Cross-Lingual and Cross-Genre Transfer for Arabic Schizophrenia Detection in Social Media

Amal Alqahtani [1,2] and Mona Diab[3]
[1]The George Washington University, Washington D.C., USA
[2]King Saud University, Riyadh, KSA
[3]Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
amalqahtani@gwu.edu, mdiab@andrew.cmu.edu

Abstract

Language-based schizophrenia detection has focused predominantly on English, leaving Arabic underrepresented in mental health NLP. We introduce an Arabic schizophrenia social media dataset (228 users, 228 controls; Cohen's $\kappa = 0.76$) and evaluate cross-lingual and cross-genre transfer across five primary settings and four reversed direction counterparts using XLM-RoBERTa-base, mBERT, and MARBERT. In-domain Arabic supervision yields the strongest performance, with MARBERT reaching 0.7706 macro-F1. In contrast, no primary transfer condition satisfies our criterion for robust transfer (macro-F1 > 0.50 and consistent performance across all three encoders). Reversed direction experiments suggest that genre transfer is directional: social to clinical transfer partially succeeds in the Arabic-only setting, whereas clinical to social transfer collapses. Results from translated clinical data further indicate that translation-based evaluation can introduce artifacts specific to encoder and direction. Overall, the findings support the need for native target language and target genre resources for reliable multilingual mental health evaluation.

1 Introduction

Computational analysis of language has become an important approach for studying mental health, including schizophrenia, depression, posttraumatic stress disorder, and related conditions. Schizophrenia is especially relevant to language-based modeling because the condition can involve disruptions in coherence, reference, and discourse organization (Covington et al., 2005). Prior work has shown that computational features extracted from speech, writing, and social media can capture schizophrenia-related signals (Elvevåg et al., 2007; Bedi et al., 2015; Corcoran et al., 2018; Mitchell et al., 2015; Birnbaum et al., 2017; Bae et al., 2021; Sarioglu Kayi et al., 2017).

Despite this progress, language-based mental health detection remains heavily centered on English (Alasmari, 2025; Bucur et al., 2026). This creates a substantial gap for Arabic-speaking populations, where annotated mental health resources are limited and schizophrenia detection has received little attention. Arabic also poses well-known NLP challenges, including rich morphology, dialectal variation across Gulf, Egyptian, Levantine, Maghrebi, and Modern Standard Arabic (MSA) varieties, frequent code switching with English, and limited supervision for this task. These factors have motivated Arabic pretraining efforts such as AraBERT (Antoun et al., 2020) and MARBERT (Abdul-Mageed et al., 2021).

The challenge, however, is not only linguistic. Clinical essays and social media timelines differ in length, register, vocabulary, communicative purpose, topic distribution, and discourse structure. A model trained on English clinical essays may therefore fail on Arabic social media because of language mismatch, genre mismatch, or both. Understanding these sources of degradation is important for multilingual mental health NLP, where translation and cross-lingual transfer are often used as substitutes for native data.

This paper evaluates these dimensions using a language and genre framework with four datasets: Arabic Schizophrenia X (introduced here), English Schizophrenia X (Sarioglu Kayi et al., 2017), English LabWriting (Sarioglu Kayi et al., 2017), and machine-translated Arabic LabWriting. We compare XLM-RoBERTa-base (Conneau et al., 2020), mBERT (Devlin et al., 2019), and MARBERT (Abdul-Mageed et al., 2021), holding model configuration fixed across settings. Our goal is not to identify the optimal architecture for Arabic schizophrenia detection, but to evaluate how language shift, genre shift, transfer direction, and translation artifacts affect generalization.

The paper makes three contributions.

Proceedings of the 1st Workshop on Multimodal, Multilingual, and Multicultural Mental Health and Psychotherapy (MultiPsyche), pages 105–117
November 9, 2026 ©2026 Association for Computational Linguistics

- We introduce Arabic Schizophrenia X, a user-level Arabic social media dataset collected from X using three complementary query strategies and annotated by two native Arabic speakers with substantial inter-annotator agreement ($\kappa = 0.76$).
- We present a controlled evaluation framework for cross-lingual and cross-genre mental health NLP, including both primary and reversed direction transfer settings.
- We show that in-domain Arabic supervision substantially outperforms transfer-based alternatives, that no primary transfer condition is robust across encoders, and that translated corpora can produce artifacts specific to encoder and direction. The results motivate native target language and target genre resources for Arabic mental health NLP.

2 Related Work

Schizophrenia language detection. Computational studies have shown that schizophrenia-related language can be detected from clinical speech, writing, and social media (Elvevåg et al., 2007; Bedi et al., 2015; Corcoran et al., 2018; Mitchell et al., 2015; Birnbaum et al., 2017; Bae et al., 2021; Sarioglu Kayi et al., 2017). These studies establish that language contains predictive signal for schizophrenia-related classification, but they do not determine how reliably such signal transfers across languages, platforms, or genres. That transfer question is the focus of this work.

Arabic mental health NLP and cross-lingual transfer. Arabic mental health NLP remains underexplored relative to English (Alasmari, 2025; Bucur et al., 2026). Existing Arabic resources primarily target depression, suicidal ideation, obsessive-compulsive symptoms, or general mental health questions (Hassib et al., 2022; El-Ramly et al., 2021; Maghraby and Ali, 2022; Al-Haider et al., 2024; Abdulsalam et al., 2024; Baghdadi et al., 2022; Alhuzali et al., 2024). To our knowledge, no dedicated Arabic schizophrenia social media dataset existed prior to this work. Cross-lingual mental health transfer remains limited: Lifelo et al. (2024) study cross-lingual prediction with meta-training and in-context learning, while Skianis et al. (2024) use translated depression data and report variable performance across languages. Our work differs by explicitly separating language, genre, and direction of transfer for schizophrenia detection.

Dataset	Genre	SZ	Ctrl
Arabic Schizophrenia X	Social media	228	228
English Schizophrenia X	Social media	135	135
English LabWriting	Clinical essays	93	95
Arabic LabWriting	Clinical essays	93	95

Table 1: Datasets. SZ = schizophrenia users; Ctrl = controls.

Generalization and translationese. Mental health models trained on one platform often fail to generalize to others (Harrigian et al., 2020), indicating that genre and platform are central evaluation variables rather than nuisance factors. Machine-translated corpora introduce a second concern: translationese artifacts can preserve source-language structure and distort cross-lingual estimates (Vanmassenhove et al., 2021; Artetxe et al., 2020). We therefore treat translated clinical data as a useful but confounded approximation, and explicitly analyze how translation direction interacts with encoder behavior.

3 Datasets

Our evaluation uses four datasets spanning two languages and two genres (Table 1).

3.1 Arabic Schizophrenia X

We collected Arabic tweets from X using three complementary query strategies. The first, *narrative disclosure*, pairs personal narrative anchors (قصتي, تجربتي, حياتي) with condition terms. The second, *diagnosis disclosure*, matches co-occurrence of healthcare provider terms, diagnosis verbs, and schizophrenia keywords. The third, *profile-based disclosure*, identifies schizophrenia-related patterns in user biographies. All strategies share 22 primary keywords covering clinical Arabic, colloquial expressions, English code switching, and transliterations (Appendix C). Retweets were excluded throughout.

The three strategies retrieved 1,541 candidate tweet-profile pairs, independently labeled by two native Arabic speakers following detailed guidelines (Appendix D). Six categories are collapsed to a binary label. Positive categories are *Genuine Personal Disclosure* (the user describes symptoms, treatment, or impact of the condition on their life) and *Help and Treatment Seeking* (the user requests

Category	Count	%
Both agree Positive	508	33.0
Both agree Negative	860	55.8
Disagreement (excluded)	173	11.2
Total pairs	1,541	100
Cohen's κ	0.76	
Unique positive users	251	

Table 2: Inter-annotator agreement on 1,541 tweet-profile pairs.

Statistic	SZ	Ctrl
Users	228	228
Total tweets	22,878	19,226
Mean tweets/user	100.3	84.4
Median tweets/user	120	120
Mean chars/user	9,635	10,333
Median chars/user	9,528	9,539

Table 3: User-level dataset statistics after preprocessing and balancing.

financial assistance, doctor recommendations, or prayers for their own condition). Negative categories are *Humor or Metaphorical Use*, *Media or Professional Content*, *Third Party or Hypothetical*, and *Ambiguous Context*. Annotators considered both the tweet text and the user profile before assigning a label. Cohen's $\kappa = 0.76$ (88.8% agreement; Landis and Koch, 1977), reflecting substantial agreement. We retain users only when both annotators assign a positive label to at least one candidate tweet-profile pair (508 agreed-positive pairs, 251 unique users), maximizing label precision at the cost of recall. Agreed-negative candidate pairs are used to characterize collection noise but are not used as controls. The high proportion of agreed-negative labels (55.8%) reflects the noisiness of keyword-based social media collection (Harrigian et al., 2021).

For each positive user, the full public timeline is retrieved excluding self-disclosure posts and retweets. Controls are sampled from a public Arabic tweet stream, filtered to Arabic profiles and keyword-free. Both cohorts are preprocessed identically. Steps include Arabic-only filtering, URL and mention removal, emoji-to-text conversion, whitespace normalization, and removal of short or duplicate tweets. Users with fewer than 5 tweets are excluded; the dataset is balanced to 228 per class and tweets capped at 120 per user (the control median), yielding 42,104 tweets from 456 users. Users are represented as single [SEP]-concatenated documents (Mitchell et al., 2015; Coppersmith et al., 2015). Table 3 summarises the resulting dataset statistics.

To prevent leakage, classification is performed at the user level so that all text from a user appears in only one split. Self-disclosure and diagnosis-triggering posts are removed before aggregation, and per-user deduplication removes repeated tweets. Transfer experiments train on one dataset and evaluate on another, ensuring no overlap between source and target users.

3.2 Other Datasets

English Schizophrenia X is the Twitter schizophrenia dataset introduced by Sarioglu Kayi et al. (2017), consisting of users with apparently genuine self-reported diagnoses of a schizophrenia-related condition collected via the Twitter API using regular-expression matching on the root *schizo*, with each diagnosis verified by a human annotator. We use a subset of 135 schizophrenia users and 135 matched controls, preprocessed identically to Arabic Schizophrenia X (120-tweet cap, [SEP]-concatenated). **English LabWriting** (Sarioglu Kayi et al., 2017) consists of personal essays by 93 patients and 95 controls on two prompted topics (an average Sunday and what makes them angriest), yielding up to two essays per participant. Essay length statistics are reported in Appendix E; essays average 125 words (634 characters) per text, roughly 15 times shorter in character volume than the social media timelines in Arabic Schizophrenia X. **Arabic LabWriting** is the LabWriting corpus translated by GPT-4o (Hurst, Aaron et al., 2024) using a prompt designed to preserve structural irregularities, fragmentation, and spelling errors rather than normalize them (Appendix B); the quality and characteristics of this translation directly affect performance estimates in experiments involving this corpus.

4 Experimental Design

4.1 Transfer Settings

We define five primary settings, comprising one in-domain monolingual baseline and four transfer experiments, plus four reversed direction counterparts used to test transfer asymmetry (full details in Appendix A). The baseline evaluates Arabic Schizophrenia X in-domain. Exp 1 approximates genre shift within Arabic using machine-translated

clinical text as source; because the source is machine-translated, translation effects cannot be fully disentangled from genre effects. Exp 2 approximates language shift within the clinical genre, but because the target is machine-translated Arabic, translation effects cannot be fully excluded. Exp 3 evaluates language shift within the social media genre. Exp 4 combines both language and genre shift. The four reversed direction experiments swap source and target to test transfer asymmetry; these were planned as part of the evaluation design before results were inspected. A fully balanced 2×2 language and genre factorial would require a native Arabic clinical corpus, which does not exist for schizophrenia. The translationese analysis in Section 5.5 characterises the artifact this gap introduces.

4.2 Model and Training

All experiments use three multilingual encoders, XLM-RoBERTa-base (Conneau et al., 2020), mBERT (Devlin et al., 2019), and MARBERT (Abdul-Mageed et al., 2021). XLM-RoBERTa-base and mBERT are general-purpose multilingual encoders; MARBERT is pretrained specifically on 1 billion Arabic tweets (Abdul-Mageed et al., 2021) and serves as an Arabic-specialized reference point. We use the same fine-tuning configuration across all models and settings, with learning rate 2×10^{-5}, batch size 16, weight decay 0.01, warmup ratio 0.1, 3 epochs, and random seed 42. We hold all hyperparameters fixed so that performance differences reflect the transfer condition and model pretraining rather than tuning differences.

For social media datasets, user timelines exceed the 256-token input window. We introduce **chunk averaging** as our document representation strategy, in which documents are split into overlapping 256-token chunks (stride 128), each classified independently, and probabilities averaged to a user-level prediction. One chunk is sampled per user per training epoch; all chunks are averaged at evaluation. The ablation in Section 5.7 validates this choice against simple truncation.

4.3 Evaluation

The baseline uses 5-fold stratified cross-validation (mean ± std reported). Transfer experiments use a single fixed seed (42). We acknowledge that this creates an asymmetry between the baseline (which has variance estimates from five folds) and transfer results (which are single-run point estimates). Exact transfer gaps should therefore be interpreted as indicative rather than definitive; replication with multiple seeds is noted as a limitation (see the Limitations section).

We report accuracy, macro-F1 (primary metric), and AUC-ROC. We define two evaluation thresholds used throughout the paper.

- **Prediction collapse** (macro-F1 $\leq$ 0.40) denotes performance close to a degenerate single-class predictor. On a perfectly balanced binary dataset, all-positive prediction yields a macro-F1 of exactly 0.333 (positive F1 = 0.667, negative F1 = 0.0). We use 0.40 as a heuristic threshold slightly above this floor; results at or below this value should be interpreted as providing little usable discriminative signal.
- **Robust transfer performance** (macro-F1 $>$ 0.50 and consistent across all three encoders) denotes above-chance performance for the balanced binary task that replicates across model variants rather than appearing only as a single-encoder anomaly.

5 Results

Table 4 reports classification performance across the three encoders and five primary settings. The in-domain Arabic baseline substantially outperforms all transfer conditions. Across the primary transfer settings, most macro-F1 values fall near the degenerate single-class region (0.333–0.40), and no condition satisfies our criterion for robust transfer performance, defined as macro-F1 $>$ 0.50 with consistent above-chance performance across all three encoders.

5.1 In-Domain Arabic Baseline

The in-domain baseline shows that schizophrenia-related linguistic patterns are detectable in Arabic social media when supervision is available in the same language and genre. MARBERT achieves the highest performance (0.7706 ± 0.0222 macro-F1; 0.8644 ± 0.0296 AUC-ROC), followed by mBERT (0.7174 ± 0.0335 macro-F1) and XLM-RoBERTa-base (0.6965 ± 0.0791). This ranking is consistent with the benefit of Arabic Twitter pretraining for in-domain Arabic social media classification. The larger variance for XLM-RoBERTa-base suggests greater sensitivity to fold composition at this dataset scale, which is expected for small mental health corpora (Harrigian et al.,

Exp.	Source	Target	Setting	Acc.	Macro-F1	AUC
XLM-RoBERTa-base						
Baseline	AR Schiz. X	AR Schiz. X	Mono., same genre	0.7038 ± 0.0751	0.6965 ± 0.0791	0.7898 ± 0.0716
Exp 1	AR LabWriting	AR Schiz. X	Cross-genre	0.5000	0.3333	0.3413
Exp 2	EN LabWriting	AR LabWriting	Cross-ling., same genre (clinical)	$0.6541^{\dagger}$	$0.6256^{\dagger}$	$0.7092^{\dagger}$
Exp 3	EN Schiz. X	AR Schiz. X	Cross-ling., same genre (social)	0.5000	0.3333	0.5506
Exp 4	EN LabWriting	AR Schiz. X	Cross-ling., cross-genre	0.3925	0.3919	0.3665
mBERT						
Baseline	AR Schiz. X	AR Schiz. X	Mono., same genre	0.7192 ± 0.0330	0.7174 ± 0.0335	0.7987 ± 0.0347
Exp 1	AR LabWriting	AR Schiz. X	Cross-genre	0.5066	0.3515	0.4948
Exp 2	EN LabWriting	AR LabWriting	Cross-ling., same genre (clinical)	0.5378	0.4845	0.6348
Exp 3	EN Schiz. X	AR Schiz. X	Cross-ling., same genre (social)	0.4035	0.3865	0.3553
Exp 4	EN LabWriting	AR Schiz. X	Cross-ling., cross-genre	0.5066	0.3478	0.5788
MARBERT						
Baseline	AR Schiz. X	AR Schiz. X	Mono., same genre	$\mathbf{0.7742} \pm 0.0220$	$\mathbf{0.7706} \pm 0.0222$	$\mathbf{0.8644} \pm 0.0296$
Exp 1	AR LabWriting	AR Schiz. X	Cross-genre	0.5022	0.3599	0.5693
Exp 2	EN LabWriting	AR LabWriting	Cross-ling., same genre (clinical)	0.4919	0.3297	0.6043
Exp 3	EN Schiz. X	AR Schiz. X	Cross-ling., same genre (social)	0.5000	0.3483	0.3670
Exp 4	EN LabWriting	AR Schiz. X	Cross-ling., cross-genre	0.5000	0.3333	0.5112

Table 4: Classification results for all three encoders. The baseline reports mean $\pm$ standard deviation across 5 folds. Transfer experiments report single-run scores (seed 42). **Bold** marks the overall best result per metric (MARBERT baseline). † marks the only above-chance transfer result (XLM-RoBERTa Exp 2), which does not replicate across encoders and is discussed as a likely translation-related artifact (see Section 5.5). Macro-F1 ≤ 0.40 indicates proximity to the prediction collapse floor (true single-class macro-F1 = 0.333; see Section 4.3). Chance level = 0.50 macro-F1 for the balanced binary task.

2020).

5.2 Cross-Lingual Transfer Within Social Media

Exp 3 evaluates language shift while holding genre constant (EN social $\rightarrow$ AR social). All three encoders perform at or near collapse: 0.3333 for XLM-RoBERTa, 0.3865 for mBERT, and 0.3483 for MARBERT. AUC-ROC values are also weak or inconsistent (0.5506, 0.3553, and 0.3670, respectively), indicating unreliable ranking signal. These results suggest that same-genre cross-lingual transfer does not yield reliable schizophrenia detection under this setup, even when the target-language encoder is Arabic-specialized.

5.3 Cross-Genre Transfer Within Arabic

Exp 1 evaluates genre shift within Arabic (AR clinical $\rightarrow$ AR social), using machine-translated clinical essays as source data. All models again perform near collapse: 0.3333 for XLM-RoBERTa, 0.3515 for mBERT, and 0.3599 for MARBERT. These results show that clinical to social transfer is ineffective in this setting. Because the Arabic clinical source is machine-translated and smaller than the Arabic social media target, the experiment should be interpreted as an approximation of genre shift rather than a clean causal estimate of genre alone.

5.4 Combined Language and Genre Shift

Exp 4 combines language and genre shift (EN clinical $\rightarrow$ AR social). Performance remains at or below the collapse threshold across all encoders: 0.3919 for XLM-RoBERTa, 0.3478 for mBERT, and 0.3333 for MARBERT. The combined-shift condition is not substantially worse than the Arabic clinical to social condition in Exp 1. This pattern is consistent with the hypothesis that either mismatch can severely limit transfer, but the present design does not fully disentangle language, genre, translation, and dataset-size effects.

5.5 Translation-Based Clinical Transfer

Exp 2 evaluates cross-lingual transfer within the clinical genre using machine-translated Arabic LabWriting as the target. XLM-RoBERTa achieves the only above-chance transfer result in the primary experiments (0.6256 macro-F1), while mBERT (0.4845) and MARBERT (0.3297) do not replicate the effect. This lack of cross-encoder consistency means the result does not meet our robustness criterion. One plausible interpretation is that XLM-RoBERTa exploits translation artifacts or English structural traces preserved in the Arabic translation (Artetxe et al., 2020). Another possibility is that XLM-RoBERTa captures a genuine cross-lingual clinical signal that the other encoders miss. The reversed-direction results are consistent

Exp.	Model	Orig. F1	Rev. F1	Δ
Rev-Exp 1 AR Social → EN Social[†]				
	XLM-RoBERTa	0.3333	0.4947	+0.161
	mBERT	0.3865	0.4831	+0.097
	MARBERT	0.3483	0.4095	+0.061
Rev-Exp 2 AR Clinical → EN Clinical[‡]				
	XLM-RoBERTa	**0.6256**	0.4383	−0.187
	mBERT	0.4845	0.4797	−0.005
	MARBERT	0.3297	**0.6073**	+0.278
Rev-Exp 3 AR Social → AR Clinical[§]				
	XLM-RoBERTa	0.3333	0.4717	+0.138
	mBERT	0.3515	**0.5863**	+0.235
	MARBERT	0.3599	**0.5217**	+0.162
Rev-Exp 4 AR Social → EN Clinical[¶]				
	XLM-RoBERTa	0.3919	0.3615	−0.030
	mBERT	0.3478	**0.5837**	+0.236
	MARBERT	0.3333	0.4415	+0.108

[†]Reverses Exp 3 (EN Social→AR Social)
[‡]Reverses Exp 2 (EN Clinical→AR Clinical)
[§]Reverses Exp 1 (AR Clinical→AR Social)
[¶]Reverses Exp 4 (EN Clinical→AR Social)

Table 5: Reversed direction transfer results. Orig. F1 = original direction; Rev. F1 = reversed direction. **Bold** marks above-chance results (F1 $>$ 0.50). Chance level = 0.50 macro-F1.

with the artifact interpretation, although they do not establish it conclusively.

5.6 Reversed Direction Transfer and Asymmetry

To test whether transfer behavior is symmetric, we evaluate the four reversed direction conditions (Rev-Exp 1–4; Appendix A). Table 5 reports the results.

Genre transfer is directional. Rev-Exp 3 reverses the Arabic cross-genre condition and therefore provides the clearest test of genre direction, since no language shift is involved. Clinical-to-social transfer in Exp 1 collapses for all models, whereas social to clinical transfer in Rev-Exp 3 achieves above-chance performance for mBERT (0.5863) and MARBERT (0.5217), with XLM-RoBERTa remaining below chance (0.4717). This suggests that social media timelines may provide more transferable supervision for short clinical essays than the reverse. However, the pattern is not uniformly robust across all encoders or all settings: in Rev-Exp 4, only mBERT remains above chance. We therefore interpret the asymmetry as suggestive evidence rather than a definitive conclusion. Dataset size, document length, topic coverage, and translation effects may all contribute.

Exp.	Model	Chunk Avg F1	Truncation F1
Baseline	XLM-RoBERTa	**0.6965**	0.6416
	mBERT	**0.7174**	0.7134
	MARBERT	**0.7706**	0.7328
Exp 1	XLM-RoBERTa	0.3333	0.3430
	mBERT	0.3515	0.3599
	MARBERT	0.3599	**0.3840**
Exp 2	XLM-RoBERTa	0.6256	0.6256
	mBERT	0.4845	0.4845
	MARBERT	0.3297	0.3297
Exp 3	XLM-RoBERTa	0.3333	0.3333
	mBERT	0.3865	**0.4007**
	MARBERT	0.3483	0.3430
Exp 4	XLM-RoBERTa	0.3919	**0.4452**
	mBERT	**0.3478**	0.3333
	MARBERT	0.3333	0.3333

Table 6: Macro-F1 comparison of chunk averaging versus simple truncation at 256 tokens. **Bold** marks the higher value per row. Exp 2 results are identical because both source and target are short LabWriting documents, so chunk averaging is never activated.

Translation effects are encoder- and direction-dependent. The Exp 2 anomaly mirrors in Rev-Exp 2. XLM-RoBERTa drops from 0.6256 in the original direction to 0.4383 in reverse, while MARBERT rises from 0.3297 to 0.6073. This reversal is consistent with a translationese account: encoders may benefit when the translated side contains structural traces that align with their pretraining or representation biases (Artetxe et al., 2020). The result cautions against interpreting translated-corpus performance as direct evidence of language-independent clinical signal without additional diagnostic analysis.

Same-genre cross-lingual transfer fails in both directions. Rev-Exp 1 reverses the social media cross-lingual condition (AR social → EN social). Although scores are higher than in the original direction for all three encoders, no model exceeds 0.50 macro-F1. Thus, same-genre cross-lingual transfer fails to produce reliable classification performance in either direction under the current protocol.

5.7 Ablation: Chunk Averaging versus Truncation

Table 6 compares chunk averaging with simple 256-token truncation.

Chunk averaging improves the in-domain baseline for XLM-RoBERTa (+5.5 F1) and MARBERT (+3.8 F1), with only a negligible gain for mBERT (+0.4 F1). This indicates that useful in-domain sig-

nal is distributed across user timelines rather than concentrated in the first 256 tokens. In transfer settings, neither representation strategy consistently dominates. This suggests that transfer failures are driven primarily by language, genre, and translation mismatch rather than by the user-document aggregation method.

6 Discussion

Native in-domain supervision is critical. The strongest contrast is between matched Arabic supervision and transfer alternatives: MARBERT reaches 0.7706 macro-F1 in-domain, while no primary transfer condition is robust across encoders. This supports investment in native Arabic mental health datasets rather than using English supervision or translated corpora as substitutes.

Language and genre mismatch both degrade transfer. Cross-lingual social media transfer (Exp 3), Arabic clinical to social transfer (Exp 1), and combined transfer (Exp 4) all produce near-collapse macro-F1. These results show that both mismatches are major obstacles under the present setup. Because the Arabic clinical data are machine-translated and the datasets differ in size, length, and topic structure, we do not claim independent causal effects; rather, each mismatch is associated with severe degradation.

Arabic-specialized pretraining helps in-domain but not primary transfer. MARBERT's in-domain advantage is consistent with Arabic Twitter pretraining, but the same specialization does not improve primary transfer. In-domain fit and cross-lingual transferability therefore appear separable in this task.

Genre transfer appears directional. Clinical-to-social transfer collapses, whereas social to clinical transfer partially succeeds for mBERT and MARBERT in the Arabic-only condition. Longer social media timelines may provide more diverse supervision than short prompted essays, but dataset size, document length, topic coverage, and annotation differences remain plausible alternatives. Future benchmarks should therefore evaluate transfer direction explicitly.

Translated corpora can mislead. The translated clinical experiments show an interaction between encoder and direction: XLM-RoBERTa benefits in the English-to-machine-translated-Arabic direction, whereas MARBERT benefits in reverse. This pattern is consistent with translationese artifacts and cautions against treating translated-corpus performance as direct evidence of genuine cross-lingual clinical generalization without diagnostic analysis.

Aggregation matters in-domain but does not explain transfer failure. Chunk averaging improves in-domain Arabic performance for two encoders, indicating that signal is distributed across user timelines. It does not systematically improve transfer settings, so language, genre, and translation mismatch remain the primary explanations for transfer failure.

7 Conclusion

We evaluated schizophrenia detection across language and genre using Arabic and English social media and clinical datasets. In-domain Arabic social media supervision produced the strongest performance, with MARBERT achieving 0.7706 macro-F1 and 0.8644 AUC-ROC. In contrast, no primary transfer condition achieved robust performance across XLM-RoBERTa-base, mBERT, and MARBERT. Reversed direction experiments suggest that genre transfer is directional, with Arabic social to clinical transfer partially succeeding while clinical to social transfer collapses. Translation-based clinical experiments further show that translated corpora can introduce artifacts specific to encoder and direction. Overall, the results indicate that reliable Arabic schizophrenia detection requires native target-language and target-genre resources, and that multilingual mental health evaluation must explicitly control for genre direction, transfer direction, and translation artifacts.

Limitations

Transfer experiments use a single random seed, so fine-grained differences between transfer settings should be treated as indicative until replicated across seeds. The Arabic clinical condition is machine-translated because no native Arabic schizophrenia clinical corpus is available, leaving translation as a confound. The datasets also differ in size, document length, genre, and topic structure, preventing full causal isolation of language and genre. Finally, Arabic social media labels rely on self-reported disclosures rather than clinical diagnosis, as is common in social media mental health research (Coppersmith et al., 2015; Cohan et al., 2018); demographic, dialectal, and

qualitative error analysis is needed before drawing fairness or deployment conclusions.

Ethics Statement

Mental health classification is high stakes because false positives may stigmatize users, while false negatives may delay support. This study is a methodological evaluation, not a deployment-ready screening system. Future clinical use would require institutional review, privacy-preserving data handling, demographic and dialectal fairness evaluation, and clinical oversight. Raw user text is not released; any dataset release should use access controls, data-use agreements, and safeguards against re-identification.

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.

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.

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.

Ashwag Alasmari. 2025. A scoping review of arabic natural language processing for mental health. In *Healthcare*, volume 13, page 963. MDPI.

Hassan Alhuzali, Ashwag Alasmari, and Hamad Alsaleh. 2024. Mentalqa: An annotated arabic corpus for questions and answers of mental healthcare. *IEEE Access*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684.

Yi Ji Bae, Midan Shim, and Won Hee Lee. 2021. Schizophrenia detection using machine learning approach from social media content. *Sensors*, 21(17):5924.

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.

Gillinder Bedi, Facundo Carrillo, Guillermo A Cecchi, Diego Fernández Slezak, Mariano Sigman, Natália B Mota, Sidarta Ribeiro, Daniel C Javitt, Mauro Copelli, and Cheryl M Corcoran. 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*, 1(1):15030.

Michael L Birnbaum, Sindhu Kiranmai Ernala, Asra F Rizvi, Munmun De Choudhury, and John M Kane. 2017. A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *Journal of medical Internet research*, 19(8):e7956.

Ana-Maria Bucur, Marcos Zampieri, Tharindu Ranasinghe, and Fabio Crestani. 2026. A survey on multilingual mental disorders detection from social media data. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 893–918.

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

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.

Cheryl M Corcoran, Facundo Carrillo, Diego Fernández-Slezak, Gillinder Bedi, Casimir Klim, Daniel C Javitt, Carrie E Bearden, and Guillermo A Cecchi. 2018. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, 17(1):67–75.

Michael A Covington, Congzhou He, Cati Brown, Lorina Naçi, Jonathan T McClain, Bess Sirmon Fjordbak, James Semple, and John Brown. 2005. Schizophrenia and the structure of language: the linguist's view. *Schizophrenia research*, 77(1):85–98.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

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.

Brita Elvevåg, Peter W Foltz, Daniel R Weinberger, and Terry E Goldberg. 2007. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophrenia research*, 93(1-3):304–316.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. Do models of mental health based on social media data generalize? In *Findings of the association for computational linguistics: EMNLP 2020*, pages 3774–3788.

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: Improving Access*, pages 15–24, Online. 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.

Hurst, Aaron et al. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

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

Zita Lifelo, Huansheng Ning, and Sahraoui Dhelim. 2024. Adapting mental health prediction tasks for cross-lingual learning via meta-training and in-context learning with large language model. *arXiv preprint arXiv:2404.09045*.

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

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 11–20.

Efsun Sarioglu Kayi, Mona Diab, Luca Pauselli, Michael Compton, and Glen Coppersmith. 2017. Predictive linguistic features of schizophrenia. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 241–250, Vancouver, Canada. Association for Computational Linguistics.

Konstantinos Skianis, John Pavlopoulos, and A. Seza Doğruöz. 2024. Severity prediction in mental health: Llm-based creation, analysis, evaluation of a novel multilingual dataset. *Preprint*, arXiv:2409.17397.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: Main volume*, pages 2203–2213.

A Experimental Settings

Table 7 lists all nine experimental conditions, comprising the in-domain baseline, four primary transfer experiments, and four reversed direction counterparts.

Exp.	Source	Target	Transfer condition
Primary experiments			
Baseline	Arabic Schizophrenia X	Arabic Schizophrenia X	Monolingual, same genre
Exp 1	Arabic LabWriting	Arabic Schizophrenia X	Cross-genre (Arabic)
Exp 2	English LabWriting	Arabic LabWriting	Cross-lingual, same genre (clinical)
Exp 3	English Schizophrenia X	Arabic Schizophrenia X	Cross-lingual, same genre (social)
Exp 4	English LabWriting	Arabic Schizophrenia X	Cross-lingual, cross-genre
Reversed direction experiments (source and target swapped)			
Rev-Exp 1	Arabic Schizophrenia X	English Schizophrenia X	Cross-lingual, same genre (social)
Rev-Exp 2	Arabic LabWriting	English LabWriting	Cross-lingual, same genre (clinical)
Rev-Exp 3	Arabic Schizophrenia X	Arabic LabWriting	Cross-genre (Arabic)
Rev-Exp 4	Arabic Schizophrenia X	English LabWriting	Cross-lingual, cross-genre

Table 7: Full experimental settings. Exp 1 uses machine-translated Arabic as source, so translation effects cannot be fully disentangled from genre effects in that cell. Exp 2 and Exp 3 vary language while holding genre constant. Exp 4 varies both simultaneously. Reversed direction experiments were planned before results were inspected.

B English-to-Arabic Translation Prompt

The following prompt was used to translate the English LabWriting essays into Arabic. It was designed to preserve clinically meaningful linguistic irregularities rather than normalize the text into fluent Arabic.

SYSTEM ROLE
You are a professional Arabic translator specializing in informal and colloquial English. You will translate short personal essays written by native English speakers, some of whom are patients with schizophrenia or related psychotic disorders. The texts are personal narratives covering two topics, the writer's average Sunday and what makes the writer angriest.

TRANSLATION INSTRUCTIONS

1. **Preserve the original voice and register.** These are informal, first-person narratives. Do not elevate or formalize the language. Use informal Modern Standard Arabic (MSA). Avoid full regional dialect. If the source text uses casual, fragmented, or colloquial English, the Arabic translation should reflect a similarly informal register.
2. **Preserve linguistic irregularities.** Some texts contain spelling errors, run-on sentences, grammatical mistakes, repetition, incomplete thoughts, or incoherent passages. Do not correct, normalize, or smooth these features. Reproduce their structural, syntactic, and stylistic features in Arabic as faithfully as possible, since these elements may carry clinical or analytical significance. Preserve exact repetitions, including word-for-word restatements, without consolidating or varying them.
3. **Preserve emotional tone.** Anger, sadness, confusion, calm, joy, or emotional instability expressed in the original should be conveyed with equivalent emotional weight in Arabic.
4. **Do not add, omit, or interpret.** Translate only what is written. Do not explain, clarify, summarize, soften, intensify, reorganize sentences, or reorder ideas. If a passage is unclear or ambiguous in English, keep it equally unclear or ambiguous in Arabic.
5. **Preserve incomplete sentences.** If a sentence ends mid-thought or is syntactically incomplete, leave it incomplete in Arabic. Do not complete or close the thought.
6. **Handle incoherent passages.** Translate all passages as written, regardless of how fragmented or incoherent they appear. Do not flag, skip, or treat any passage differently based on coherence level. Apply the same word-for-word translation approach to incoherent text as to coherent text.
7. **Handle cultural references carefully.** Proper nouns (e.g., "Rite-Aid," "McDonald's," "MetroCard") should be transliterated. Cultural concepts with no Arabic equivalent should be transliterated and briefly noted with a bracketed translator's note.
8. **Preserve profanity and slang.** Preserve profanity, slang, and offensive language at equivalent intensity in Arabic. Do not censor or soften them.

OUTPUT FORMAT
Return only the Arabic translation. Do not include the original English text unless specifically asked.

INPUT PLACEHOLDER

Here is the text to translate.

C Schizophrenia Query Keywords

Type	Examples
Clinical Arabic	فصام، الفصام، الفصامي
Colloquial Arabic	انفصام الشخصية، انفصام
Arabic transliteration	شيزوفرينيا
English (code switching)	schizophrenia

Table 8: Representative schizophrenia keywords used across all three query strategies. The full set of 22 keywords additionally covers hamza variants, taa marbuta variants, and schizoaffective disorder terms.

D Human Annotation Guidelines

The following guidelines were provided to both annotators in Arabic and English, with worked examples, prior to annotation.

ANNOTATION GUIDELINE PROVIDED TO ANNOTATORS

Purpose
This project aims to build a dataset of Arabic tweets containing expressions related to schizophrenia. The main task is to distinguish tweets that represent genuine disclosures by individuals who suffer from the condition from tweets that use the term "schizophrenia" in other contexts, such as humour, metaphor, news, professional commentary, or third-party reports.

Collection Context
Tweets were collected using search phrases such as أنا مصاب بالفصام (*I have schizophrenia*), أعاني من الفصام (*I suffer from schizophrenia*), and عندي فصام (*I have schizophrenia*), among others. Annotators were instructed to read the full tweet together with the account biography before assigning a label.

Classification System
Six categories map to a binary label.

Positive (Patient / Yes)

1. **Genuine Personal Disclosure.** The user shares a personal experience with the condition, describes symptoms, treatment, duration, or its impact on their life.
   Example. أنا مصاب بالفصام منذ ٩ سنوات والحمد لله الدواء ساعدني (*I have had schizophrenia for nine years, and thankfully the medication helped me.*)
2. **Help and Treatment Seeking.** The user requests financial help, treatment, doctor recommendations, or prayers for their own condition.
   Example. أنا مريض فصام، هل تعرفون دكتور جيد؟ (*I am a schizophrenia patient; do you know a good doctor?*)

Negative (Not Patient / No)

3. **Humor / Metaphorical Use.** The word "schizophrenia" is used for humour, sarcasm, or conflicting feelings in a non-literal way.
   Example. أنا عندي فصام نص يحب القهوة ونص يحب الشاي (*I have schizophrenia: one half likes coffee and one half likes tea.*)
4. **Media / Professional Content.** News articles, awareness campaigns, or tweets from professionals, such as doctors or psychologists.
   Example. هل أنا مصاب بالفصام؟ الطب النفسي يجيبك [link] (*Am I affected by schizophrenia? Psychiatry answers you.*)
5. **Third Party / Hypothetical.** Community accounts relaying anonymous questions, hypothetical self-questioning from anxiety, or explicit denial.
   Example. من الخاص أنا مصاب بالفصام، ما هي حقوقي؟ (*From private messages: I have schizophrenia; what are my rights?*)
6. **Ambiguous Context.** Tweets that cannot be clearly classified due to insufficient context, unclear meaning, or conflicting indicators.

Positive Indicators

- Specific treatment details, such as medication names, dosages, or monthly injections.
- Illness duration, such as "for 10 years" or "since 2013."
- Specific symptoms, such as auditory or visual hallucinations and delusions.
- Requests for help from charities, doctors, public figures, or the wider public.
- The phrase "schizophrenia patient" appearing in the account biography.

Negative Indicators

- Questions such as "Am I schizophrenic?"
- Laughing emojis or clearly joking context alongside a schizophrenia mention.
- External links to articles, awareness pages, or websites.
- Professional accounts, including doctors, specialists, journalists, or institutions.
- Political, sarcastic, or metaphorical use.
- Community accounts relaying anonymous messages, such as "from DMs."

Important Notes

- Read the tweet and account biography together before classifying.
- Context matters: the same phrase may be genuine or metaphorical depending on use.
- When in doubt, choose "Ambiguous Context" (Negative).
- Do not assume. Classify only based on what is written.

E LabWriting Essay Length Statistics

Statistic	EN LabWriting	AR LabWriting
Essays (n)	370	370
Mean words/essay	125	94
Median words/essay	124	93
Mean chars/essay	634	516
Median chars/essay	627	507
Min words/essay	11	9
Max words/essay	255	215
For reference: Arabic Schizophrenia X		
Mean chars/user	≈9,500	

Table 9: LabWriting essay length statistics (computed from the corpus) compared to the mean character volume per user in Arabic Schizophrenia X. Social media timelines are roughly 15 times longer than individual LabWriting essays in character volume.

F Preprocessing Summary

The social media preprocessing pipeline applied the following steps. Arabic-only filtering using the platform language tag. URL, mention, and hashtag-symbol removal while retaining lexical content. Emoji-to-text normalization (e.g., grinning_face) to preserve emotional expressiveness signals. Whitespace normalization. Removal of tweets shorter than 10 characters after cleaning. Per-user de-duplication. Exclusion of users with fewer than 5 remaining tweets. User-level aggregation with [SEP] separators. The Arabic schizophrenia dataset was balanced to 228 schizophrenia users and 228 controls with a per-user tweet cap of 120 after this filtering.