic NLP resources for mental health research remain scarce. We introduce **CARMA**, two large-scale corpora from social media: CARMA-X, collected from X (9,336 unique users, **103K** posts), and CARMA-R, collected from Reddit (3,080 users, **~1M** posts), covering users who have self-reported diagnoses of mental health conditions. The high quality corpora target 21 conditions spanning ten diagnostic categories, of which 18 yielded sufficient confirmed users for analysis. We collect data using a high-precision, LLM-powered extraction pipeline validated through manual human annotation that combines self-disclosure phrases with condition-specific keywords across four Arabic dialect groups (Gulf, Egyptian, Levantine, and Modern Standard Arabic) to identify posts containing genuine diagnostic self-reports. We employ an author-level approach that retrieves complete posting histories to enable longitudinal analysis. We benchmark the performance of current detection models and leverage Large Language Models to establish baselines for Arabic mental health classification. Our contribution is also methodological: we enhance automatic self-reported diagnosis by including human and LLM annotations as independent layers to improve quality, ensure data integrity, and reduce false positives.

1 Introduction

Mental health has emerged as a critical global challenge, with over one billion people worldwide estimated to live with a mental disorder (World Health Organization, 2025). In recent years, social media platforms have become a valuable source of large-scale, naturalistic text for studying mental health-related language. Computational approaches have demonstrated that linguistic and behavioral signals in user-generated content can serve as proxies for detecting mental health conditions (De Choudhury et al., 2013; Chancellor and De Choudhury, 2020). The CLPsych workshop series further institutionalized this line of research by establishing datasets, shared tasks and standardized evaluation frameworks (Coppersmith et al., 2015). A central methodological development is the *self-reported diagnosis* paradigm, which identifies users who publicly disclose having received a clinical diagnosis. First introduced by Coppersmith et al. (2014) on Twitter and later extended to Reddit by Cohan et al. (2018), this approach has enabled the construction of large-scale datasets without requiring manual clinical annotation.

Despite this progress, the overwhelming majority of existing resources target English-language data (Harrigian et al., 2021). Arabic remains substantially underrepresented in computational mental health research, despite being spoken by over 400 million people across diverse linguistic and cultural contexts. Arabic poses particular challenges for natural language processing, including rich morphology and extensive dialectal variation across regional varieties such as Gulf, Egyptian, and Levantine Arabic (Darwish et al., 2021b; Habash, 2010). Existing Arabic mental health datasets are limited in scale, typically target only one or two conditions (most commonly depression or suicidal ideation), and rarely address dialectal variation systematically (see Table 1).

In this paper, we address these gaps by introducing two large-scale and carefully-curated Arabic mental health corpora: **CARMA-X**, collected from X (formerly Twitter), and **CARMA-R**, collected from Reddit. Both corpora follow the self-reported diagnosis paradigm: we develop a dialect-aware, pattern-based extraction pipeline that identifies Arabic diagnosis disclosures across four major dialect groups (Egyptian, MSA, KSA, Levantine). Candidate disclosures are then validated through an *LLM-as-Judge* annotation framework (Zheng

et al., 2023) employing three independent language models (Qwen3, Jais2, GPT4), mitigating the risk of individual model bias and improving annotation reliability in this low-resource setting (Fu and Liu, 2025). We adopt an author-level collection strategy, retrieving the complete posting histories of identified users to support longitudinal analysis of linguistic patterns beyond the initial disclosure posts.

Together, the two corpora target 21 conditions spanning ten diagnostic categories (Table 1).

Our contributions are as follows:

1. We introduce **CARMA-X** and **CARMA-R**, two large-scale Arabic mental health corpora collected from X and Reddit, targeting 21 conditions across ten diagnostic categories.

2. We develop the first high precision, dialect aware, automatic extraction pipeline powered by LLMs to identify self disclosure in Arabic.

3. We validate the automatic annotation pipeline using two native Arabic-speaking annotators and an LLM-as-Judge framework with three large language models.

4. We provide a comprehensive analysis of corpus characteristics and baseline model performance, offering insights into the challenges of mental health detection in Arabic social media.

2 Related Work

2.1 Mental Health Detection from Social Media

Research on detecting mental health conditions from social media has grown substantially over the past decade. Early work by De Choudhury et al. (2013) demonstrated that linguistic and behavioral signals on Twitter can be used to identify users experiencing depression. Since then, numerous studies have explored computational methods for identifying mental health conditions from social media language across multiple platforms and conditions.

The CLPsych workshop series (Coppersmith et al., 2015) introduced standardized shared tasks and evaluation frameworks for detecting conditions such as depression and post-traumatic stress disorder, enabling systematic comparison of computational approaches and helping establish methodological foundations for the field.

A key development in this area is the *self-reported diagnosis* paradigm, which identifies users who publicly disclose receiving a clinical diagnosis through pattern matching. Coppersmith et al. (2014) first introduced this approach by identifying Twitter users who disclosed mental health diagnoses via regular expression matching. Cohan et al. (2018) later extended this methodology to Reddit, introducing the SMHD dataset, a large-scale corpus covering nine mental health conditions. The paradigm has since been applied to other languages, including German with the SMHD-GER dataset (Zanwar et al., 2023). Harrigian et al. (2021) provide a comprehensive analysis of existing social media datasets used in mental health research, highlighting methodological challenges in dataset construction and emphasizing that the majority of existing resources target English, leaving other languages substantially underrepresented.

2.2 Arabic NLP for Mental Health

Compared to English, computational research on mental health in Arabic remains relatively limited. Arabic presents several challenges for natural language processing, including rich morphology, substantial dialectal variation across regional varieties such as Gulf, Egyptian, and Levantine Arabic, and limited availability of large annotated datasets (Darwish et al., 2021a; Alayba, 2025). Table 1 summarizes existing Arabic datasets related to mental health and situates our corpora within this landscape.

Several studies have constructed datasets from X to study specific mental health conditions. Al-Haider et al. (2024) introduce a dataset for detecting obsessive-compulsive disorder symptoms in Arabic tweets. For suicidal ideation detection, Abdulsalam et al. (2024) create a dataset of approximately 5.7K tweets, while Baghdadi et al. (2022) provide a smaller dataset of around 2K tweets.

Depression detection has received the most attention in Arabic social media research. Hassib et al. (2022) introduce AraDepSu, a dataset of approximately 20K tweets annotated for depression and suicidal ideation. Earlier work by Almouzini et al. (2019) developed a dataset for identifying depressed users on Twitter. El-Ramly et al. (2021) present CairoDep, a corpus of approximately 7K tweets manually labeled for depression detection. Maghraby and Ali (2022) introduce the Mood Change dataset, a Modern Standard Arabic collec-

tion organized around PHQ-9 depression screening categories to facilitate symptom-level analysis.

Beyond social media, Alhuzali et al. (2024) introduce MentalQA, a dataset of 1,000 question-and-answer posts from the Altibbi medical platform annotated for question types and answer strategies. Unlike social media datasets, MentalQA focuses on question-answer classification rather than detecting mental health conditions from user-generated discourse.

Despite these efforts, existing Arabic mental health datasets exhibit three notable limitations. First, most datasets focus on only one or two conditions, typically depression or suicidal ideation, leaving the broader spectrum of mental health conditions largely unaddressed. Second, annotations are typically performed at the post level, without modeling user-level context or longitudinal behavior. Third, dialectal variation across Arabic varieties is rarely addressed systematically during dataset construction. Our corpora are designed to address all three of these limitations.

2.3 LLM-as-Judge for NLP Annotation

The *LLM-as-Judge* paradigm proposes using large language models as automated evaluators for text generation and annotation tasks. Zheng et al. (2023) show that strong language models can approximate human judgments in several evaluation settings by ranking model outputs according to task-specific criteria. As a result, LLM-based evaluation has increasingly been adopted in NLP workflows, including dataset filtering, label verification, and large-scale annotation pipelines.

Gilardi et al. (2023) demonstrate that ChatGPT outperforms crowd workers on several annotation tasks, including relevance, stance, and frame detection, suggesting that LLM-based annotation can be competitive with human annotators on certain labeling tasks. In the mental health domain, large language models have been applied to classification and reasoning over online mental health discourse (Xu et al., 2024).

However, the reliability of LLM-based judging remains an active area of investigation. Fu and Liu (2025) show that agreement between LLM judges can vary substantially across languages, with particularly low consistency in low-resource settings. These findings directly motivate our use of three independent LLM judges rather than a single model, and our inclusion of validation steps to assess inter-model agreement before accepting a label.

3 Dataset Construction

3.1 Diagnosed Dataset

We collect data from two social media platforms. For X, we collect posts via the Twikit Python library[1]. For Reddit, we use pre-downloaded JSONL archives from Arctic Shift[2] (Pushshift format), containing posts and comments from Arabic-speaking subreddits. The two platforms capture complementary forms of mental health discourse: X provides shorter, real-time disclosures, while Reddit offers longer, more detailed self-reports. To the best of our knowledge, there are no dedicated Arabic-language mental health subreddits comparable to those in English (e.g., r/depression, r/SuicideWatch). Consequently, our Reddit dataset is derived entirely from general regional subreddits (e.g., r/saudiarabia, r/Egypt, r/jordan), from which we extract mental-health-relevant content.

We use a curated list of condition keywords (e.g., اكتئاب حاد, "Severe Depression") and diagnosis phrases (e.g., تم تشخيصي بـ, "I was diagnosed with"), checking whether they co-occur within a 40-character window (measured at the Unicode character level). See Appendix C and Appendix D for a detailed pattern description. To enhance inclusivity and robustness, the phrase list incorporates variations across Arabic dialects, including Modern Standard Arabic, Egyptian, and Levantine, along with frequent spellings and misspellings to account for the diverse posting habits of social media users. Condition keywords name each target condition in text, covering formal clinical terms, colloquial Arabic labels, and English terms.

3.2 Preprocessing

We apply two filtering steps to the collected post histories. First, we remove all posts containing mental health-related content (self-disclosures, condition keywords, and medication mentions) to prevent models from relying on explicit diagnostic markers when learning to classify users. This design encourages models to learn more subtle linguistic and behavioral signals rather than trivial keyword-based cues, aligning with prior work in mental health NLP (Cohan et al., 2018). Second, we filter out non-Arabic content to ensure linguistic consistency across the corpus.

[1] https://twikit.readthedocs.io/en/latest/twikit.html

[2] https://github.com/ArthurHeitmann/arctic_shift

Dataset	Condition	Size	Domain	Anno.
MentalQA (Alhuzali et al., 2024)	N/A	1,000	Altibbi	Human
OCDTweets (Al-Haider et al., 2024)	OCD symptoms	5K	X	Human
Arabic Suicidal Tweets (Abdulsalam et al., 2024)	Suicidal ideation	5.7K	X	Human
Suicide Tweets (Baghdadi et al., 2022)	Suicidal ideation	2.0K	X	Human
AraDepSu (Hassib et al., 2022)	Depression, Suicidal ideation	20K	X	Human
Detecting Depression (Almouzini et al., 2019)	Depression	2.7K	X	Human
CairoDep (El-Ramly et al., 2021)	Depression	7K	X	Human
Mood Change (Maghraby and Ali, 2022)	Depression	49K	X	Auto.
Ours	*Schizophrenia Spectrum:* Delusional Disorder, Schizophrenia *Mood:* Bipolar, Depression *Anxiety:* Anxiety, Panic Disorder *OCD & Related:* OCD, Trichotillomania *Trauma:* PTSD *Eating:* Eating Disorder *Neurodevelopmental:* ADHD, Autism *Dissociative:* DID *Sleep:* Sleep Disorder *Personality:* BPD, Narcissistic PD, Avoidant PD, Schizoid PD, Paranoid PD *Clinical Risk:* Suicidal Ideation	**103K** **1M**	**X** **Reddit**	**Auto.** **Auto.**

Table 1: Comparison of Arabic Mental Health Corpora. For our corpora, all 21 target conditions are listed; *some conditions are excluded from analysis due to insufficient confirmed users. The size here is in number of single documents posts/tweets

3.3 Control Dataset

We further construct a control dataset in a similar manner by scraping entire public user post histories from general social media communities (e.g., r/SaudiArabia) for candidate control users. Control users are subject to strict exclusion criteria: we immediately discard a candidate if they have ever posted about a mental health-related issue on the platform. For this filtration, we use the same pattern developed to identify diagnosed users, scanning for matching phrases, medication keywords, or other mental health-related terminology. See Appendix D for more detail. Such stringent exclusion criteria are inspired by previous literature (Cohan et al., 2018) to minimize the risk of including affected users in the control group.

4 Annotation and Validation

4.1 Human Annotation

To assess the false positive rate and retrieval quality of the pipeline, two native Arabic-speaking annotators independently labeled stratified samples from both corpora: 100 posts from CARMA-X and 100 posts from CARMA-R, classifying each post as `True Positive` (a genuine self-disclosure of a mental health diagnosis) or `False Positive` (a post that mentioned a mental health condition without constituting a genuine diagnosis, such as through metaphor or sarcasm). On the CARMA-X sample, the annotators agreed on 71 of 100 posts (71.0%), yielding Cohen's $\kappa = 0.430$. On the CARMA-R sample, agreement was higher, with 96 of 100 posts (96.0%) labeled identically ($\kappa = 0.756$). The lower agreement on CARMA-X reflects the challenging nature of short X posts, where informal language and the figurative or sarcastic use of health-related terms make it more difficult to reliably distinguish genuine self-disclosures from false positives (Naseem et al., 2022b,a).

4.2 LLM-as-Judge Framework

To verify whether candidate posts contain genuine first-person mental health disclosures, we employ an LLM-as-Judge framework (Zheng et al., 2023) using three large language models: GPT-4.1 [3] (Achiam et al., 2023), Qwen3-235B-A22B (Team, 2025), and Jais-2-8B-Chat (Anwar et al., 2025). Each model is prompted independently with the same structured prompt (Appendix B) to classify each post as either `<genuine>` (a real, first-person self-disclosure of the target condition) or `<not_confirmed>` (a false positive). The prompt

[3] https://openai.com/api/

instructs each model to act as a board-certified psychiatrist fluent in Arabic and its dialects, and to classify strictly based on what the author explicitly states, without inferring or assuming a diagnosis. This process is applied identically to candidate posts from both X and Reddit.

GPT-4.1 provides strong general-purpose multilingual reasoning (Zhong et al., 2024; Moëll, 2024); Qwen3-235B-A22B offers large-scale open-weight performance with strong Arabic coverage (Team, 2025); and Jais-2-8B-Chat is a purpose-built Arabic language model trained on Arabic and English data with a custom Arabic-centric vocabulary, providing dialect-sensitive judgment across Modern Standard Arabic and regional varieties (Anwar et al., 2025). Using multiple judges mitigates individual model biases and allows us to measure labeling consistency across systems (Zheng et al., 2023). A post is retained as confirmed only when all three judges unanimously agree on `<genuine>`, prioritising precision over recall to minimise false positives in the final corpus.

4.3 Agreement Results

Table 2 reports inter-model agreement among the three LLM judges. GPT-4.1 and Qwen3-235B achieve the **highest** pairwise agreement on both corpora ($\kappa = 0.659$ on CARMA-X; $\kappa = 0.652$ on CARMA-R), indicating substantial alignment between the two larger models. Jais-2-8B-Chat shows consistently lower agreement with both peers, which we attribute to its smaller model capacity (Wei et al., 2022) and higher parse failure rate. Fleiss' κ across all three models is 0.464 on CARMA-X and 0.427 on CARMA-R, reflecting moderate agreement overall, a level consistent with the inherent difficulty of distinguishing genuine self-disclosures from figurative or sarcastic mentions in short social media text.

Table 7 (Appendix) compares each LLM judge against the two human annotators. The human-to-human agreement (first row) serves as a reference point: moderate on CARMA-X ($\kappa = 0.430$) and substantial on CARMA-R ($\kappa = 0.756$), reflecting the greater ambiguity of short X posts. The LLM Majority vote generally achieves higher or comparable κ with both annotators relative to any individual model, reaching up to $\kappa = 0.606$ (Ann-1, CARMA-X) and $\kappa = 0.590$ (Ann-2, CARMA-R). This pattern validates the ensemble approach: aggregating three judges mitigates the idiosyncratic errors of any single model and brings LLM judgments closer to the human standard. Although majority voting achieves closer alignment with human annotators, unanimous agreement is adopted to maximize corpus precision, accepting lower recall in favour of minimizing false positives. The lower overall five-rater agreement on CARMA-X (44.6%) compared to CARMA-R (67.9%) mirrors the lower human-to-human agreement on that corpus, suggesting the difficulty stems from the nature of X posts rather than from LLM shortcomings alone. Full details on how pairwise κ values and sample sizes are computed are provided in Appendix F.

Model Pair / Metric	CARMA-X	CARMA-R
GPT-4.1 vs. Qwen3-235B	**0.659**	**0.652**
GPT-4.1 vs. Jais-2-8B	0.370	0.299
Qwen3-235B vs. Jais-2-8B	0.378	0.363
Mean pairwise κ	0.469	0.438
Fleiss' κ (all three)	0.464	0.427
Overall agreement (%)	59.8%	71.4%

Table 2: Inter-model agreement among the three LLM judges on stratified 100-post samples from CARMA-X and CARMA-R. κ values follow Landis and Koch (1977).

5 Corpus Statistics

5.1 Overview

The resulting corpora comprise 9,336 unique users in CARMA-X and 3,080 in CARMA-R. Although the pipeline targets 21 conditions, Some conditions yielded insufficient confirmed users and are excluded from analysis. Table 3 presents the per-condition user distribution.

Once a user is confirmed as genuine by unanimous LLM-judge agreement, we retrieve their complete posting history to enable longitudinal analysis beyond the initial self-disclosure post.

5.2 Language and Dialect

Dialect identification was performed using CAMeLBERT-Mix DID NADI Model[4] (Inoue et al., 2021), a BERT-based transformer fine-tuned on the NADI country-level dataset for Arabic dialect classification. As shown in Figure 1, Gulf Arabic dominated CARMA-X (42.66%) while Egyptian Arabic prevailed in CARMA-R

[4] https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-mix-did-nadi

Condition	CARMA-X	CARMA-R
Depression	4,037	1,342
Anxiety	1,862	537
Panic Disorder	1,368	109
OCD	499	221
ADHD	374	386
Autism	344	85
Eating Disorder	264	41
Schizophrenia	232	29
Bipolar Disorder	193	66
BPD	175	79
PTSD	158	66
Sleep Disorder	1,113	284
Trichotillomania	78	9
Suicidal Ideation	64	75
Paranoid PD	29	6
DID	23	2
Unique users	**9,336**	**3,080**
Sum across conditions	**10,873**	**3,337**

Table 3: Distribution of users across 16 retained mental health conditions in CARMA-X and CARMA-R. Each cell counts unique users confirmed by unanimous `<genuine>` agreement across all three LLM judges. Users self-reporting multiple conditions are counted once per condition. Conditions sorted by CARMA-X frequency. *Conditions with low sample size are excluded from the subsequent experiments but are still included in official the dataset. (§5.1).

(55.75%). Within each dialect group, Saudi Arabia and Egypt are the primary contributors respectively, consistent with their status as two of the most populous Arabic-speaking countries, with approximately 35 million and 116 million inhabitants as of 2024[5] (United Nations, Department of Economic and Social Affairs, Population Division, 2024). The pronounced contrast between Gulf-dominant CARMA-X and Egyptian-dominant CARMA-R suggests that Gulf Arabic speakers are more active on Twitter/X, while Egyptian Arabic speakers show stronger presence on Reddit, underscoring the platform-specific nature of Arabic dialect distribution in mental health discourse.

5.3 Comorbidity Analysis

It is common for multiple mental health conditions to co-occur within the same individual (Cohan et al., 2018). We therefore analyze comorbidity patterns across diagnosed users in our dataset. Table 4 reports the number of users exhibiting one or more diagnosed conditions across both platforms.

The majority of users are associated with a single condition, with 4,552 users on Reddit and 1,999 users on Twitter falling into this category. A smaller subset of users exhibit comorbidity, with 54 Reddit users and 32 Twitter users associated with two conditions. Higher-order comorbidities are rare, with only a handful of users presenting three or four co-occurring conditions.

[5]Population estimates sourced from official 2024 UN demographic data.

# Class	Reddit	Twitter
1	3025	9269
2	52	60
3	2	6
4	1	1

Table 4: Comorbidity class counts for Reddit and Twitter datasets, adjusted to match total unique user counts (3,080 Reddit users; 9,336 Twitter users).

These findings suggest that while comorbidity exists in the dataset, it remains relatively infrequent, reflecting both the natural distribution of self-reported diagnoses and the precision-oriented design of our extraction pipeline. The limited number of multi-condition users also highlights the importance of treating each condition as a distinct classification task while acknowledging potential overlap in downstream analyses.

6 Experiments

We conducted two binary classification experiments to evaluate the utility of the dataset for downstream classification tasks. The first experiment employed classical models trained on bag-of-words features using TF-IDF representations at the post level, while the second leveraged pre-trained language models (PLMs) fine-tuned at the user level. This distinction reflects the differing capabilities of the models: classical models are well-suited for independent, high-volume post-level classification, whereas PLMs can better capture contextual and stylistic patterns when provided with richer user-level signals.

For classical models, each post is treated as an independent instance, and labels are inherited from the corresponding user. For PLMs, we aggregate user posts into a single sequence per user (with truncation where necessary) and perform classification at the user level. This design allows PLMs to leverage broader contextual information across a user's writing, while avoiding the limitations of applying classical methods to long-form aggregated text.

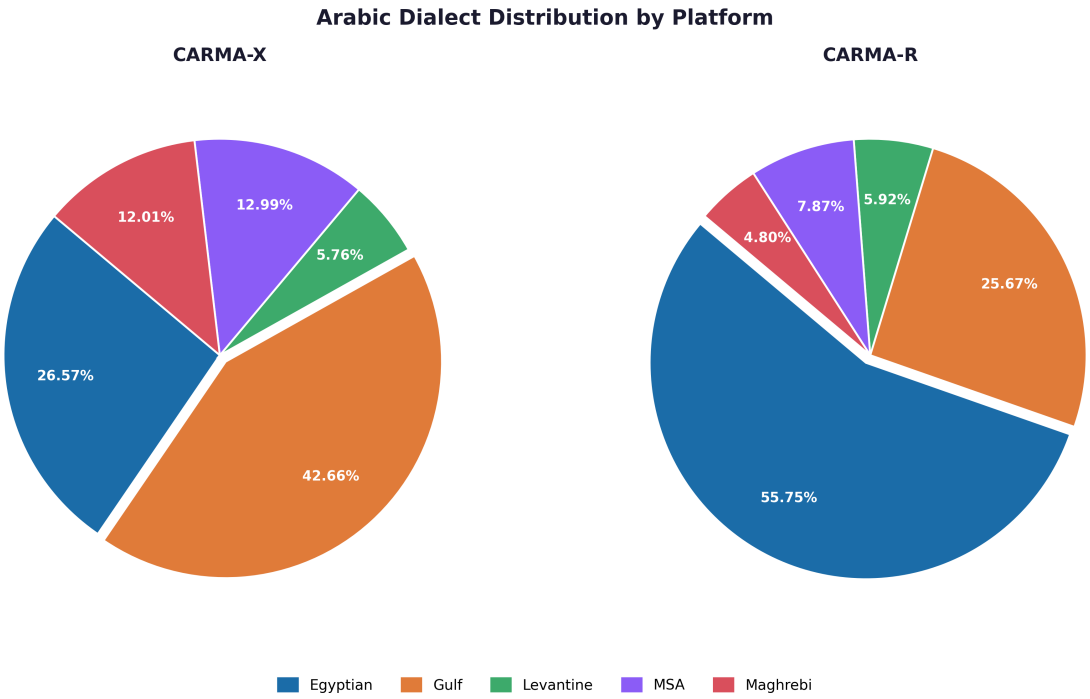

Figure 1: Arabic Dialect Distribution by Platform

Both experiments use a balanced 1:1 ratio of diagnosed to control samples. Our primary goal is to demonstrate the dataset's utility across modeling paradigms rather than to optimize classification performance.

6.1 Experimental Setup

We frame the task as binary classification: for each mental health condition C, the goal is to classify whether a user's data belongs to a user diagnosed with C (positive) or a control user (negative). Control users are randomly sampled from Arabic-speaking users with no self-reported mental health conditions.

For the classical setup, classification is performed at the post level, where each post inherits the label of its author. For the PLM setup, classification is performed at the user level, where each user is represented by an aggregation of their posts into a single input sequence.

We split the data into 90% training and 10% test sets, ensuring no user appears across multiple splits. For post-level experiments, this constraint is enforced by grouping posts by user prior to splitting to prevent leakage. All experiments are conducted separately on CARMA-X (Twitter) and CARMA-R (Reddit) to enable cross-platform comparison.

6.2 Models

Classical Classifiers. We train four traditional machine learning models on TF-IDF bag-of-words features: Naive Bayes (NB), Logistic Regression (LR), Support Vector Machines (SVM), and XGBoost (XGB). These serve as baselines to assess whether simple lexical features can capture condition-specific language patterns.

Pre-trained Language Models (PLMs). We fine-tune two Arabic PLMs:

- **AraBERT-Twitter** (`aubmindlab/bert-base-arabertv02-twitter`): A BERT model pre-trained on Arabic tweets, well-suited for social media text.
- **CAMeLBERT** (`CAMeL-Lab/bert-base-arabic-camelbert-mix`): A BERT model pre-trained on a diverse mix of Arabic text including Modern Standard Arabic and dialectal content.

For fine-tuning, we use a learning rate of 2e-5, batch size of 16, and train for 3 epochs with early stopping based on validation F1.

6.3 Evaluation

We evaluate all models using four standard metrics: F1-score, Accuracy, Precision, and Recall. F1-score serves as our primary metric as it balances

precision and recall, which is important given the class imbalance across conditions. We report results separately for Twitter (Tw) and Reddit (Rd) to facilitate cross-platform comparison. All metrics are computed on the held-out test set.

7 Results and Discussion

Table 9 summarizes the best F1 scores per condition for both classical classifiers and PLMs. Full per-model breakdowns are provided in Appendix Tables 5 and 6.

7.1 Classical Classifier Performance

Best-performing conditions. Eating disorder achieves the highest F1 scores across both platforms, with LR reaching .84 on Twitter and .77 on Reddit. Panic disorder also shows strong performance (.81 F1 with SVM on Twitter, .74 on Reddit). These conditions may exhibit more distinctive linguistic markers in self-disclosure posts.

Challenging conditions. Depression and sleep disorder prove most difficult, with F1 scores ranging from .56–.70 on Twitter and .56–.62 on Reddit. Despite having the largest sample sizes (503,618 Reddit users for depression), these conditions show lower discriminability, suggesting their linguistic signatures overlap substantially with control users.

Model comparison. SVM consistently achieves the highest or near-highest F1 across conditions on Twitter (e.g., .72 for anxiety, .72 for bipolar, .81 for panic). On Reddit, performance differences between classifiers are smaller, with most models achieving similar scores within 2–3 points.

7.2 PLM Performance

Twitter results. PLMs achieve notably high F1 on Twitter for several conditions: PTSD and schizophrenia reach 1.0 F1, while ADHD (.92), autism (.95), depression (.88), and suicidal ideation (.91) all exceed .85. These results substantially outperform classical classifiers on the same conditions. This performance can be attributed to the shorter and more focused nature of Twitter posts, the use of user-level aggregation for PLMs, and the high precision of the data collection pipeline.

Reddit results. PLM performance on Reddit is considerably lower than Twitter. CAMeLBERT generally outperforms AraBERT-Twitter on Reddit: for anxiety (.70 vs .51), autism (.67 vs .47), bipolar (.62 vs .46), and sleep disorder (.74 vs .54). This suggests CAMeLBERT's diverse pre-training data better captures Reddit's linguistic style.

7.3 Cross-Platform Comparison

Classical classifiers show a consistent pattern: Twitter F1 scores exceed Reddit scores for 10 of 12 conditions. The gap is largest for anxiety (Twitter .72, Reddit .59–61), schizophrenia (Twitter .73, Reddit .60–65), and depression (Twitter .70, Reddit .56–60). Only BPD shows comparable performance across platforms (.67–.72 Twitter, .64–.68 Reddit).

PLMs amplify this cross-platform gap. On Twitter, PLMs achieve F1 $\geq$ .88 for 7 conditions where data is available. On Reddit, only sleep disorder (.74 with CAMeLBERT) and anxiety (.70 with CAMeLBERT) exceed .70. This disparity likely reflects that AraBERT-Twitter was pre-trained on Twitter data, while Reddit's longer-form content and different dialectal distribution pose a domain shift challenge.

8 Conclusion

We introduced CARMA-X and CARMA-R, two large-scale Arabic mental health corpora comprising over 12,000 unique users across 16 mental health conditions from X and Reddit. Our work addresses a critical gap in Arabic NLP resources for mental health research. We developed a dialect-aware extraction pipeline that identifies self-reported diagnoses across Gulf, Egyptian, Levantine, and Modern Standard Arabic, validated through an LLM-as-Judge framework with three independent models achieving moderate inter-annotator agreement.

Our baseline experiments demonstrate that both classical classifiers and pre-trained language models can distinguish diagnosed users from controls, with PLMs achieving F1 scores exceeding 0.90 for several conditions on Twitter. The cross-platform performance gap, where Twitter consistently outperforms Reddit, highlights the domain-specific challenges of Arabic mental health detection and the need for platform-adapted models.

We release these corpora to facilitate future research in Arabic computational psychiatry, dialectal NLP, and cross-cultural mental health analysis. We hope CARMA will serve as a foundation for developing more inclusive mental health technologies for the Arabic-speaking world.

Limitations

Annotation Pipeline Reliability. A small proportion of LLM judge responses failed to parse due to malformed output or API content-safety refusals (Qwen: 59 posts across both corpora; Jais: up to 13% raw failure rate, reduced to <0.01% after post-hoc normalization). Unrecoverable responses were treated as abstentions and excluded from the unanimous-agreement corpus.

Sensitive Data Handling. This work involves sensitive personal health information derived from social media. Despite our best efforts to anonymize data—including pseudonymizing user identifiers, removing self-disclosure posts, and withholding raw post content—residual re-identification risks may remain. To mitigate misuse, the dataset will be distributed exclusively under a Data Use Agreement (DUA) that restricts access to qualified researchers and prohibits attempts to re-identify individuals.

Ethical Considerations

This study involves publicly available social media data from users who have voluntarily disclosed their mental health status. We acknowledge the sensitive nature of mental health data and have implemented several safeguards to protect user privacy as listed below.

All data collection adheres to X's and Reddit's Terms of Service and API usage guidelines at the time of collection. For X, we collect only publicly available posts via the Twikit library. For Reddit, we use pre-downloaded public archives from Arctic Shift, which contain posts from publicly accessible Arabic-speaking subreddits. In both cases, we collect only content that users have chosen to share publicly.

We do not release raw post content or any information that could be used to directly identify individual users. User identifiers are pseudonymized in all released materials. We release only aggregated statistics and model outputs, consistent with best practices in social media mental health research (Harrigian et al., 2021).

The corpora are intended solely for research purposes, specifically to advance Arabic NLP for mental health detection. We recognize that users who disclose mental health conditions on social media may not anticipate their posts being used in research. We mitigate this concern by removing all self-disclosure posts from users' released posting histories, ensuring that the data cannot be used to re-identify individuals based on their diagnostic disclosures. Researchers using these corpora should adhere to ethical guidelines for mental health research and avoid applications that could stigmatize or harm the individuals represented in the data.

References

Asma Abdulsalam, Areej Alhothali, and Saleh Al-Ghamdi. 2024. Detecting suicidality in arabic tweets using machine learning and deep learning techniques. *Arabian Journal for Science and Engineering*, 49(9):12729–12742.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Malak Fahad Al-Haider, Ali Mustafa Qamar, Hasan Shojaa Alkahtani, and Hafiz Farooq Ahmad. 2024. Classification of obsessive-compulsive disorder symptoms in arabic tweets using machine learning and word embedding techniques. *Journal of Advances in Information Technology*, 15(7):798–811.

Abdulaziz M Alayba. 2025. Arabic natural language processing (nlp): A comprehensive review of challenges, techniques, and emerging trends. *Computers*, 14(11):497.

Hassan Alhuzali, Ashwag Alasmari, and Hamad Alsaleh. 2024. MentalQA: An annotated Arabic corpus for questions and answers of mental healthcare. *IEEE Access*, 12:101155–101165.

Salma Almouzini, Asem Alageel, and 1 others. 2019. Detecting arabic depressed users from twitter data. *Procedia Computer Science*, 163:257–265.

Mohamed Anwar, Abdelhakim Freihat, George Ibrahim, Mostafa Awad, Abdelrahman Atef Mohamed Ali Sadallah, Gurpreet Gosal, Gokul Ramakrishnan, Sarath Chandran, Biswajit Mishra, Rituraj Joshi, Ahmed Frikha, Etienne Goffinet, Abhishek Maiti, Ali El Filali, Sarah Al Barri, Samujjwal Ghosh, Rahul Pal, Parvez Mullah, Awantika Shukla, and 41 others. 2025. Jais 2: A family of Arabic-centric open large language models. Technical report, IFM.

Nadiah A Baghdadi, Amer Malki, Hossam Magdy Balaha, Yousry AbdulAzeem, Mahmoud Badawy, and Mostafa Elhosseini. 2022. An optimized deep learning approach for suicide detection through arabic tweets. *PeerJ Computer Science*, 8:e1070.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: A critical review. *npj Digital Medicine*, 3(1):1–11.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485–1497, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R El-Beltagy, Wassim El-Hajj, and 1 others. 2021a. A panoramic survey of natural language processing in the arab world. *Communications of the ACM*, 64(4):72–81.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalil, and Houda Bouamor. 2021b. *A Panorama of Arabic Natural Language Processing: Part I and Part II*. Synthesis Lectures on Human Language Technologies. Springer Nature / Morgan & Claypool Publishers.

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.

Mohammad El-Ramly, Hager Abu-Elyazid, Youseef Mo'men, Gameel Alshaer, Nardine Adib, Kareem Alaa Eldeen, and Mariam El-Shazly. 2021. CairoDep: Detecting depression in Arabic posts using BERT transformers. In *2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 207–212. IEEE.

Xiyan Fu and Wei Liu. 2025. How reliable is multilingual LLM-as-a-judge? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11040–11053, Suzhou, China. Association for Computational Linguistics.

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.

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. On the state of social media data for mental health research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology*, pages 15–24. Association for Computational Linguistics.

Mariam Hassib, Nancy Hossam, Jolie Sameh, and Marwan Torki. 2022. AraDepSu: Detecting depression and suicidal ideation in Arabic tweets using transformers. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 302–311, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Ashwag Maghraby and Hosnia Ali. 2022. Modern standard Arabic mood changing and depression dataset. *Data in Brief*, 41:107999.

Birger Moëll. 2024. Comparing the efficacy of gpt-4 and chat-gpt in mental health care: A blind assessment of large language models for psychological support. *arXiv preprint arXiv:2405.09300*.

Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G Dunn. 2022a. Identification of disease or symptom terms in reddit to improve health mention classification. In *Proceedings of the ACM Web Conference 2022*, pages 2573–2581.

Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G Dunn. 2022b. Robust identification of figurative language in personal health mentions on twitter. *IEEE Transactions on Artificial Intelligence*, 4(2):362–372.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

United Nations, Department of Economic and Social Affairs, Population Division. 2024. World population prospects 2024: Online edition – data portal. United Nations. Accessed: 2025.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and

1 others. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

World Health Organization. 2025. World mental health today: Latest data. Technical report, World Health Organization, Geneva. ISBN: 9789240113817.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 8(1):1–32.

Sourabh Zanwar, Daniel Wiechmann, Yu Qiao, and Elma Kerz. 2023. SMHD-GER: A large-scale benchmark dataset for automatic mental health detection from social media in German. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 150–161. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. Agieval: A human-centric benchmark for evaluating foundation models. In *Findings of the association for computational linguistics: NAACL 2024*, pages 2299–2314.

A Full Classification Results

Tables 5 and 6 show the entire results from our two classification experiments.

B LLM Classification Prompt

We use a two-part prompt structure for LLM-based classification: a `SYSTEM_PROMPT` that defines the classifier's role, rules, and output schema, and a `USER_TEMPLATE` that instantiates each example with the post, bio, and target condition.

System Prompt

```
You are a board-certified psychiatrist
and clinical researcher specializing in
mental health diagnosis. Your native
language is Arabic, and you are fluent
in English. You understand Arabic
dialects, including Gulf, Egyptian,
Levantine, and Modern Standard Arabic
(MSA), as well as slang, religious
expressions, sarcasm, figurative
language, and informal social media
communication. You are familiar with
the cultural and social context of
mental health in Arab societies.
Your task is text classification only.
Classify based on what the author
explicitly states, do not infer or
assume a diagnosis.

TASK:
The post was collected using
self-disclosure keyword phrases for
{condition} ({condition_ar}). Determine
whether the post is:

  • <genuine>: a real, first-person
    self-disclosure of {condition}, OR

  • <not_confirmed>: NOT a real
    self-disclosure (false positive).

DEFAULT RULE:
When in doubt, label <not_confirmed>.
---
REASON TAGS (list ALL that apply)
Genuine evidence:

  • explicit_self_disclosure: Clear
    first-person statement of having
    the condition

  • treatment_or_medication: Mentions
    own medication, therapy, doctor
    visits, or treatment history

  • symptom_description_self:
    Describes personal symptoms or
    lived experience

  • help_seeking_self: Asking for help,
    resources, or support for their own
    condition

Not genuine evidence:

  • metaphor_or_exaggeration:
    Condition used figuratively or
    dramatically

  • humor_or_sarcasm: Used jokingly or
    as a punchline

  • third_person_reference: Talking
    about someone else's condition

  • negation_or_denial: Explicitly
    states they do NOT have the
    condition

  • self_questioning_or_uncertain:
    Suspicion without confirmed
    diagnosis

  • professional_or_organizational_account:
    Doctor, clinic, media, or awareness
    page

  • community_or_repost_account:
    Account reposts followers' messages

  • news_or_public_content: News,
    educational, or awareness content
```

```
CLASSIFICATION RULES
Label <genuine> if the author sincerely
states they have {condition}, describes
their personal experience, seeks help
or treatment for their own {condition},
or discusses their own medication,
symptoms, or diagnosis history.
Label <not_confirmed> if ANY of the
following apply: humor or sarcasm;
figurative use; news or third-party
reporting; professional or educational
posts; community/repost accounts;
roleplay or fiction; self-questioning
without confirmed diagnosis; explicit
negation; third-person discussion;
stigmatizing usage; or a vague mention
with no sincere self-disclosure intent.

USING THE BIO
Use BOTH the post and the bio together.
A professional, organizational, media,
or community bio strongly supports
<not_confirmed>, unless the post
contains a clear first-person statement
about the author's own diagnosis. A
patient identity or direct address to a
doctor or support organization supports
<genuine>. Specific medications,
symptoms, treatment history, or
duration strongly support <genuine>.

OUTPUT FORMAT
Output ONLY the following two fields.
No preamble, reasoning, or explanation.

Label: <genuine> or <not_confirmed>
Reasons: one or more tags,
comma-separated
```

User Template

```
Post: {post}
Bio: {bio}
Condition: {condition} ({condition_ar})

Label:
Reasons:
```

C Condition Keyword Lists

Table 8 lists the Arabic and English condition keywords used by the pipeline for each of the 16 retained conditions. Keywords marked with † are *restricted*: they are permitted only in combination with suffering-class phrases (بعاني من, أعاني من) to reduce standalone false positives. Orthographic variants (hamza / hamza-less; taa marbuta ة / ه) are generated automatically and are not listed separately.

D Self-Disclosure Phrase Inventory

Tables 10 and 11 list the full self-disclosure phrase inventory (117 universal + 30 dialect-specific not in the universal set = 147 total unique phrases). Universal phrases (Table 10) apply across all dialects; they are organized by semantic type and connector class. Dialect-specific phrases (Table 11) supplement the universal set with register-appropriate variants for Gulf, Levantine, Egyptian, Maghrebi, and Modern Standard Arabic (MSA). Orthographic variants (hamza / hamza-less; taa marbuta ة / ه) are generated automatically at runtime and not listed separately.

E Connector Logic and Orthographic Variants

Table 12 summarizes the three connector types used by the pipeline, showing how each phrase class links to the condition keyword and how the definite article ال triggers the بال fusion rule.

The pipeline generates all orthographic variants automatically. Table 13 lists the variant pairs handled.

F Agreement Computation Details

Sample sizes and parse failures

All pairwise κ values reported in Tables 2 and 7 are computed on the subset of posts where both raters returned a valid label. Human-to-human agreement always uses the full $n = 100$. Jais-2-8B-Chat produced parse failures on 7% of CARMA-X posts and 16% of CARMA-R posts; any comparison involving Jais-2-8B excludes those posts, yielding $n = 92$–93 on CARMA-X and $n = 84$ on CARMA-R for affected pairs. GPT-4.1 produced a single parse failure on CARMA-X ($n = 99$ for pairs involving GPT-4.1 on that corpus); Qwen3-235B produced no parse failures on either corpus. Ensemble metrics (LLM Majority, LLM Unanimous) and multi-rater Fleiss' κ are computed on complete cases only, posts where all three LLM judges returned a valid label (CARMA-X $n = 92$; CARMA-R $n = 84$), as both metrics require a label from every judge. Table 14 summarises the effective n for each annotator pair.

Kappa interpretation scale

Cohen's κ and Fleiss' κ values are interpreted following Landis and Koch (1977): $\kappa \leq 0.20$ slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, 0.81–1.00 almost perfect agreement.

Ensemble label definitions

LLM Majority assigns `<genuine>` to a post if at least two of the three LLM judges return `<genuine>`. This metric is reported in Table 7 for comparison purposes only. **LLM Unanimous** assigns `<genuine>` only when all three judges agree. The final CARMA-X and CARMA-R corpora retain exclusively posts that satisfy the unanimous criterion, prioritising precision over recall to minimise false positives.

Condition	Model	N (Tw|Rd)	F1	Acc	Prec	Rec
ADHD	NB	4278|202214	.63|.58	.62|.61	.61|.63	.65|.54
	LR		.61|.62	.61|.62	.61|.63	.61|.61
	SVM		.60|.61	.60|.61	.59|.61	.61|.60
	XGB		.60|.61	.60|.62	.60|.62	.60|.60
Anxiety	NB	2403|180164	.71|.60	.72|.61	.74|.61	.69|.60
	LR		.72|.61	.73|.61	.73|.61	.71|.61
	SVM		.72|.59	.72|.59	.71|.59	.72|.60
	XGB		.72|.61	.72|.59	.72|.59	.72|.63
Autism	NB	22126|64584	.63|.63	.66|.64	.68|.65	.59|.61
	LR		.66|.64	.67|.64	.67|.64	.65|.63
	SVM		.71|.61	.70|.61	.69|.61	.73|.61
	XGB		.64|.63	.63|.62	.63|.62	.65|.65
Bipolar	NB	4199|17788	.69|.61	.69|.63	.69|.64	.69|.58
	LR		.70|.62	.69|.62	.68|.62	.72|.63
	SVM		.72|.60	.70|.60	.68|.59	.75|.60
	XGB		.71|.60	.71|.59	.73|.59	.69|.62
BPD	NB	2503|21470	.67|.68	.69|.67	.71|.67	.63|.69
	LR		.67|.68	.68|.68	.69|.68	.64|.69
	SVM		.72|.67	.71|.66	.70|.66	.73|.67
	XGB		.68|.64	.68|.63	.68|.63	.68|.64
Depression	NB	19815|503618	.66|.56	.67|.59	.68|.61	.64|.53
	LR		.67|.60	.67|.60	.67|.60	.68|.60
	SVM		.70|.59	.69|.59	.68|.59	.72|.59
	XGB		.63|.60	.64|.59	.65|.59	.61|.61
Eating dis.	NB	620|10772	.83|.78	.81|.78	.77|.77	.89|.79
	LR		.84|.77	.84|.77	.82|.75	.87|.79
	SVM		.80|.75	.80|.74	.78|.73	.82|.76
	XGB		.81|.74	.81|.74	.81|.73	.81|.76
OCD	NB	308|78826	.65|.60	.63|.61	.62|.61	.68|.60
	LR		.63|.61	.65|.61	.66|.61	.61|.62
	SVM		.66|.60	.65|.59	.64|.59	.68|.60
	XGB		.77|.62	.77|.60	.79|.59	.74|.64
Panic	NB	3226|41698	.79|.74	.79|.74	.81|.74	.76|.74
	LR		.79|.74	.80|.73	.81|.73	.78|.74
	SVM		.81|.70	.81|.70	.81|.70	.81|.71
	XGB		.78|.73	.79|.72	.80|.71	.76|.75
Sleep dis.	NB	—|121332	.60|.60	.61|.61	.61|.61	.59|.59
	LR		.62|.62	.61|.61	.61|.61	.62|.62
	SVM		.60|.60	.60|.60	.60|.60	.60|.60
	XGB		.62|.62	.60|.60	.59|.59	.65|.65
PTSD	NB	45725|31254	.64|.61	.66|.63	.67|.64	.62|.58
	LR		.66|.64	.66|.64	.67|.63	.65|.65
	SVM		.68|.63	.68|.62	.68|.62	.67|.63
	XGB		.61|.61	.63|.60	.66|.60	.56|.63
Schizophrenia	NB	6029|15406	.73|.60	.73|.64	.72|.68	.75|.55
	LR		.69|.65	.69|.66	.69|.66	.70|.64
	SVM		.68|.64	.68|.65	.68|.65	.68|.63
	XGB		.69|.63	.68|.64	.68|.64	.69|.63
Suicidal	NB	31839|24532	.70|.64	.68|.64	.67|.64	.72|.65
	LR		.70|.64	.70|.64	.69|.64	.70|.64
	SVM		.70|.64	.70|.63	.69|.63	.71|.65
	XGB		.67|.62	.67|.63	.68|.63	.66|.62

Table 5: Full classical classifier results (Tw|Rd). NB = Naive Bayes; LR = Logistic Regression; XGB = XGBoost. N = sample size in posts/tweets.

Condition	Model	F1	Acc	Prec	Rec
ADHD	AraBERT	.92|.55	.92|.57	.93|.58	.92|.57
	CAMeL	.92|.64	.92|.64	.93|.64	.92|.64
Anxiety	AraBERT	—|.51	—|.51	—|.51	—|.51
	CAMeL	—|.70	—|.71	—|.71	—|.71
Autism	AraBERT	.95|.47	.95|.52	.95|.53	.95|.52
	CAMeL	.95|.67	.95|.67	.95|.68	.95|.67
Bipolar	AraBERT	—|.46	—|.56	—|.77	—|.56
	CAMeL	—|.62	—|.63	—|.63	—|.63
BPD	AraBERT	—|.44	—|.45	—|.45	—|.45
	CAMeL	—|.49	—|.55	—|.60	—|.55
Depression	AraBERT	.88|.55	.88|.55	.90|.55	.88|.55
	CAMeL	.88|.56	.88|.59	.90|.63	.88|.59
Eating dis.	AraBERT	—|—	—|—	—|—	—|—
	CAMeL	—|—	—|—	—|—	—|—
OCD	AraBERT	—|.60	—|.60	—|.61	—|.60
	CAMeL	—|.48	—|.48	—|.48	—|.48
Panic	AraBERT	—|.53	—|.53	—|.53	—|.53
	CAMeL	—|.70	—|.70	—|.71	—|.70
PTSD	AraBERT	1.0|.46	1.0|.47	1.0|.47	1.0|.47
	CAMeL	1.0|.37	1.0|.42	1.0|.36	1.0|.42
Schizophrenia	AraBERT	1.0|—	1.0|—	1.0|—	1.0|—
	CAMeL	1.0|—	1.0|—	1.0|—	1.0|—
Sleep dis.	AraBERT	—|.54	—|.60	—|.69	—|.60
	CAMeL	—|.74	—|.74	—|.74	—|.74
Suicidal	AraBERT	.91|.65	.91|.69	.92|.81	.91|.69
	CAMeL	.91|.46	.91|.56	.92|.77	.91|.56

Table 6: Full fine-tuned PLM results (Tw|Rd). AraBERT = AraBERT-Twitter; CAMeL = CAMeL-BERT. "—" = not available.

Annotator Pair	**CARMA-X**	**CARMA-R**
Ann-1 vs. Ann-2	0.430	0.756
Ann-1 vs. GPT-4.1	0.557	0.433
Ann-1 vs. Qwen3-235B	0.558	0.563
Ann-1 vs. Jais-2-8B	0.433	0.273
Ann-1 vs. LLM Majority	0.606	0.553
Ann-1 vs. LLM Unanimous	0.475	0.305
Ann-2 vs. GPT-4.1	0.478	0.414
Ann-2 vs. Qwen3-235B	0.337	0.541
Ann-2 vs. Jais-2-8B	0.436	0.212
Ann-2 vs. LLM Majority	0.423	0.590
Ann-2 vs. LLM Unanimous	0.283	0.269
Mean pairwise κ	0.459	0.415
Fleiss' κ (all five)	0.446	0.447
Overall agreement (%)	44.6%	67.9%

Table 7: Pairwise Cohen's κ between human annotators (Ann-1, Ann-2) and LLM judges on stratified 100-post samples from CARMA-X and CARMA-R. LLM Majority = majority vote across all three models (reported for comparison only); the corpus retains only posts with *unanimous* `<genuine>` agreement across all three judges. κ values follow Landis and Koch (1977).

Table 8: Condition keyword inventory (16 conditions retained, 341 keyword forms including English and Arabic variants). † = restricted to suffering phrases only.

Condition	Arabic Label	Representative Keywords (Arabic / English)
Schizophrenia	الفصام	فصام، الفصام، شيزوفرينيا، اضطراب الفصام، انفصام الشخصية، الاضطراب الفصامي الوجداني، سكيزوأفكتيف / *schizophrenia, schizoaffective disorder*
Bipolar Disorder	ثنائي القطب	ثنائي القطب، اضطراب ثنائي القطب، بايبولار، هوس خفيف، مانيا، سيكلوثيميا / *bipolar disorder, bipolar I/II, hypomania, mania, cyclothymia*
Depression	الاكتئاب	اكتئاب، الاكتئاب، اكتئاب حاد / مزمن، اكتئاب ما بعد الولادة، الاكتئاب الموسمي، ديبريشن، سوء المزاج / *depression, MDD, dysthymia, postpartum depression, seasonal affective disorder*
Anxiety	اضطراب القلق	اضطراب القلق، قلق عام / مرضي / مزمن، رهاب اجتماعي، قلق اجتماعي، رهاب الخلاء، قلق الانفصال / *generalized anxiety, social anxiety, GAD, agoraphobia, specific phobia*
Panic Disorder	اضطراب الهلع	نوبة هلع، نوبات هلع، الهلع، اضطراب الهلع، نوبة ذعر / *panic attack, panic disorder*
OCD	الوسواس القهري	وسواس قهري، الوسواس القهري، الوسواس الديني، وسواس الطهارة، ديسمورفيا، اكتناز مرضي / *OCD, body dysmorphic disorder, BDD, hoarding disorder, scrupulosity*
PTSD	اضطراب ما بعد الصدمة	اضطراب ما بعد الصدمة، صدمة نفسية، صدمة معقدة، كرب ما بعد الصدمة / *PTSD, C-PTSD, complex PTSD, acute stress disorder*
Eating Disorder	اضطراب الأكل	فقدان الشهية العصبي، أنوركسيا، الشره المرضي، بوليميا، نهم الطعام، $ARFID$ / *anorexia, bulimia, binge eating disorder, ARFID, orthorexia*
ADHD	فرط الحركة وتشتت الانتباه	فرط حركة، تشتت انتباه، نقص الانتباه، اضطراب الانتباه، $ADHD$ / *ADHD, attention deficit hyperactivity disorder; ADD*†
Autism	التوحد	التوحد، توحد، اضطراب طيف التوحد، طيف التوحد، أسبرجر، متلازمة أسبرجر / *autism spectrum disorder, Asperger's*
DID	اضطراب الهوية التفارقي	اضطراب الهوية التفارقي، تعدد الشخصيات / *dissociative identity disorder, DID, multiple personality disorder*
Sleep Disorder	اضطراب النوم	الأرق، أرق، انقطاع النفس النومي، فرط النوم، الخدار / *insomnia, sleep apnea, hypersomnia, narcolepsy*
Trichotillomania	اضطراب شد الشعر	شد الشعر، نتف الشعر، نتف الجلد، قضم الأظافر / *trichotillomania, excoriation disorder, onychophagia*
BPD	اضطراب الشخصية الحدية	اضطراب الشخصية الحدية، الحدية، بوردرلاين / *borderline personality disorder, BPD*
Paranoid PD	اضطراب الشخصية الزوراني	اضطراب الشخصية الزوراني، الشخصية البارانوية، البارانويا / *paranoid personality disorder*
Suicidal Ideation	الأفكار الانتحارية	أفكار انتحارية، الفكر الانتحاري / *suicidal thoughts, suicidal ideation;* محاولة انتحار † إيذاء النفس† انتحار† / *suicide attempt*†, *self-harm*†

Condition	Classical (Best)		PLM (Best)	
	Tw	Rd	Tw	Rd
ADHD	.63	.62	.92	.64
Anxiety	.72	.61	—	.71
Autism	.71	.64	.95	.67
Bipolar	.72	.62	—	.63
BPD	.72	.68	—	.55
Depression	.70	.60	.88	.59
Eating dis.	.84	.78	—	—
OCD	.77	.62	—	.60
Panic	.81	.74	—	.71
PTSD	.68	.64	1.0	.47
Schizophrenia	.73	.65	1.0	—
Sleep dis.	.62	.62	—	.74
Suicidal	.70	.64	.91	.65

Table 9: Summary of best F1 scores per condition. Tw = Twitter (CARMA-X), Rd = Reddit (CARMA-R). Full results in Appendix Tables 5 and 6.

Table 10: Universal self-disclosure phrases by type and connector class. **B** = بـ connector; **M** = من connector; **D** = direct (no connector).

Type	Conn.	Phrases
Affliction	B	أنا مصاب / انا مصاب، أنا مصابة، أنا مريض، أنا مريضة، كشخص مصاب
Suffering	M	أعاني من / اعاني من، أنا أعاني من، أنا أتعالج من، أنا أعيش مع، أصبت
Clinical Diagnosis	B	تم تشخيصي، شخصوني، شخصني، شخصتني، شخصت، شخصني الطبيب، شخصني الدكتور، شخصتني الطبيبة، شخصتني الدكتورة، الأخصائي شخصني، الأخصائية شخصتني
Possessive	D	عندي، أنا عندي
Discovery /	B	طلعت مصاب، لقيت نفسي مصاب، اكتشفت إني مصاب، عرفت إن أنا مصاب، اتأكدت إني مصاب، كنت مصاب
Confirmation	D	تشخيصي هو، تشخيصي كان، طلع تشخيصي، طلعت عندي، لقيت نفسي عندي، اكتشفت إن عندي، عرفت إني عندي، اتأكدت إن عندي، بعد التشخيص طلع عندي، التحليل أثبت إن عندي
Doctor Result	B	أكد الطبيب إصابتي، الدكتور أكد إصابتي، تم التأكيد على إصابتي، تم إثبات إصابتي، الدكتور أكد التشخيص، الطبيب أثبت التشخيص
	D	تشخيص الطبيب هو، تشخيص الدكتور هو، تشخيص حالتي هو، تشخيص المرض هو، تم تشخيص حالتي على أنها، التشخيص الذي حصلت عليه هو، طلع إن التشخيص هو، اتضح إن التشخيص كان

Table 11: Dialect-specific self-disclosure phrases. **B** = بـ connector; **M** = من connector; **D** = direct.

Dialect	Conn.	Phrases
Gulf	B	انا مصاب، مبتلي، مبتلية، ابتليت
	D	عندي، فيني، انا متعايش، جاني، جاتني
	M	اعاني من
Levantine	B	انا مصاب
	D	عندي، معي، طلعلي، طلعتلي، صار معي، صارت معي، اجاني، اجاتني
	M	بعاني من، عم بعاني
Egyptian	B	انا مصاب، طلعت مصاب، لقيت نفسي مصاب، اكتشفت إني مصاب، اتأكدت إني مصاب
	D	عندي، معايا، جالي، جاتلي، طلعلي، طلعت عندي، لقيت نفسي عندي، اكتشفت إن عندي، اتأكدت إن عندي، عرفت إني عندي، أنا بواجه مشكلة اسمها
	M	بعاني من، اتشخصت
Maghrebi	B	انا مصاب
	D	عندي، كاين عندي، جاني، جاتني، طلعلي، تشخصت
MSA	B	أنا مصاب، أنا مصابة، شُخِّصت، شخصت
	D	لدي، تشخيصي هو، طلع تشخيصي، أتعايش مع
	M	أعاني من، تم تشخيصي

Table 12: Connector rules: phrase type, connector morpheme, keyword form generated, and example.

Phrase Type	Conn.	Keyword Form	Example
Affliction, Diagnosis, Confirmation	بـ	Base kw: بـ+kw ال-kw: بال+stem English: بـ kw	مصاب باكتئاب مصاب بالاكتئاب مصاب بـ $PTSD$
Suffering	من	Phrase ends in من: bare kw Phrase lacks من: من+kw	أعاني من الاكتئاب عم بعاني من اكتئاب
Possessive, Discovery	Direct	Bare or definite kw	عندي اكتئاب تشخيصي هو الفصام

Table 13: Orthographic variant pairs generated automatically.

Feature	Variant A	Variant B
Hamza on alef	أنا	انا
Hamza in conjunction	إن	ان
Taa marbuta	مصابة	مصابه
Taa marbuta	مريضة	مريضه
Hamza on ا	أعاني	اعاني

Annotator Pair	CARMA-X	CARMA-R
Ann-1 vs. Ann-2	100	100
Ann-1/2 vs. GPT-4.1	99	100
Ann-1/2 vs. Qwen3-235B	100	100
Ann-1/2 vs. Jais-2-8B	93	84
Ann-1/2 vs. LLM Majority	92	84
Ann-1/2 vs. LLM Unanimous	92	84
GPT-4.1 vs. Qwen3-235B	99	100
GPT-4.1 vs. Jais-2-8B	92	84
Qwen3-235B vs. Jais-2-8B	93	84
Fleiss' κ (all)	92	84

Table 14: Effective sample size n for each annotator pair after excluding posts with parse failures from the relevant model(s).