Cross-Lingual Transfer Learning for Mental Health Applications

Javid Alakbarli
George Washington University
Washington, D.C., USA
javidalakbarli@gwu.edu

Saad Mankarious
George Washington University
Washington, D.C., USA
saadm@gwmail.gwu.edu

Chuhui Qiu
George Washington University
Washington, D.C., USA
chqiu@gwmail.gwu.edu

Nikita Ravi
George Washington University
Washington, D.C., USA
nikitaravi@gwmail.gwu.edu

December 11, 2025

Abstract

Mental health disorders represent a global crisis, yet automated detection tools remain predominantly focused on English, leaving linguistically diverse populations underserved. This study investigates the efficacy of Cross-Lingual Transfer Learning (CLTL) in adapting large-scale English mental health models to Arabic and Russian contexts. We propose a sequential transfer learning framework that leverages high-resource English corpora—specifically the *Mindset* and merged Reddit depression datasets—to improve depression detection in the *CARMA* (Arabic) and *VKontakte* (Russian) datasets.

We employed XLM-RoBERTa to conduct two primary experiments: (1) fine-tuning an English-pretrained model for Arabic, and (2) evaluating zero-shot, few-shot, and full-shot adaptation strategies for Russian using Low-Rank Adaptation (LoRA). Our results highlight the nuance of cross-lingual transfer. In the Arabic domain, transfer learning improved accuracy and precision but faced challenges in recalling negative samples compared to a tuned monolingual baseline, suggesting significant cultural or linguistic barriers. Conversely, in the Russian domain, cross-lingual transfer proved critical. A monolingual baseline failed to generalize, exhibiting severe mode collapse (F1-score 0.11 for the control class), whereas our transfer learning approach achieved robust performance (Macro F1 0.77) with as few as 100 target samples, eventually converging to an F1-score of 0.97. These findings

demonstrate that while cross-lingual transfer is a powerful tool for mitigating data scarcity and model collapse, its effectiveness is heavily influenced by the linguistic and cultural distance between source and target domains.

1 Introduction

This project focuses on cross-lingual transfer learning for mental health prediction tasks, specifically examining how knowledge transfers from large English mental health datasets to smaller datasets in low-resource languages such as Russian and Arabic. The research addresses the critical need for multilingual mental health screening tools that overcome language barriers and offer culturally sensitive support across diverse linguistic communities. We focus on Large Language Models for their well-established ability to capture linguistic patterns at scale. We employ finetuning (E1) and prompting (E2) on two datasets, D1 and D2, in Arabic and Russian, respectively, with an intermediary large English corpus from which we transfer knowledge. We curated all datasets from social media due to the availability of annotated samples. Our results highlight the difficulty of transferring mental health knowledge from English to Arabic, which we attribute to the cultural context of Arabic and the stigma of mental health that shape unique linguistic expressions. More specifically, transferring knowledge through finetuning achieves higher recall on a binary classification task but lower accuracy, maintaining the same F1 score as the baseline. For English-to-Russian transfer learning, we observe better knowledge transfer by prompting on the same binary classification task, achieving stronger performance from transfer learning than from the baseline.

1.1 Research Questions

Our goal is to assess the potential of large annotated corpus in English in augmenting mental health analysis in other languages where the availability of such corpora is unavailable. The primary research questions that guide this investigation are as follows:

1. RQ1: How can the knowledge learned from large-scale English mental health datasets be effectively transferred to mental health prediction tasks in low-resource languages?

2. RQ2: What are the optimal cross-lingual transfer learning strategies to maintain predictive accuracy when adapting from high-resource to low-resource language contexts?

3. RQ3: How do linguistic and cultural differences between the source and target languages impact the effectiveness of mental health prediction models?

1.2 Hypothesis of the Research

The central hypothesis of this research is that models pre-trained on large English mental health datasets can be effectively adapted to perform mental health prediction tasks in low-resource languages through appropriate cross-lingual transfer learning techniques. Specifically, we hypothesize that:

1. Cross-lingual language models (such as mBERT and XLM-R) will serve as effective bridges for transferring mental health-related linguistic patterns across languages
2. Fine-tuning strategies that preserve language-specific mental health expressions while leveraging cross-lingual semantic similarities will outperform monolingual approaches in low-resource settings
3. The effectiveness of transfer learning will vary depending on the linguistic similarity between the source and target languages, with closer language families showing better transfer performance

1.3 Motivation

Mental health disorders affect millions of people worldwide, yet the majority of mental health research and resources remain concentrated in English-speaking populations. This creates significant barriers for non-English speakers seeking mental health support through digital platforms.

1.4 Contributions

Table 1.4 shows how each of the team members contributed to the project.

Full Name	**Description**
Javid Alakbarli	Sections 2, 3, 4, 5, 6
Saad Mankarious	Sections 3, 4
Chuhui Qiu	Seciton 4, 5, 6
Nikita Ravi	Section 3

1. **Javid Alakbarli:** I used a merged English Reddit corpus and the XLM-RoBERTa model to develop a cross-lingual transfer learning pipeline for the Russian VKontakte dataset. I implemented zero-shot, few-shot, and full-shot adaptation strategies because the monolingual baseline failed to capture non-depressive signals effectively, resulting in a collapsed model. I obtained results showing a dramatic recovery in performance, where my transfer learning approach improved the control class F1-score from 0.11 (baseline) to 0.77 with just 100 samples.

1.5 GitHub Repository

`https://github.com/fibonacci-2/masterout.git`

2 Related Work

2.1 Cross-Lingual Offensive Language Detection: A Systematic Review (Jiang & Zubiaga, 2023)

Jiang and Zubiaga [1] present the first comprehensive systematic review focusing exclusively on Cross-Lingual Transfer Learning (CLTL) techniques for detection of offensive languages in social media contexts. Their survey analyzes 67 relevant papers and provides a fine-grained taxonomy of CLTL approaches organized into three main transfer categories: instance transfer, feature transfer, and parameter transfer. The study identifies key challenges that include language diversity, dataset scarcity, cultural variation, and modeling limitations in low-resource scenarios.

The authors emphasize that Cross-Lingual Transfer Learning emerges as a promising direction to mitigate challenges associated with data scarcity by leveraging domain knowledge from high-resource to low-resource languages. Their framework provides valuable insights into the cross-lingual workflow, which consists of four stages: ***data preparation, cross-lingual training, model adaptation, and detection*** on target language. This systematic approach to CLTL provides a foundational framework that can be adapted for mental health applications, particularly to understand how offensive language patterns might relate to mental health indicators across linguistic boundaries.

2.2 UniBridge: A Unified Approach to Cross-Lingual Transfer Learning for Low-Resource Languages (Pham et al., 2024)

Pham et al. [2] introduce UniBridge, a framework specifically designed to improve the effectiveness of Cross-Lingual Transfer Learning for low-resource languages. Their approach addresses two fundamental elements of language models: embedding initialization and optimal vocabulary size determination. The study proposes a novel embedding initialization method that leverages both lexical and semantic alignment, combined with an automated vocabulary size optimization algorithm.

The UniBridge framework demonstrates significant improvements in F1-scores across multiple languages and tasks, including Named Entity Recognition (NER), Part-of-Speech (POS) tagging, and Natural Language Inference (NLI). Their methodology includes five key stages: vocabulary size searching using Average Log Probability (ALP), language-specific embedding initialization, model adaptation to new languages, downstream task

training, and multi-source transfer learning. The authors' approach to multi-source transfer learning, where knowledge from multiple source languages is aggregated rather than relying on single-source transfer, provides valuable insights for mental health applications that could benefit from diverse linguistic perspectives.

2.3 A Survey on Multilingual Mental Disorders Detection from Social Media Data (Bucur et al., 2025)

Bucur et al. [3] present the first comprehensive survey specifically focused on mental health disorder detection using multilingual social media data. Their work addresses the critical gap in mental health NLP research, where most existing studies focus exclusively on English data, overlooking important mental health signals present in non-English texts. The survey investigates cultural nuances that influence online language patterns and self-disclosure behaviors across different linguistic communities.

The authors provide a comprehensive catalog of multilingual mental health datasets and identify several key research approaches: *translation-based methods, multilingual approaches using cross-lingual embeddings, and few-shot learning techniques.* They highlight that methods developed for multiple languages simultaneously utilize cross-lingual embeddings and leverage information from resource-rich languages such as English to make predictions on data in languages such as Spanish, Korean, and Chinese. This survey establishes the foundation for understanding how mental health expressions vary across cultures and languages, providing crucial context for developing effective cross-lingual transfer learning approaches.

2.4 Adapting Mental Health Prediction Tasks for Cross-lingual Learning via Meta-Training and In-context Learning (Lifelo et al., 2024)

Lifelo et al. [4] introduce two frameworks designed for cross-lingual transfer and adaptation of mental health prediction tasks in low-resource languages. Their study focuses on African languages, particularly Swahili, addressing prediction tasks on stress, depression, severity of depression, and suicidal ideation. The authors propose both model-agnostic meta-learning (MAML) approaches and large language model-based in-context learning strategies.

Their meta-learning approach achieves significant improvements over traditional fine-tuning methods, outperforming baseline approaches by **18%** in ***macro F1 score*** with ***XLM-R*** and **0.8%** with ***mBERT***. The study also explores in-context learning capabilities of large language models across different cross-lingual prompting approaches, finding

that native language prompts perform better than cross-lingual prompts but less effectively than English prompts.

3 Datasets

To analyze the effectiveness of cross-lingual transfer, we utilized a combination of high-resource English datasets and specific target-language datasets. Table 1 summarizes their characteristics.

Dataset	**Language**	**Description**	**Usage**
Mindset [5]	English	Large-scale, automatically annotated Reddit dataset containing 7 mental health conditions.	Source data for Exp 1 (English → Arabic).
Merged English Depression [7, 8]	English	A composite dataset created by merging and cleaning two distinct Reddit depression repositories.	Source data for Exp 2 (English → Russian).
CARMA [6]	Arabic	Automatically annotated Reddit dataset covering 6 conditions.	Target data for Exp 1.
VKontakte Depression	Russian	Dataset collected from CIS-region social networks, labeled for sentiment and depression.	Target data for Exp 2. [9]

Table 1: Overview of datasets used in this study.

4 Methods

4.1 Data Preparation and Pre-processing

For **Experiment 1**, we filtered the Mindset dataset and the CARMA dataset by removing posts with conditions other than depression. This allows for a binary classification task that can be cross-compared against Experiment 2.

For **Experiment 2**, we constructed a robust English source dataset by merging two publicly available repositories: the *Depression Reddit Cleaned* dataset [7] and the *Reddit Depression Dataset* [8]. This merging process allowed us to aggregate a large-scale corpus of depression-related discourse.

Additionally, given the noise inherent in social media text, we applied a rigorous preprocessing pipeline to all datasets:

- **Filtering:** Posts containing fewer than 10 words were removed to exclude samples with insufficient semantic context.

- **Normalization:** We converted all text to lowercase and removed URLs, non-alphanumeric characters, and platform-specific artifacts.
- **Stopword Removal:** Common stopwords were filtered out to focus the model's attention on content-bearing terms.

4.2 Experiment 1: English to Arabic Transfer

As shown in figure 1, pipeline for implementing transfer learning. Without pre-training, each model follows the top path shown, each training a model from its given input and output features alone.

When the fine-tuning dataset is (1) similar to the pre-training dataset in some underlying pattern and (2) has significantly better quality or quantity than the fine-tuning dataset, it is possible to initialize the finetuned model to that of the pretraing path, a technique known as "knowledge transfer" that potentially helps the desired model's learning. We used **roberta-base** on Mindset dataset (English) and then fine-tuned the trained model to detect depression samples from CARMA (Arabic). Our goal is to beat the performance of the same model on Arabic data only (F1 0.22), as outlined in table 3.

We partitioned the CARMA (Arabic) dataset in 0.8:0.2 fraction for training and validation. For the Mindset (English) dataset, all samples are used for pre-training and none are utilized during validation.

Split	**Mindset** (English)	**CARMA** (Arabic)
Training Set	50,000	16,000
Validation Set	-	4,000
Total	**50,000**	**20,000**

Table 2: Data distribution for Experiment 1.

Experiment	**Dataset**	**Performance (F1)**
English Baseline	Mindset	0.80
Pretrained from English	CARMA	0.58
Arabic preliminary baseline	CARMA	0.22

Table 3: `roberta-base` performance on depression binary classification: (1) pretraining on English corpus; (2) training the model obtained from (1) on Arabic data; (3) training a base model on Arabic data alone (without pretraining on english corpus)

4.3 Experiment 2: English-to-Russian Transfer

This experiment evaluates the transferability of mental health signals from the merged English corpus to the Russian *VKontakte* dataset. The experimental workflow is illus-

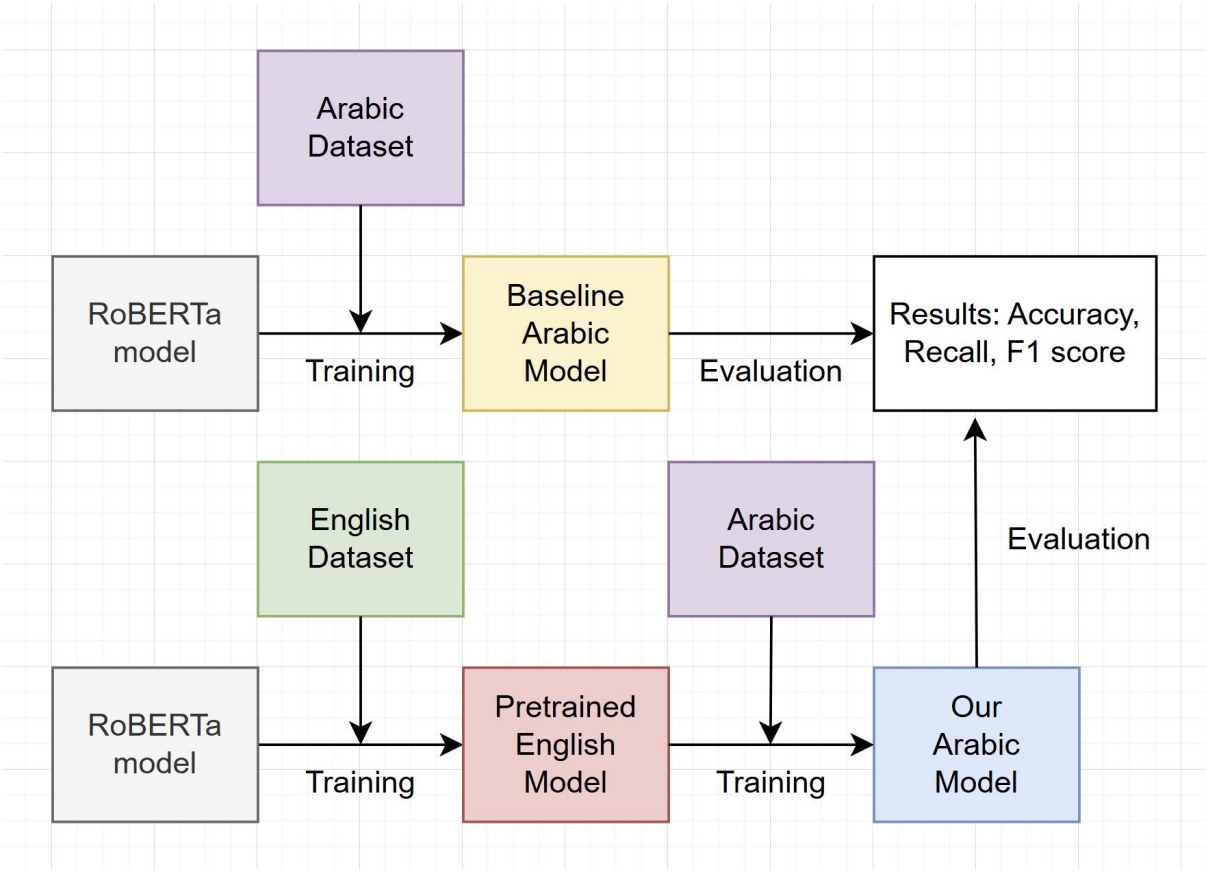

Figure 1: Transfer Learning Pipeline

trated in Figure 2.

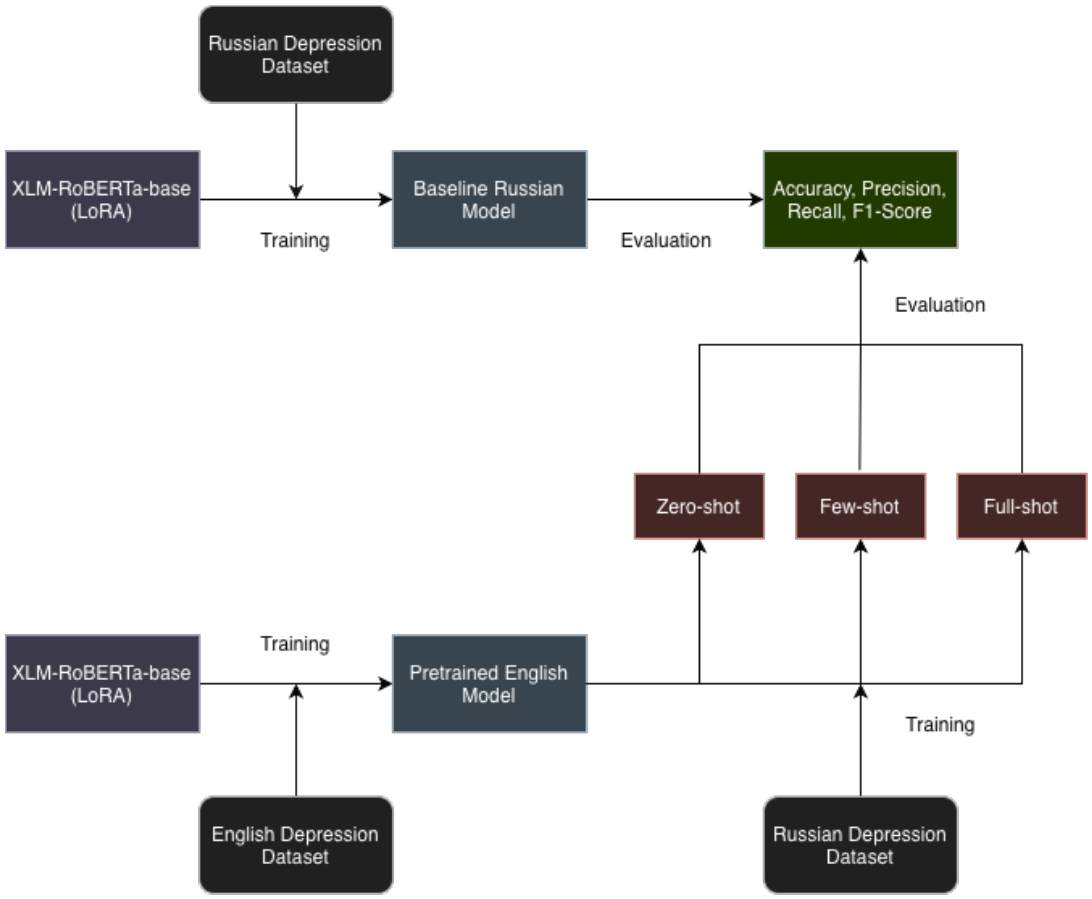

Figure 2: Overview of the experimental pipeline for Experiment 2. The workflow compares a baseline Russian model against a transfer learning approach where an English-pretrained model is adapted to Russian via Zero-shot, Few-shot, and Full-shot strategies using LoRA.

4.3.1 Dataset Splits

To ensure robust evaluation, we partitioned both the merged English source dataset and the target Russian dataset into training, validation, and test sets. Table 4 details the specific sample counts.

Split	**Merged English Source**	**Russian Target (VKontakte)**
Training Set	1,100,850	43,435
Validation Set	235,896	7,665
Test Set	235,897	9,018
Total	**1,572,643**	**60,118**

Table 4: Data distribution for Experiment 2. The English data is a merge of [7] and [8].

4.3.2 Experimental Strategy

We compared a monolingual baseline against a transfer learning approach consisting of two stages:

Stage 1: Source Domain Training. We trained the XLM-R model with LoRA adapters on the merged English dataset. This allowed the model to learn binary depression classification patterns from a massive corpus of over 1.5 million samples.

Stage 2: Target Domain Adaptation. We evaluated the transfer of these English-trained adapters to the Russian dataset using three strategies:

1. **Zero-shot Evaluation:** The model trained on the English corpus is applied directly to the Russian test set.

2. **Few-shot Adaptation:** The model is further fine-tuned on a limited subset of Russian samples (k-shots).

3. **Full-shot Fine-tuning:** The model is fine-tuned on the complete Russian training set to assess initialization benefits.

5 Evaluation & Results

To comprehensively assess model performance, we utilized standard classification metrics: **Precision**, **Recall**, and **F1 Score**.

- **Precision** measures the accuracy of positive predictions, which is crucial to minimize false alarms in mental health screening.

- **Recall** measures the model's ability to identify all relevant cases. High recall is particularly critical for depression detection, as missing a positive case (false negative) can have severe consequences.

- **F1 Score** provides a harmonic mean of precision and recall, serving as the primary metric for overall performance comparison.

5.1 Experiment 1: English-to-Arabic Transfer

We evaluate the performance of our English-to-Arabic transfer strategy using fine-tuning. Table 5 summarizes the results.

Table 5: Experiment 1 Results: Performance comparison of Monolingual Baseline vs. Transfer Learning strategies on the Arabic dataset.

Method	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Baseline	0.5059	0.5030	0.9267	0.6521
Pre-trained	0.5560	0.5507	0.6047	0.5764

5.2 Experiment 2: English-to-Russian Transfer

We evaluated the performance of our English-to-Russian transfer strategy across different data regimes, comparing the monolingual baseline against Few-Shot ($N = 100, 250, 500, 1000$) and Full-Shot adaptation. Table 6 summarizes the results.

Table 6: Experiment 2 Results: Performance comparison of Monolingual Baseline vs. Transfer Learning strategies on the Russian dataset. 'Control' and 'Depression' columns represent F1-Scores.

Method	**Control Class**		**Depression Class**		**Macro Average**	
	Recall	**F1-Score**	**Recall**	**F1-Score**	**Precision**	**F1-Score**
Baseline (Monolingual)	0.07	0.11	0.89	0.62	0.43	0.37
Few-Shot (N=100)	0.69	0.76	0.85	0.78	0.78	0.77
Few-Shot (N=250)	0.85	0.87	0.90	0.87	0.88	0.87
Few-Shot (N=500)	0.93	0.92	0.92	0.92	0.92	0.92
Few-Shot (N=1000)	0.95	0.94	0.93	0.94	0.94	0.94
Full-Shot (Transfer)	**0.95**	**0.97**	**0.99**	**0.97**	**0.97**	**0.97**

The **Baseline model**, trained solely on Russian data without English pre-training, exhibited a severe collapse in the 'Control' class, achieving a recall of only 0.07 and an F1-score of 0.11. In contrast, the **Transfer Learning** approach demonstrated immediate stability. With only $N = 100$ samples (Few-Shot), the model achieved balanced performance (Control F1: 0.76, Depression F1: 0.78), drastically ouprtperforming the baseline.

As the number of few-shot samples increased to $N = 1000$, performance rapidly converged toward Full-Shot metrics, reaching a macro F1-score of 0.94. The Full-Shot transfer model achieved near-perfect classification with F1-scores of 0.97 for both classes.

6 Discussion & Analysis

6.1 Experiment 1 Analysis

Although the use of pretraining significantly improved the result (F1 score: 0.58) over the preliminary result on Arabic data alone (F1 score: 0.22). Upon ablation experiment, we found that many of the hyperparameter tuning improvements equally improved the Arabic baseline (F1 score: 0.65). However, the improved result in the baseline method is due to a strong performance over recall score (0.93 vs. 0.60 in pre-trained) while the pre-trained method offers improved accuracy (0.56 vs. 0.51) and precision (0.55 vs. 0.50). This suggests that transfer learning has positively impacted the model's ability to recognize positive samples, but it has also negatively affected the model's ability to recognize negative samples. In summary, our transfer learning did not show a significant improvement over training directly using Arabic alone, and it hints at a significant linguistic, cultural, or other difference between English and Arabic speakers when posts are analyzed for the purpose of mental health analysis.

6.2 Experiment 2 Analysis

6.2.1 The Failure of Monolingual Baselines

The most striking finding from Experiment 2 is the failure of the monolingual baseline to learn a robust decision boundary. Despite having access to the full Russian training set, the baseline biased heavily toward the 'Depression' class (Recall 0.89) while failing to identify 'Control' samples (Recall 0.07). This suggests that without the "scaffolding" provided by the English source data, the model struggled to differentiate subtle linguistic markers of non-depressive Russian text, likely converging to a trivial solution.

6.2.2 Efficiency of Cross-Lingual Transfer

Our results validate the hypothesis that mental health signals are transferable across languages. The **Few-Shot (N=100)** experiment is particularly revealing: by showing the model just 100 Russian examples, we corrected the massive class imbalance seen in the baseline. This indicates that the English pre-training effectively taught the model *what* depression looks like structurally, requiring only a minimal "vocabulary alignment" step to apply that knowledge to Russian.

6.2.3 Data Scarcity Implications

The trajectory from $N = 100$ to $N = 1000$ demonstrates diminishing returns, suggesting that for low-resource languages, a small, high-quality dataset (< 1000 samples) combined

with English transfer learning is sufficient to build a deployment-ready model (F1 > 0.90). This is a crucial finding for languages where gathering tens of thousands of labeled samples is infeasible.

7 Conclusion & Future Work

This study investigated the efficacy of Cross-Lingual Transfer Learning (CLTL) in democratizing mental health detection tools for low-resource languages. By leveraging large-scale English corpora, we sought to determine if linguistic patterns associated with depression could be effectively transferred to Arabic and Russian contexts using XLM-RoBERTa.

Our findings offer a nuanced answer to our research questions. Regarding **RQ1** (transferability) and **RQ2** (optimal strategies), Experiment 2 demonstrated that transfer learning is not only effective but essential for preventing model collapse in data-scarce scenarios. In the Russian domain, the parameter-efficient adaptation (LoRA) of English representations allowed the model to recover from a failing baseline (Control F1: 0.11) to a robust classifier (Control F1: 0.77) with as few as 100 target samples. This confirms that English pre-training provides critical structural "scaffolding" that monolingual models struggle to learn from small datasets alone.

However, addressing **RQ3** (cultural impact), Experiment 1 revealed the limitations of this approach. In the Arabic domain, while transfer learning improved precision, it degraded the recall of negative samples compared to a tuned baseline. This suggests that the semantic and cultural distance between English and Arabic mental health discourse creates interference, preventing the direct mapping of depression signals.

7.1 Future Work

Future research should focus on three key areas:

1. **Cultural Alignment:** Investigating intermediate pre-training stages using culturally closer languages to bridge the gap between English and distinct linguistic families like Semitic languages.

2. **Qualitative Error Analysis:** Conducting a linguistic analysis of the "false negatives" in the Arabic model to understand specifically which cultural expressions of depression are being lost during transfer.

3. **Multimodal Adaptation:** Extending the LoRA-based few-shot approach to other mental health conditions (e.g., anxiety, PTSD) to verify if the high efficiency observed in the Russian experiment holds across different psychopathologies.

References

[1] A. Jiang and A. Zubiaga, "Cross-lingual Offensive Language Detection: A Systematic Review of Dataset, Approach and Challenge," *Journal of the ACM*, vol. 37, no. 4, Article 127, 2023.

[2] T. Pham, K. M. Le, and A. T. Luu, "UniBridge: A Unified Approach to Cross-Lingual Transfer Learning for Low-Resource Languages," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 3168–3184, 2024.

[3] A. M. Bucur, M. Zampieri, T. Ranasinghe, and F. Crestani, "A Survey on Multilingual Mental Disorders Detection from Social Media Data," *arXiv preprint arXiv:2505.15556*, 2025.

[4] Z. Lifelo, H. Ning, and S. Dhelim, "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*, 2024.

[5] S. Mankarious, A. Zirikly, D. Wiechmann, E. Kerz, E. Kempa, and Y. Qiao, "MindSET: Advancing Mental Health Benchmarking through Large-Scale Social Media Data," *arXiv preprint arXiv:2511.20672*, 2025.

[6] S. Mankarious and A. Zirikly, "CARMA: Comprehensive Automatically-annotated Reddit Mental Health Dataset for Arabic," *arXiv preprint arXiv:2511.03102*, 2025.

[7] InFamousCoder, "Depression: Reddit Dataset (Cleaned)," Kaggle, 2022. [Online]. Available: `https://www.kaggle.com/datasets/infamouscoder/depression-reddit-cleaned`.

[8] R. Kausish, "Reddit Depression Dataset," Kaggle, 2023. [Online]. Available: `https://www.kaggle.com/datasets/rishabhkausish/reddit-depression-dataset`.

[9] B. Omarov, S. Narynov, and D. Mukhtarkhanuly, "Dataset of depressive and suicidal posts," *Mendeley Data*, V1, 2019. doi: 10.17632/838dbcjpxb.1.