A Comparative Study of Using Pre-trained Language Models for Mental Healthcare Q&A Classification in Arabic

Anonymous ACL submission

Abstract

This study explores Pre-trained Language Models (PLMs) for Arabic mental health question answering using the novel MentalQA dataset. We establish a baseline for future research and compare PLMs to classical models. Fine-tuned PLMs outperform classical models, with MARBERT achieving the best results (0.89 F1-score). Few-shot learning with GPT models also shows promise. This work highlights PLMs' potential for Arabic mental health applications while identifying areas for further development.

1 Introduction

Mental health disorders pose a significant global burden, impacting nearly one billion people across demographics (Organization, 2022). Despite its prevalence, access to effective care remains limited, with only half receiving treatment (Consortium et al., 2004). The economic impact is substantial, with mental health costing the global economy trillions annually (Marquez and Saxena, 2016). Natural Language Processing (NLP) offers promising solutions for early intervention and resource allocation (Le Glaz et al., 2021). Recent advancements in Pre-trained Language Models (PLMs) show exceptional performance in NLP tasks (Devlin et al., 2019).

Recent advancements in PLMs have revolutionized applications in various fields, including healthcare (He et al., 2023). However, research on optimizing PLMs specifically for mental health applications remains in its early stages. Integrating PLMs into mental health services offers exciting possibilities for both patients seeking support (Liu et al., 2023; Brocki et al., 2023) and healthcare professionals aiming to improve their services (Sharma et al., 2023). However, PLMs' effectiveness in mental health depends on their ability to understand the nuances of human language including the subjectivity and variability of symptoms, and the need for specialized communication and empathy skills. This challenge is particularly pronounced for languages like Arabic, with its richness, complexity, and vast number of speakers (Guellil et al., 2021).

Despite progress in applying PLMs to mental health in other languages (Atapattu et al., 2022; Kabir et al., 2022; Sun et al., 2021), Arabic remains understudied in this area. A study by (Zhang et al., 2022) highlights a significant imbalance in the availability of mental health datasets across languages. English datasets dominate, making up 81% of the total, followed by Chinese at 10%. Conversely, Arabic datasets are scarce, accounting for a mere 1.5% of available resources.

We explore the potential of pre-trained large language models (PLMs) for Arabic mental health by investigating their effectiveness on a novel question-answering dataset, MentalQA. This dataset is the first of its kind for Arabic and focuses on mental health related interactions. Our work contributes to this domain in three key ways: 1) We conduct the first experiments on MentalQA, establishing a baseline for future research. 2) We perform a comparative analysis between classical machine learning models and PLMs, highlighting their strengths and weaknesses for this specific task. 3) By showcasing the current capabilities and limitations of PLMs in a mental health context, this work aims to promote their further development and refinement for improved mental healthcare applications.

2 Experiments

2.1 Dataset and Task Description

We leverage the MentalQA dataset (Alhuzali et al., 2024) for our analysis. This dataset contains 500 question-answer pairs on mental health from the Altibbi platform (2020-2021). Questions cover di-

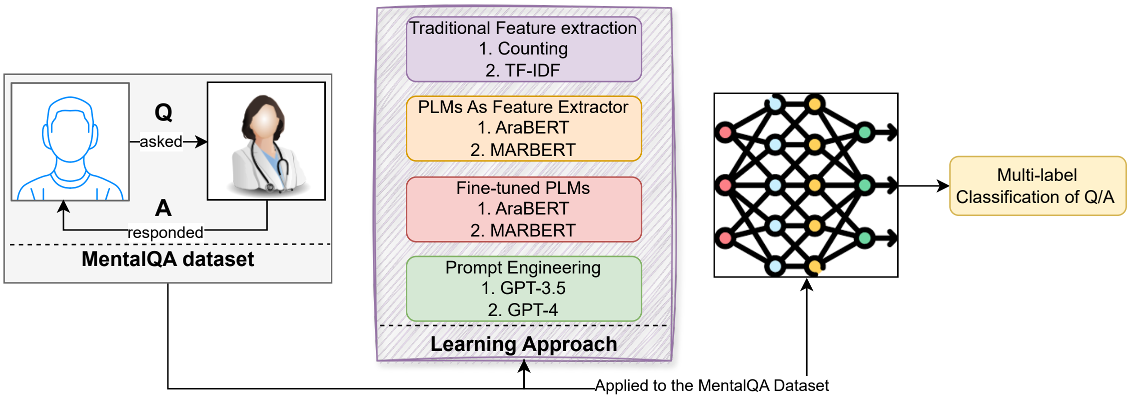

Figure 1: An overview of our experimental design.

agnosis, treatment, etc., while answers offer information, guidance, or emotional support. The MentalQA dataset encompasses two tasks: the classification of question types and answer types. Both tasks allow for the assignment of multiple labels, employing a multi-label classification approach.

2.2 Experimental Setup

We compared four approaches for question-answer multi-label classification on the MentalQA dataset: 1) Feature extraction with SVM: Traditional methods like TF-IDF convert text into numerical features for classification by an SVM. 2) PLMs as feature extractors: Pre-trained PLMs encode text into vectors capturing semantic information, which are then fed into an SVM for classification. 3) Fine-tuning PLMs: We fine-tune PLMs on the MentalQA dataset to improve their performance for this specific task. 4) Prompting Engineering: We explore using GPT-3.5 and GPT-4 with specifically designed prompts to classify questions and answers.

2.3 Implementation Details

We conducted experiments using PyTorch on a T4-GPU (15GB memory). For feature extraction and fine-tuning of PLMs, we used Hugging-Face transformers. Due to resource limitations, we focused on three Arabic PLMs: AraBERT, MARBERT (strong performance in depression detection according to (Guo et al., 2024), and CAMeLBERT-DA. For prompting engineering, we used OpenAI API with GPT-3.5 and GPT-4 variants.

We evaluated models using Micro F1-score, weighted F1-score, and Jaccard index (common metrics for multi-label classification (Alhuzali and Ananiadou, 2021; Mohammad et al., 2018). Since the MentalQA dataset has imbalanced classes, we used the weighted F1-score to account for class distribution.

2.4 Experimental Results

We evaluated four approaches for classifying question and answer types in the MentalQA dataset. Traditional feature extraction with SVM achieved reasonable performance, but PLMs as feature extractors (AraBERT, CAMelBERT, MARBERT) were more competitive, with MARBERT achieving a weighted F1-score of 0.78 for question type classification and 0.82 for answer type classification. Fine-tuning PLMs further improved results, with fine-tuned MARBERT reaching a weighted F1-score of 0.85 for question types and 0.89 for answer types. GPT-3.5 and GPT-4 showed promise in a few-shot learning setting, achieving a 7% improvement in F1-score compared to zero-shot learning for question type classification. These findings highlight the effectiveness of contextualized representations and fine-tuning for this task.

3 Conclusion

This study investigated the effectiveness of PLMs and machine learning models for Arabic mental health question answering using the MentalQA dataset. PLMs like MARBERT exhibited superior performance, suggesting their potential for future mental health applications like intervention services or resource allocation. Additionally, few-shot learning with PLMs showed promise, highlighting further exploration for developing accessible and culturally-sensitive mental health resources for Arabic language.

References

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

Hassan Alhuzali and Sophia Ananiadou. 2021. Spanemo: Casting multi-label emotion classification as span-prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1573–1584.

Thushari Atapattu, Mahen Herath, Charitha Elvitigala, Piyanjali de Zoysa, Kasun Gunawardana, Menasha Thilakaratne, Kasun de Zoysa, and Katrina Falkner. 2022. Emoment: An emotion annotated mental health corpus from two south asian countries. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6991–7001.

Lennart Brocki, George C Dyer, Anna Gładka, and Neo Christopher Chung. 2023. Deep learning mental health dialogue system. In *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 395–398. IEEE.

WHO World Mental Health Survey Consortium et al. 2004. Prevalence, severity, and unmet need for treatment of mental disorders in the world health organization world mental health surveys. *Jama*, 291(21):2581–2590.

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.

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2021. Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*, 33(5):497–507.

Zhijun Guo, Alvina Lai, Johan Hilge Thygesen, Joseph Farrington, Thomas Keen, and Kezhi Li. 2024. Large language model for mental health: A systematic review. *arXiv preprint arXiv:2403.15401*.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2023. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv preprint arXiv:2310.05694*.

Muhammad Khubayeeb Kabir, Maisha Islam, Anika Nahian Binte Kabir, Adiba Haque, and Md Khalilur Rhaman. 2022. Detection of depression severity using bengali social media posts on mental health: Study using natural language processing techniques. *JMIR Formative Research*, 6(9):e36118.

Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C Ryan, Jonathan Marsh, Jordan Devylder, Michel Walter, Sofian Berrouiguet, and Christophe Lemey. 2021. Machine learning and natural language processing in mental health: systematic review. *Journal of medical Internet research*, 23(5):e15708.

June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.

Patricio V Marquez and Shekhar Saxena. 2016. Making mental health a global priority. In *Cerebrum: the Dana forum on brain science*, volume 2016. Dana Foundation.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

World Health Organization. 2022. World mental health report: Transforming mental health for all.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023. Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. Psyqa: A chinese dataset for generating long counseling text for mental health support. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1489–1503.

Tianlin Zhang, Annika M Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: a narrative review. *NPJ digital medicine*, 5(1):46.