MarsadLab at AraHealthQA: Hybrid Contextual–Lexical Fusion with AraBERT for Question and Answer Categorization

Mabrouka Bessghaier[2], Shimaa Amer Ibrahim [2], Md. Rafiul Biswas[1], Wajdi Zaghouani[2]
[1]Hamad Bin Khalifa University, Qatar, [2]Northwestern University in Qatar, Qatar
mbiswas@hbku.edu.qa
{mabrouka.bessghaier,shimaa.ibrahim,wajdi.zaghouani}@northwestern.edu

Abstract

This paper presents the MarsadLab submission to Track 1 of the AraHealthQA 2025 Shared Task, addressing two subtasks: (A) multi-label question categorization and (B) multi-label answer categorization in Arabic mental health discourse. Our approach employs a hybrid contextual–lexical fusion architecture built on AraBERTv2, enriched with task-specific handcrafted features such as lexical indicators, linguistic cues, and domain-informed keyword signals. On the official test set, the system achieved a weighted F1 score of 0.55 (Jaccard 0.41) for Task A and 0.79 (Jaccard 0.67) for Task B.

1 Introduction

Mental health strongly shapes how people think, feel, and function, and untreated conditions such as anxiety, depression, or cognitive disorders can severely reduce quality of life. This growing societal need has motivated the use of computational methods to support mental health understanding and intervention.

Meanwhile, advances in Large Language Models (LLMs) such as GPT-3 (Brown et al., 2020), LAMDA (Thoppilan et al., 2022), PaLM (Chowdhery et al., 2023), and LLaMA (Touvron et al., 2023) have transformed NLP and shown promise in healthcare applications (Sakai and Lam, 2025). Yet their potential remains underexplored in Arabic, particularly for mental health.

Research on Arabic mental health NLP is still in its early stages. For instance, (Mezzi et al., 2022) used BERT-based intent recognition (Devlin et al., 2019) with the MINI framework to diagnose conditions such as depression, suicidality, and panic disorder, achieving nearly 90% accuracy. Nevertheless, benchmarks and resources remain scarce, highlighting the need for community-driven initiatives in this area.

2 Background

The AraHealthQA 2025 Shared Task (Alhuzali et al., 2025) introduces the first benchmark for Arabic medical question answering, with two tracks: Mental Health QA (MentalQA) and General Health QA (MedArabiQ). Our work focuses on MentalQA, specifically **Subtask A: Question Categorization** and **Subtask B: Answer Categorization**. In Task A, the system classifies questions into categories such as Diagnosis or Epidemiology. For example:

هل يعتبر الخوف من عدم الإنجاب مستقبلاً
حالة عادية خاصةً لما أكون متعلقة بأطفال كثيراً
وأنا على وجه جواز أنا خايفة جداً
(Is the fear of not being able to have children in the future normal, especially since I am very attached to children and about to get married?)

The expected output is **A (Diagnosis)** and **D (Epidemiology)**. In Task B, the system instead classifies answers by strategy, such as **1 (Information)** and **2 (Direct Guidance)** for the same example.

The dataset for Track 1 is a newly introduced resource comprising approximately 500 manually annotated Arabic question–answer pairs in the mental health domain (Alhuzali et al., 2024). The data is primarily in Modern Standard Arabic with some dialectal variation, collected from user-generated online content. Each instance may carry multiple overlapping labels, reflecting the complexity of real mental health communication.

Recent efforts in Arabic health-related NLP and conversational AI highlight both the opportunities

Proceedings of The Third Arabic Natural Language Processing Conference, pages 233–238

and challenges for building robust health QA systems. Prior research on Arabic chatbots emphasizes the design of dialog systems for clinical intents, leveraging techniques such as intent classification, NER, and slot filling, while also noting gaps in evaluation protocols, resources, and ethical considerations such as privacy and bias (Ahmed et al., 2022). Parallelly, the fight against health misinformation, particularly during COVID-19, has driven the development of annotation frameworks, credibility signals, and check-worthiness pipelines across multiple languages (Alam et al., 2021; Nakov et al., 2022), providing valuable methodologies for grounding QA in trustworthy evidence. On the mental health side, studies analyzing Arabic social media discourse, such as depression expression (Mohamed and Zaghouani, 2024) and COVID-19-related loneliness (Shurafa and Zaghouani, 2025), demonstrate the feasibility of corpus-driven modeling of emotional and psychological signals, while underscoring ethical concerns around data sensitivity. More recent research has focused on the intersection of Arabic NLP and mental health, including comprehensive surveys of methods and resources (Alasmari, 2025), empirical evaluations of pre-trained language models for Arabic Q/A classification in mental health (Alhuzali and Alasmari, 2025), and applied systems such as the bilingual MindWave app for AI-driven support (Bensalah et al., 2024). Additionally, large-scale evaluations of LLMs in the Arabic mental health domain (Zahran et al., 2025) shed light on both the promise and limitations of current models. The AraHealthQA 2025 shared task is situated within this growing body of work, aiming to foster resources and benchmarks for Arabic mental health and medical QA.

Our contribution lies in integrating transformer-based contextual embeddings with carefully designed task-specific features, tailored to capture the linguistic and psychological nuances of Arabic mental health discourse. Specifically, we employ a hybrid contextual–lexical fusion approach that integrates AraBERTv2 representations with handcrafted lexicon and keyword features. The lexicons are automatically derived from the training data, capturing frequent tokens associated with each category, while the keyword lists are manually curated to reflect pragmatic markers of diagnosis, treatment, guidance, and emotional support. This design allows the model not only to benefit from deep contextual semantics but also to leverage interpretable and domain-relevant signals that are particularly valuable under the low-resource conditions of this shared task.

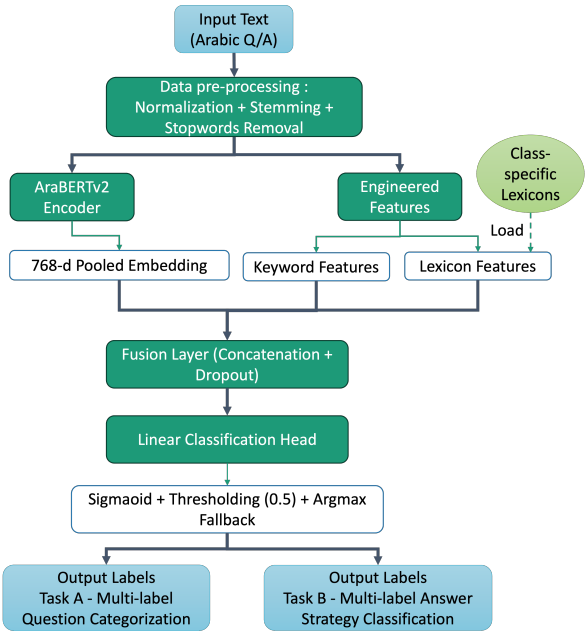

Figure 1: Hybrid Classification Architecture (instantiated separately for Task A and Task B)

3 System Overview

Our approach to both Task A and Task B is based on a hybrid architecture that integrates deep contextual embeddings with handcrafted features. The pipeline consists of three main components: (i) contextual embeddings obtained from AraBERTv2, (ii) handcrafted features capturing lexical, linguistic, and pragmatic information, and (iii) a fusion and classification layer that combines both representations.

As illustrated in Figure 1, the model follows a dual-branch design: AraBERTv2 encodes contextual semantics, while a parallel handcrafted feature block encodes lexical, linguistic, and pragmatic indicators. The two representations are concatenated, regularized with dropout, and passed through a linear classification head. Outputs are produced via sigmoid activation with a 0.5 threshold and an argmax fallback to ensure at least one label. This hybrid pipeline is applied across both tasks, with the same backbone architecture but task-specific lexicons and keyword lists tailored to Question Categorization (Task A) and Answer Strategy Classification (Task B).

3.1 Preprocessing

All input text was normalized (removing diacritics, unifying variants of alif, ya, and taa marbuta), tokenized, cleaned of non-Arabic characters and stopwords, and stemmed with the ISRI stemmer[1] to reduce words to their roots. This preprocessing ensured consistent lexical representations and allowed inflected forms to be collapsed into a single token.

3.2 Features Extraction

We incorporated two types of handcrafted features in parallel to AraBERT embeddings:

Lexicon Features. For each label set, we built lexicons by extracting the top 40 most frequent non-stopword tokens from the preprocessed training data. During feature extraction, these lexicons were used as lookup tables. For each input and each class, we checked whether at least one token from the class lexicon appeared in the text: if true, the feature value was set to 1; otherwise, it was 0. Importantly, this means that multiple overlaps do not increase the score—the feature encodes only the binary presence or absence of a lexicon match. The resulting binary vector was then concatenated with other features and AraBERT embeddings.

Keyword Features. We defined manually curated keyword lists to capture domain-relevant and pragmatic expressions (e.g., definitional markers, directive verbs, supportive phrases). These keywords were stemmed and matched in the input text, with binary features assigned to indicate their presence.

Both lexicon-based and keyword-based features were concatenated with AraBERT embeddings, enabling the model to exploit not only contextual semantics but also interpretable lexical and pragmatic cues.

3.3 Sub-Task A: Multi-label Question Categorization

Step 1: Contextual Embeddings. Each question was encoded with AraBERTv2[2], producing a 768-dimensional pooled embedding.

Step 2: Features Extraction. For Task A, we built class-specific lexicons by extracting the top 40 most frequent non-stopword tokens from the preprocessed, stemmed training questions associated with each label. At inference, The created lexicon is used as a lookup to compute a binary presence feature per class. In parallel, curated keyword lists were defined based on question categories, such as whether the question seeks a diagnosis, treatment, or lifestyle advice. Both lexicon and keyword indicators were concatenated with AraBERT embeddings to enrich the representation of each question. Representative examples are shown in Table 1, with the full lists in Appendix A.

Category	Keywords Examples
A	أعراض (Symptoms)
B	علاج (Treatment), دواء (Medicine)
C	جسم (Body), دماغ (Brain)
D	سبب (Cause), عوامل (Factors)
E	رياضة (Exercise), نوم (Sleep)
F	طبيب (Doctor), مستشفى (Hospital)

Table 1: Example keywords associated with each label category (Task A)

Answer Strategy	Example Keywords
1 : Information	اعراض, تشخيص, معلومة
2 : Direct Guidance	نصيحة, يجب, أنصح
3 : Emotional Support	اطمئن, تقلق, معك

Table 2: Examples of defined keywords for Answer classification (Task B)

Step 3: Fusion and Classification. The AraBERT embeddings and handcrafted features were concatenated, regularized with dropout, and passed through a linear layer. Predictions were obtained via sigmoid activation with a 0.5 threshold and argmax fallback. Besides, category Z (Other) acts as a default class whenever a question does not strongly align with any of the six primary categories (A–F).

3.4 Sub-Task B: Multi-label Answer Strategy Classification

Step 1: Contextual Embeddings. Each answer was encoded with AraBERTv2, producing a 768-dimensional pooled embedding.

Step 2: Features Extraction. For Task B, we built lexicons for each of the three answer strategies (Information, Direct Guidance, Emotional Support)

[1] https://www.nltk.org/_modules/nltk/stem/isri.html

[2] aubmindlab/bert-base-arabertv2

by extracting the top 40 tokens from the training data for each class. As with Task A, lexicon features were computed as binary indicators. Additionally, curated keyword lists were integrated, which capture pragmatic signals such as definitional markers, directive verbs, and supportive expressions. Representative examples are shown in Table 2, with full lists in Appendix B.

Step 3: Fusion and Classification. As in Task A, embeddings and handcrafted features were concatenated, passed through dropout, and classified using a linear layer followed by sigmoid activation with thresholding and argmax fallback.

4 Experimental Setup

Following the AraHealth shared task, we evaluate using weighted F1-score and Jaccard similarity. The model is optimized with binary cross-entropy loss and AdamW (learning rate 1.5). A sigmoid threshold of 0.5 is used to convert probabilities into binary predictions, and an argmax fallback ensures that at least one label is always assigned. Early stopping with patience (3–5 epochs) is applied based on validation loss to prevent overfitting.

5 Results and Discussion

5.1 Main Findings

Task A. The model demonstrates strong performance on categories with salient lexical cues such as Diagnosis and Treatment, while categories characterized by diffuse semantics including Epidemiology and Other present greater classification challenges. Our approach achieves a weighted F1-score of 0.55 and weighted Jaccard similarity of 0.41 on the official test set.

Task B. The classification hierarchy reveals that Information detection yields the highest accuracy, followed by Direct Guidance identification. Notably, Emotional Support frequently exhibits confusion with guidance categories due to overlapping pragmatic markers. The model attains superior performance on the official test set with a weighted F1-score of 0.79 and weighted Jaccard similarity of 0.67.

5.2 Discussion

We conducted comprehensive ablation studies to quantify the contribution of different feature combinations: AraBERT-only baseline, AraBERT+Lexicon features, AraBERT+Keywords, and the AraBERT+Lexicon+Keywords combination. Table 3 presents the detailed results. In addition, we also explored using lexicon and keyword features with a traditional classifier such as SVM to examine their standalone effectiveness outside the AraBERT architecture.

Task A Performance Analysis: The AraBERT baseline achieved F1=0.52 and Jaccard=0.39, with keyword features yielding the strongest gains (F1=0.56, Jaccard=0.43). Lexicon features alone slightly reduced performance, while the combined setup offered balanced improvements (F1=0.55, Jaccard=0.41).To establish a comparative baseline beyond transformer architectures, we implemented a traditional Support Vector Machine utilizing lexicon and keyword features exclusively. This classical approach demonstrated competitive performance with F1=0.45 and Jaccard=0.34, representing a notable accomplishment without the computational overhead of large language models, though performance remained consistently below all AraBERT configurations.

Task B Performance Analysis: The AraBERT baseline delivered strong performance (F1=0.79, Jaccard=0.69), with feature integration yielding minimal changes. Lexicon features slightly reduced performance, while keywords and combined features maintained near-baseline results. The SVM implementation achieved competitive performance with F1=0.74 and Jaccard=0.62, demonstrating effective classification capabilities.

The experimental findings indicate that lexicon features demonstrate optimal effectiveness for categories characterized by stable, domain-specific terminology (e.g., Diagnosis, Treatment), while keyword features exhibit particular strength in supporting pragmatic classification tasks (Direct Guidance, Emotional Support). Notably, Task A reveals that keyword features individually outperform other configurations, suggesting their superior utility for question classification compared to lexicon-based approaches. However, the dataset suffers from severe class imbalance—especially in Task A—which substantially affects overall model performance.

6 Conclusion

We proposed a hybrid architecture that integrates AraBERTv2 contextual embeddings with a compact set of handcrafted features, including lexical indicators, linguistic cues, and domain-informed

Task	Variant	F1	Jaccard
A	AraBERT only	0.52	0.39
	+ Lexicon	0.51	0.37
	+ Keywords	0.56	0.43
	+ Lexicon + Keywords	0.55	0.41
	SVM + features	0.45	0.34
B	AraBERT only	0.79	0.69
	+ Lexicon	0.77	0.67
	+ Keywords	0.78	0.68
	+ Lexicon + Keywords	0.79	0.67
	SVM + features	0.74	0.62

Table 3: Ablation study results on official test set.

keywords. Our findings demonstrate that fusing transformer-based representations with carefully engineered lexical and pragmatic features yields robust performance in both question categorization and answer strategy classification, even on small-scale, domain-specific datasets.

Limitations

Lexicon and keyword features were effective with SVM, but did not improve AraBERT, suggesting the limitation lies in the fusion strategy rather than the features themselves. Another issue is the restricted coverage of the manually defined keywords, which work well for dominant categories but miss diverse expressions; automated expansion or domain-specific terminologies could help. Finally, the dataset size and imbalance remain major challenges: Task A suffers from severe class imbalance with poor recall on minority labels, while Task B shows moderate imbalance, both limiting generalization.

Acknowledgements

This study was supported by the grant NPRP14C-0916-210015, awarded by the Qatar Research, Development and Innovation Council (QRDI).

References

Arfan Ahmed, Nashva Ali, Mahmood Alzubaidi, Wajdi Zaghouani, Alaa Abd-alrazaq, and Mowafa Househ. 2022. Arabic chatbot technologies: A scoping review. *Computer Methods and Programs in Biomedicine Update*, 2:100057.

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. 2021. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 611–649, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A. Alasmari. 2025. A scoping review of arabic natural language processing for mental health. *Healthcare*, 13(9):963.

H. Alhuzali and A. Alasmari. 2025. Pre-trained language models for mental health: An empirical study on arabic q&a classification. *Healthcare*, 13(9):985.

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.

Hassan Alhuzali, Farah Shamout, Muhammad Abdul-Mageed, Chaimae Abouzahir, Mouath Abu-Daoud, Ashwag Alasmari, Walid Al-Eisawi, Renad Al-Monef, Ali Alqahtani, Lama Ayash, Nizar Habash, and Leen Kharouf. 2025. Arahealthqa 2025 shared task description paper. In *Proceedings of ArabicNLP 2025*.

N. Bensalah, H. Ayad, A. Adib, and A. I. El Farouk. 2024. Mindwave app: Leveraging ai for mental health support in english and arabic. In *2024 IEEE 12th International Symposium on Signal, Image, Video and Communications (ISIVC)*, pages 1–6. IEEE.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

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.

Ridha Mezzi, Aymen Yahyaoui, Mohamed Wassim Krir, Wadii Boulila, and Anis Koubaa. 2022. Mental health intent recognition for arabic-speaking patients using the mini international neuropsychiatric interview (mini) and bert model. *Sensors*, 22(3):846.

Ahd Mohamed and Wajdi Zaghouani. 2024. Expression of depression among arab twitter users using arabic corpus analysis. *Procedia Computer Science*, 244:76–85. 6th International Conference on AI in Computational Linguistics.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Julia Maria Struß, Thomas Mandl, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Gautam Kishore Shahi, Hamdy Mubarak, Alex Nikolov, Nikolay Babulkov, Yavuz Selim Kartal, and Javier Beltrán. 2022. The clef-2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection. In *Advances in Information Retrieval*, pages 416–428, Cham. Springer International Publishing.

Hajar Sakai and Sarah S Lam. 2025. Large language models for healthcare text classification: A systematic review. *arXiv preprint arXiv:2503.01159*.

Chereen Shurafa and Wajdi Zaghouani. 2025. Corpus analysis of covid-19 related loneliness on twitter. In *Arabic Language Processing: From Theory to Practice*, pages 80–93, Cham. Springer Nature Switzerland.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

N. Zahran, A. E. Fouda, R. J. Hanafy, and M. E. Fouda. 2025. A comprehensive evaluation of large language models on mental illnesses in arabic context. *arXiv preprint arXiv:2501.06859*.

A Keywords List for Task A

This appendix lists the curated keywords associated with each category label for the subtask A:

Label A: Diagnosis (questions about interpreting clinical findings)

تشخيص, أعراض, علامات, تحليل, نتائج, مؤشرات, توصيف, تقرير, كشف, فحص, اضطراب, حالة, مرض

Label B: Treatment (questions about seeking treatments)

علاج, دواء, جلسات, العلاج, وصفة, معالج, خطة, نفسي, طبيعي, مسكن, مهدئ, مضاد, برنامج

Label C: Anatomy and Physiology (questions about basic medical knowledge)

إفراز, كيمياء, هرمون, أعصاب, عقل, مخ, دماغ, جسم, دوبامين, سيروتونين, فيسيولوجيا, بيولوجيا, خلايا, جهاز, عضو, وظائف, تشريح, أدرينالين, كورتيزول, آلية, بنية, تركيب

Label D: Epidemiology (questions about course, prognosis, and etiology of diseases)

انتشار, عدوى, وراثة, مخاطر, مؤثرات, عوامل, سبب, مآل, توقع, نسبة, إحصاء, احتمال, تفشي

Label E: Healthy Lifestyle (questions related to diet, exercise, and mood control)

نظام, طعام, سهر, نوم, مشي, جري, تمارين, رياضة, تأمل, استرخاء, صحة, تغذية, لياقة, رشاقة, وزن, حمية, يوغا

Label F: Provider Choices (questions seeking recommendations for medical professionals and facilities)

أخصائي, اختصاصي, عيادة, مستشفى, دكتور, طبيب, معالج, مستشار, طوارئ, استشارة, توصية, مركز

B Keywords List for Task B

This appendix lists the curated keywords associated with each category label for the subtask B

Label 1: Information (factual responses)

معلومة, تشير, يعني, تعريف, دراسات, أعراض, سبب, يظهر, علامة, يفسر, موضح, شرح, تشخيص, بيانات, دليل

Label 2: Direct Guidance (action-oriented responses)

عليك, ينصح, افضل, جرب, حاول, ينبغي, يجب, أنصح, نصيحة, اجراء, سلوك, اتبع, خطة, خطوة, لازم, قم

Label 3: Emotional Support (empathy and encouragement)

تفهم, قلب, اطمئن, تقلق, الله, اشعر, وحدك, معك, تهون, مشاعر, اهتم, يهمني, ادعمك, متفهم, احساس, ارتاح