IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 June 2024, accepted 13 July 2024, date of publication 17 July 2024, date of current version 29 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3430068

RESEARCH ARTICLE

MentalQA: An Annotated Arabic Corpus for Questions and Answers of Mental Healthcare

HASSAN ALHUZALI [1], **ASHWAG ALASMARI** [2,3], **AND HAMAD ALSALEH** [4]
[1]College of Computing, Department of Computer Science and Artificial Intelligence, Umm Al-Qura University, Makkah 21421, Saudi Arabia
[2]Department of Computer Science, King Khalid University, Abha 62521, Saudi Arabia
[3]Center for Artificial Intelligence (CAI), King Khalid University, Abha 62521, Saudi Arabia
[4]College of Computer and Information Sciences, Department of Information System, King Saud University, Riyadh 11543, Saudi Arabia

Corresponding authors: Ashwag Alasmari (aasmry@kku.edu.sa) and Hassan Alhuzali (hrhuzali@uqu.edu.sa)

This work was supported by the Deanship of Research and Graduate Studies, King Khalid University, through the Small Group Research under Grant RGP1/337/45.

ABSTRACT Mental health disorders significantly impact people globally, regardless of background, education, or socioeconomic status. However, access to adequate care remains a challenge, particularly for underserved communities with limited resources. Text mining tools offer immense potential to support mental healthcare by assisting professionals in diagnosing and treating patients. This study addresses the scarcity of Arabic mental health resources for developing such tools. We introduce MentalQA, a novel Arabic dataset featuring questioning-answering (QA) style, including a total of 1000 annotations. To ensure data quality, we conducted a rigorous annotation process using a well-defined schema with quality control measures. Data was collected from a question-answering medical platform. The annotation schema for mental health questions and corresponding answers draws upon existing classification schemes with some modifications. Question types encompass six distinct categories: diagnosis, treatment, anatomy & physiology, epidemiology, healthy lifestyle, and provider choice. Answer strategies include information provision, direct guidance, and emotional support. Three experienced annotators collaboratively annotated the data to ensure consistency. Our findings demonstrate high inter-annotator agreement, with Fleiss' Kappa of 0.61 for question types and 0.98 for answer strategies. Our in-depth analysis uncovered insightful patterns in patients behavior and doctor responses. We found a link between the types of questions asked and the strategies doctors use in their answers. Both genders focused on treatment-related questions, though with some variation in emphasis. Interestingly, the types of questions asked also differed by age group. Patients tended to express negative emotions in their questions, while doctors maintained a neutral tone in their responses. Finally, we observed differences in how long it took doctors to answer and the number of words they used depending on the type of question. MentalQA offers a valuable foundation for developing Arabic text mining tools capable of supporting mental health professionals and individuals seeking information.

INDEX TERMS Corpus creation, mental health, natural language processing, question-answering, questions classification.

I. INTRODUCTION

Mental health disorders are highly prevalent worldwide. The impact of mental health can affect individuals regardless of their age, gender, socioeconomic status, or cultural background. According to the World Health Organization, most people do not have access to effective care although one in every eight people in the world experience a mental disorder [1]. This gap results from several factors, including: a shortage of resources, a low number of mental care professionals, inefficient tools and practices in decision-making, social and cultural taboos [2], [3], [4], [5].

Effective communication is the first step towards building a meaningful relationship between doctor and patient. Yet,

The associate editor coordinating the review of this manuscript and approving it for publication was Mostafa M. Fouda.

Language barriers might become significant obstacles on the path to effective diagnosis and care. Multiple studies have reported language as a barrier in various health facilities, which have emerged as an alternative to traditional in-person healthcare and services. Especially, during and after the COVID-19 pandemic, health care has been used for frequent consultation as it saves time, resources, and service consumption [6], [7]. The language barrier exists in both traditional and modern clinical settings which impedes effective delivery of medical services.

The rapid advancement of Artificial Intelligence (AI) and Natural Language Processing (NLP) has also transformed the landscape of disease detection and treatment, including mental illnesses. Innovative tools are being developed to assist mental health care professionals by efficiently reviewing medical history, identifying different patterns, and recommending treatments [8], [9]. There has been a significant surge of interest in the development of text mining tools aimed at providing mental health support [10], [11], [12], [13]. These tools are intended to assist professionals in diagnosing a greater number of patients rather than replacing them. Likewise, by adopting these tools in mental health care, the burden on the health care system can be reduced. However, the development of such tools faces certain limitations, predominantly caused by the scarcity of available datasets. This challenge is particularly pronounced for the Arabic language, let alone in mental health.

Existing datasets on mental health are mostly focused on specific disorders, such as suicidal attempts, self-injury, loneliness, depression, or anxiety. This may limit the ability of AI models to diagnose mental health problems. Examples of those mental health datasets include [14], [15], [16], [17]. More specific datasets are being developed that focus on emotions related to specific mental health issues. For example, the CEASE dataset [18] focuses on the emotions of people who have attempted suicide, while the EmoMent dataset [19] focuses on emotions related to depression and anxiety. Other datasets focus on identifying the level of pain in mental health notes such as [20] or identifying the causal interpretation from mental health notes, such as CAMS dataset by [21].

Despite efforts worldwide to create corpora in other languages [18], [19], [20], [21], [22], [23], [24], the Arabic language is an understudied language regarding mental health disorders. To date, only a handful of works have considered mental health issues in the Arabic language [25], [26], [27]. In particular, Aldhafer and Yakhlef [25] developed depression detection models from Arabic texts on Twitter which focused on the cultural stigma surrounding depression in Arab societies. Another study by Al-Musallam and Al-Abdullatif [26] also focused on the detection of depression in which they applied various machine learning algorithms and feature extraction techniques.

We observe the following gaps in online mental health research. First, there is a lack of research in Arabic language mental health research and detection. Secondly, most research focused on depression detection in Arabic language, whereas other types of mental health were scarcely researched. The third problem is that most Arabic text research has focused on statements made by people in one-way communication and has ignored the types of questions that arise in two-way communication.

The objective of our paper is to create a novel Arabic mental health dataset. The obtained corpus comprises a total of 500 questions and answers (Q&A) posts, including both question types and answer strategies, yielding a total of 1000 annotations. This dataset encompasses interactions, including questions posed by patients and corresponding answers provided by professional doctors. To validate our corpus, we conducted an annotation study following a well-defined annotation schema and employed a quality control process. We also performed extensive analyses to gather evidence on the potential and benefits of our MentalQA dataset. These analyses included correlations between question types and response strategies, an examination of the top frequently used words, an exploration of patient demographics, an analysis of sentiment trends, and an investigation of answering patterns. MentalQA dataset provides valuable resources for constructing effective communication between patients and healthcare providers for mental health support.[1] MentalQA makes significant contributions to both the field of mental healthcare and the NLP community by:

- Richer Data Source: Existing datasets like CEASE [18] and EmoMoment [19] primarily consist of patient-generated content, such as sentences or notes expressing emotions related to suicide attempts or depression. In contrast, MentalQA offers a richer data source, including both questions from patients seeking information or clarification and answers from medical experts.
- Focus on Arabic language: Mental health resources often lack representation in languages other than English. While efforts exist for other languages, Arabic remains understudied in this domain [28]. MentalQA bridges this gap by focusing specifically on Arabic.
- Broader Data Scope: Existing datasets, such as those focusing on pain levels [20] or causal interpretation [21], address specific aspects of mental health information. MentalQA, on the other hand, offers a broader scope. It encompasses a wider range of mental health topics and concerns gleaned from real-world question-and-answer interactions.

The rest of the paper is organized as follows: Section II details the data collection, annotation schema development, and quality control procedures. Section III explores data quality through inter-annotator agreement and analyzes the dataset's characteristics. Section IV discusses the findings and potential applications of MentalQA, acknowledges limitations, proposes future research directions, and explores

[1] The MentalQA dataset is available at https://github.com/hasanhuz/MentalQA

TABLE 1. Data statistics.

Criteria	Statistics
# Questions (Q)	2,621
# Answers (A)	2,621
# Categories (Q)	7
# Categories (A)	3
# Distinct Doctors (A)	84
# Avg length words per (Q)	30
# Avg length words per (A)	31

ethical considerations. Finally, Section V concludes our work.

II. METHODS

This section details the methodology used to construct the MentalQA dataset, a resource for Arabic mental health question answering systems. Figure 1 outlines the key steps involved: the data collection, data annotations, and task definitions which are integral to our approach. First, we collected data from a reputable Arabic medical platform, focusing on mental health questions and answers. Next, an annotation schema was developed to categorize both questions types and answers strategies. Three expert annotators applied the schema, ensuring consistency through collaboration and independent evaluation. The process resulted in the MentalQA dataset, which offers three key tasks for building future question answering systems in the domain of Arabic mental health.

A. DATA COLLECTION

We collected data from the medical platform,[2] which provides reliable, up-to-date, and simplified medical information in Arabic. The website includes thousands of medical articles, a medical glossary, a Q&A section, the latest medical news, and additional health services. For this work, we collected Q&A posts from 2020 to 2021, as well as the most popular Q&A posts (i.e., those that were voted as useful by users). For the data collection, we employed a custom web scraping tool developed using Python to automatically extract question-and-answer pairs. That resulted in a dataset of $53,402$ unique Q&A pairs. We were interested in mental health questions, so we included only Q&A posts in the Mental Health category. This process resulted in a final dataset of $2,621$ questions and answers unique pairs. Table 1 presents detailed statistics about the data, including the number of questions, answers, distinct doctors who responded to questions, categories belonging to both questions and answers, and the average length of words per question and answer.

B. ANNOTATION SCHEMA DEVELOPMENT

We based our annotation schema for mental health questions and their corresponding answers on the first layer of the questions classification schema [29] and answers strategies taxonomy [23], with some modifications. The question classification includes six broad categories, which are described below. It is important to note that the scheme proposed in the study by [29] consisted of seven general categories. However, the results of our annotation study revealed that one of these categories is not applicable to our dataset. Hence, we excluded it from our annotation schema for that reason.

- Diagnosis. Questions about the interpretation of clinical findings, tests, and the criteria and manifestations of diseases.
- Treatment. Questions about seeking treatments, which may include drug therapy, how to use a drug, and the side effects and contraindications of drugs.
- Anatomy and physiology. This category includes important knowledge about basic medicine, such as tissues, organs, and metabolism.
- Epidemiology. Questions in this category are mainly about the course, prognosis, and sequelae of diseases, as well as the etiology and causation of diseases, and the association of risk factors with diseases.
- Healthy lifestyle. Questions are specified to diet, exercise, mood control and other lifestyle factors that can affect health.
- Provider choices. Questions ask for recommendations for hospitals, medical departments, doctors, and the doctor visiting process.
- Other. Questions that do not fall under the above-mentioned categories.

For answer strategies, some strategies were merged into other categories that can be considered interlinked. For example, we merged the answer strategies of Restatement and Interpretation into the Information category, as they are both likely to seek or provide more information. This choice was driven by several factors. During the development of our data annotation scheme, we observed a recurring pattern where doctors would restate a patient's inquiry using their own words and subsequently provide an interpretation. Recognizing this common practice, we concluded that merging these categories was appropriate since both instances involve doctors conveying information. Furthermore, we recognized that this consolidation would minimize potential confusion among annotators, as it simplifies the relationships between labels to some extent. We also renamed the strategy of Approval and Reassurance to "Emotional Support" to make it more inclusive of other types of non-informational support. The Self-disclosure strategy was not applicable due to the nature of our data, as the answers were provided by doctors only. The answers were not evaluated for their completeness or quality with respect to the patient's information needs. The consolidated answer strategies are as follows:

- Information. This category includes answers that provide information, resources, etc. It also includes requests for information.
- Direct Guidance. This category includes answers that provide suggestions, instructions, or advice. It also

[2] Altibbi.com.

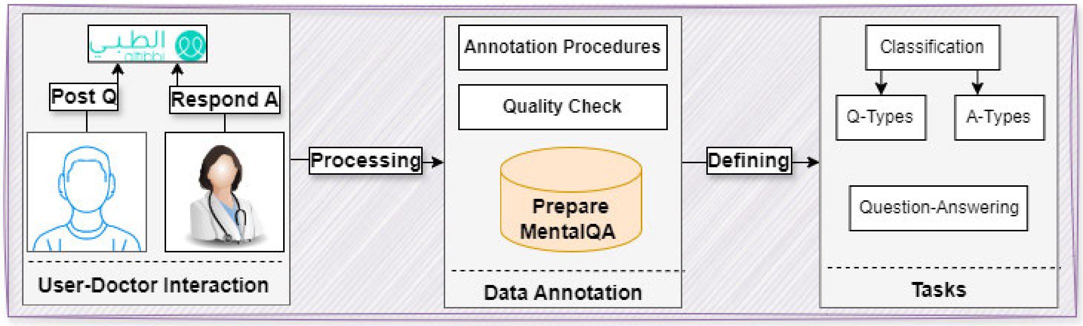


FIGURE 1. An overview of the creation of Arabic MentalQA dataset, starting from data collection, followed by detailed data annotation and the definition of tasks.

includes answers that tell the questioner what they should do to change.

- Emotional Support. This category includes answers that provide approval, reassurance, or other forms of emotional support.

C. ANNOTATION PROCEDURES

The annotation process was performed by three annotators who had experience working with biomedical text and natural language processing. They are also native speakers of Arabic. Detailed instructions outlining the annotation criteria were provided in a dedicated set of guidelines (see Supplementary File 1). The annotation process started with small batches of 20 questions. As the annotators became more familiar with the task, the batch size was gradually increased to 100 questions. The first 20 questions (trial batch) were the same for all annotators, so they could work on the task in parallel. Their annotations were first checked for quality on the trial batch, and annotators were given feedback to help correct them. Once the annotators had demonstrated that they could produce high quality annotations, they were allowed to work on the main annotation rounds. To further illustrate the annotation process, we have included a Microsoft Excel spreadsheet in Supplementary File 2. This spreadsheet served as the annotation tool used by our annotators.

The annotation process for the first 200 questions involved a collaborative approach, where all annotators worked simultaneously on the same set of questions. This initial phase aimed to ensure consistency in annotation practices. Before each new batch of data was assigned, the annotators held group meetings to discuss any disagreements that had arisen and to document the resolutions that were agreed upon. Following this collaborative effort, we evaluated the level of agreement among the annotators. Engaging in such collaborative efforts yields a substantial level of agreement among the annotators. Additionally, such collaboration assists in revisiting and correcting any disagreements that may arise among the annotators. In order to prioritize quality over quantity, the annotators were instructed to work together on the initial 200 data points. This collaborative effort aimed to ensure a thorough examination of the annotation. Once the agreement among the annotators reached a substantial level based on Fleiss' Kappa interpretation, they were then asked to label the remaining 300 data points independently. This sequential process allowed for a comprehensive evaluation of the agreement. The results of this assessment are presented in the Data Quality Check section below.

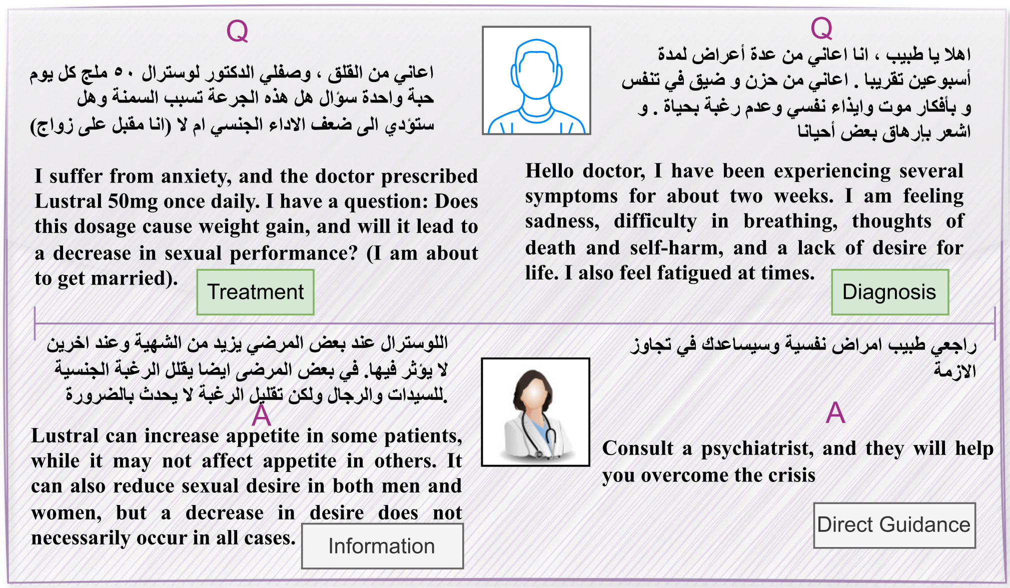
D. MENTALQA TASKS

The newly created dataset, MentalQA, comprises three interconnected tasks essential for question answering and information retrieval in the context of mental health. The first task involves classifying patients' questions into specific types, facilitating a better understanding of their intent. By categorizing questions based on diagnosis, treatment, anatomy, or others, the dataset enables the development of intelligent question-answering systems tailored to patients' needs. Figure 2 presents two examples with their annotation of both questions and answer types.

The second task focuses on classifying answers into specific strategies, ensuring the extraction of relevant information from a pool of responses. Answer strategies such as informational, direct guidance, or emotional support enhance accuracy and relevance, filtering out unreliable information and providing contextually appropriate answers to patients.

The third task involves developing a question-answering style, leveraging the classifications from the previous tasks. This style serves as the backbone for a robust question-answering system, enabling specialized responses to a wide range of mental health-related questions.

In summary, the MentalQA dataset encompasses three tasks: question classification, answer classification, and question-answering with a single, exact answer. Categorizing questions and answers improves the accuracy, relevance, and reliability of question-answering systems, advancing the field of Arabic mental health and facilitating the development of effective information retrieval systems. MentalQA represents a significant step forward in this domain, paving the way for improved question-answering and information retrieval in mental health settings. In future works, we plan to conduct experiments on these tasks to further validate the effectiveness and applicability of the MentalQA dataset. These experiments will contribute to advancing the field of Arabic mental health and refining question-answering and



اعاني من القلق ، وصفلي الدكتور لوسترال ٥٠ ملج كل يوم
حبة واحدة سؤال هل هذه الجرعة تسبب السمنة وهل
ستؤدي الى ضعف الاداء الجنسي ام لا (انا مقبل على زواج)

I suffer from anxiety, and the doctor prescribed Lustral 50mg once daily. I have a question: Does this dosage cause weight gain, and will it lead to a decrease in sexual performance? (I am about to get married).

اهلا يا طبيب ، انا اعاني من عدة أعراض لمدة
أسبوعين تقريبا . اعاني من حزن و ضيق في تنفس
و بأفكار موت وايذاء نفسي وعدم رغبة بحياة . و
اشعر بإرهاق بعض أحيانا

Hello doctor, I have been experiencing several symptoms for about two weeks. I am feeling sadness, difficulty in breathing, thoughts of death and self-harm, and a lack of desire for life. I also feel fatigued at times.

اللوسترال عند بعض المرضي يزيد من الشهية وعند اخرين
لا يؤثر فيها. في بعض المرضى ايضا يقلل الرغبة الجنسية
للسيدات والرجال ولكن تقليل الرغبة لا يحدث بالضرورة.

Lustral can increase appetite in some patients, while it may not affect appetite in others. It can also reduce sexual desire in both men and women, but a decrease in desire does not necessarily occur in all cases.

راجعي طبيب امراض نفسية وسيساعدك في تجاوز
الازمة

Consult a psychiatrist, and they will help you overcome the crisis

FIGURE 2. Example of two annotated Q&A posts, with each Q&A post translated into English for better readability. The first row represents the questions, while the second row represents the corresponding answers. Additionally, the categories for each question and answer are included.

information retrieval systems for improved patient support and care.

III. RESULTS

This section outlines the key findings obtained from analyzing the MentalQA dataset. We assessed inter-annotator agreement, explored the distribution of question types and answer strategies, and investigated the relationship between these elements. Additionally, we analyzed user behavior patterns through gender and age demographics, and explored sentiment within questions and answers. Finally, we examined answering behavior regarding response time, word count, and the language used across different answer strategies.

A. DATA QUALITY CHECK

In all the rounds of annotation mentioned, inter-annotator agreement was computed using the Fleiss' Kappa, a statistical measure for assessing the reliability of agreement between multiple annotators when assigning categorical ratings to a number of items [30]. Table 2 shows the results of our annotation study. The agreement for our annotated Q&A posts is found to be 0.61 and 0.96 for question types and answers strategies, respectively. We also looked at the total number of questions and answers where all annotators agreed on at least one category, as well as the number of questions and answers where two or more annotators agreed on at least one category. The results showed that (96%) and (100%) of the questions and answers, respectively, had at least one category agreed upon by two or more annotators. These findings indicate a substantial level of agreement among annotators in selecting the same category. To generate the final annotations, we employ a majority voting mechanism.

TABLE 2. The findings from our annotation study.

Criteria	Stat.
# Annotated (Q)	500
# Annotated (A)	500
# posts where all three annotators annotate	200
# posts where >= 2 annotators agreed on at least 1 category (Q)	192
# posts where all 3 annotators agreed on at least 1 category (Q)	100
# posts where >= 2 annotators agreed on at least 1 category (A)	200
# posts where all 3 annotators agreed on at least 1 category (A)	152
Fleiss' K (Q)	0.61
Fleiss' K (A)	0.96

The majority of questions (82%) had only one label, followed by (18%) of questions with more than one label. For answers, (69%) had one label compared to (31%) which included more than one label. Table 3 shows the number of questions and answers in each category of question types and answer strategies. According to Table 3, treatment is the most common question type (57%) in the corpus, followed by diagnosis (55%). Then, health lifestyle and epidemiology question types represent (24%) and (22%) of the annotated corpus, respectively. The least represented category in the question types is provider choices (4%). For answers, the most common answer strategy is information (75%), followed by direct guidance (56%) and emotional support (11%).

B. ANALYSIS OF QUESTION TYPES AND ANSWERS STRATEGIES

Figure 3 reveals intriguing insights into the relationship between the question types and answer strategies. We observe

TABLE 3. Number of Q&A posts in each category, where at least two annotators agree on a particular category.

	Categories	Counts
Q-types	Diagnosis (A)	286
	Treatment (B)	296
	Anatomy and physiology (C)	32
	Epidemiology (D)	116
	Healthy lifestyle (E)	125
	Provider choices (F)	20
A-strategies	Information (1)	388
	Direct guidance (2)	290
	Emotional support (3)	61

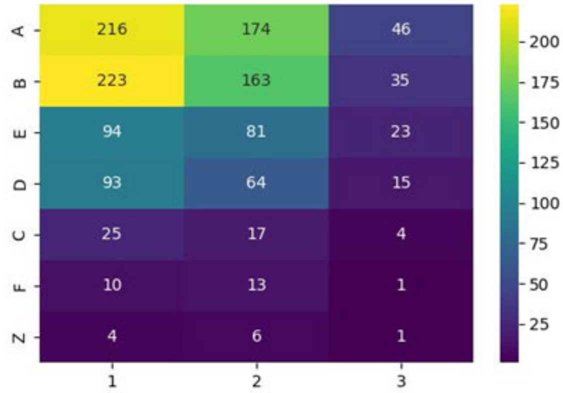


FIGURE 3. Relationship between Q types and A strategies.

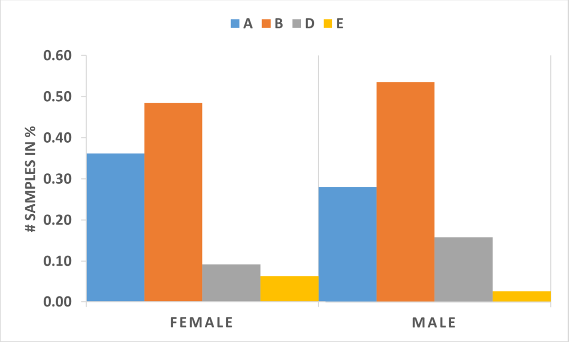


FIGURE 4. The most asked question types by patients' gender.

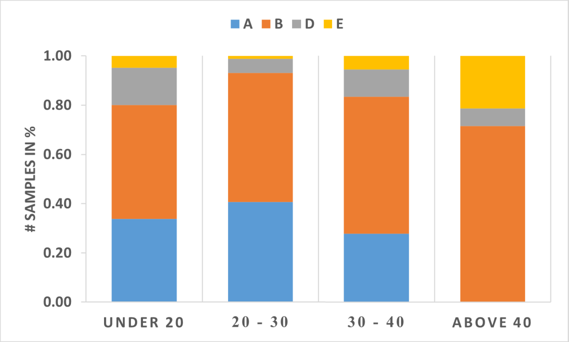


FIGURE 5. The most asked question types by patients' age.

varying degrees of correlation between the categories. For instance, question types A and B demonstrate a strong positive correlation with answer strategies 1 and 2, indicating that these question types are commonly associated with those specific answer strategies. Additionally, question types C, D, and E exhibit moderate positive correlations with answer strategies 1 and 2, indicating a tendency for these question types to be addressed using those strategies. The correlation between answer strategy 3 and question types also demonstrate a noteworthy association. Findings suggest that doctors often leverage emotional support when addressing question type of diagnosis, potentially indicating a tendency to approach the diagnosis process with a compassionate and empathetic approach. By focusing on emotional aspects, doctors may aim to connect with patients.

C. ANALYSIS OF QUESTION TYPES AND GENDER

Our analysis of the most asked questions by patients' gender, as shown in Figure 4, has revealed intriguing patterns in the topics they inquire about. This analysis demonstrates that male patients display a higher frequency of questions pertaining to treatment, with a value of 0.54, indicating a strong inclination towards seeking information and guidance on treatment options. In addition, male patients also exhibit an interest in questions about diagnosis, albeit to a lesser extent, with a value of 0.28, suggesting their desire to understand and clarify medical conditions. Conversely, female patients demonstrate a slightly different pattern in their inquiries. Their most prevalent questions predominantly revolve around treatment, with a value of 0.48, indicating a notable emphasis on exploring various treatment methods and approaches. Furthermore, female patients also exhibit a significant interest in questions about diagnosis, with a value of 0.36, highlighting their proactive approach in seeking information regarding medical evaluations.

D. ANALYSIS OF QUESTION TYPES AND AGE GROUPS

Our analysis of the most asked question types by patients, as illustrated in Figure 5, provides valuable insights into the preferences and inquiries across different age groups. The data reveals a breakdown of question types, represented by categories A, B, D, and E, corresponding to specific age groups. Examining the information presented, we observe distinct patterns among the age groups. Patients under 20 years old exhibit a relatively higher proportion of questions in categories A and B, with values of 0.34 and 0.46, respectively. This suggests a high interest in inquiries related to those question types. Conversely, questions falling under categories D and E show lower values for this age group, indicating a relatively lesser focus on those areas.

Moving to the 20-30 age group, we observe a similar trend. Categories A and B continue to have higher values, with 0.41 and 0.52, respectively, suggesting a sustained interest in those question types. However, there is a notable decrease in the values for categories D and E, indicating a reduced emphasis on those particular question types. For the 30-40 age group, the values for categories A and B decrease further, with 0.28 and 0.56, respectively. This suggests a shift in question preferences compared to the younger age groups.

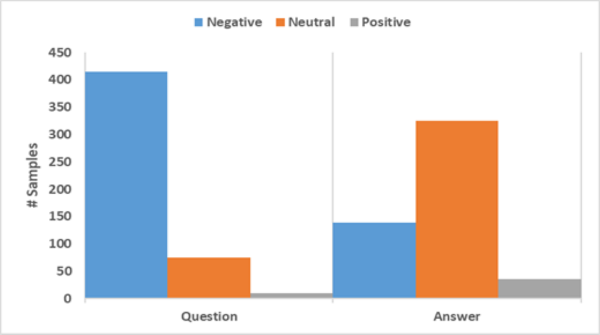


FIGURE 6. The distribution of sentiment across all annotated posts.

Interestingly, there is a slight increase in the values for categories D and E, indicating a relatively higher interest in those areas within this age group. Finally, the above 40 age group demonstrates a distinct pattern. Category A shows a value of 0.00, indicating a lack of questions related to that specific type. However, categories B, D, and E exhibit higher values, with 0.71, 0.07, and 0.21, respectively. This suggests a significant focus on questions falling within those areas, particularly category B and E.

E. SENTIMENT ANALYSIS

We conducted sentiment analysis on the data that yielded intriguing insights as shown in Figure 6. Firstly, the analysis revealed that a significant number of patient questions, including 415 posts, were classified as having a negative sentiment. In contrast, 75 posts were labeled as neutral, and only 10 posts were classified as positive. These findings suggest that patients often express negative emotions when discussing their mental health, which could indicate underlying concerns or challenges they face. Secondly, when examining the responses provided by doctors, the sentiment analysis identified a predominant trend of neutrality. Out of the analyzed answers, 325 were labeled as neutral, while 139 were classified as negative, and 36 as positive. This indicates that doctors tend to maintain a neutral tone when addressing patient inquiries, which aligns with their professional approach. However, it is worth noting that some responses do carry negative emotions, particularly when doctors' express concerns or fear regarding the patient's health, prompting them to visit a doctor as soon as possible. While the sentiment analysis serves as an initial exploration, our findings align with prior research suggesting a tendency for patients to express negative sentiment in healthcare settings, while doctors tend to use more neutral language [31]. It should be mentioned that the results of sentiment were computed based on the ''CAMelBERT model developed by Inoue et al. [32].

F. ANALYSIS OF WORD FREQUENCY

We have conducted an analysis of word frequency in posts, revealing interesting insights into the associations between different question types and the most frequently used words. Table 4 presents the top-15 words that are closely associated with specific question types. This analysis provides valuable information about the common themes and interests expressed by patients when asking different types of questions. For instance, questions related to treatment frequently include words such as ''medicine'', ''treatment'' and ''doctor''. Conversely, questions concerning diagnosis often feature words like ''feeling'', ''suffering'', and ''symptoms''.

G. ANALYSIS OF ANSWERING BEHAVIOR

We also investigated the answering behavior in terms of response time[3] and word frequency. Table 5 shows the average answer time in days and average word counts in answers across different levels of question types in the MentalQA corpus. The results reveal a few interesting patterns. Diagnosis questions (A) are answered quickly (2.61 days) with concise responses (22.33 words). Treatment questions (B) take longer (11.73 days) and have more detailed answers (31.37 words). Notably, questions combining diagnosis and treatment (A, B) take a similar amount of time to answer treatment questions (11.06 days) but require significantly more explanation (47.3 words).

To gain deep insights into the answering behavior in MentalQA corpus, we also employed word clouds, as illustrated in Figure 7, to visualize the most frequently used words using various answer strategies. Overall, the word clouds reveal distinct patterns in the terminology used across the different answer strategies. Answers providing solely information frequently employ terms such as ''treatment'', ''depression'', and ''symptoms''. Conversely, answers providing direct guidance predominantly feature terms like ''doctor'', ''essential'', and ''treatment''. Interestingly, answers offering both information and emotional support exhibit a combination of these terms, along with additional words such as ''God name'', ''no worries'', and ''feel better''.

H. ANALYSIS OF DISAGREEMENT

The insights derived from the results presented in Table 6 shed light on the disagreements among annotators and provide valuable context. The table not only presents the number of samples that experienced full disagreement. Upon analyzing Table 6, we discover that a total of 6 samples exhibited complete disagreement in terms of question types. However, none of the samples displayed complete disagreement concerning the answer strategies. Upon further investigation, we observed that these 6 samples lacked sufficient content to guide annotators in selecting the appropriate labels. To gain additional insights, we examined the word lengths of these samples and found that they ranged from 10 to 15 words, significantly shorter than the average word length reported in Table 1. This additional context deepens our understanding of the disagreement dynamics and highlights the importance of considering the content length and complexity when annotating the data.

[3] It signifies the duration in which the doctor took to respond to patients' inquiries.

TABLE 4. The top 15 words associated with each question's type.

#	A	B	C	D	E
1	I feel	I suffer	Bloating	I suffer	Myself
2	I suffer	Medication	Severe	Severe	I feel
3	When	Sleep	Daily	Anxiety	Thoughts
4	Myself	I feel	Irritable bowel	I feel	Way
5	I am	Myself	And the desire	Thinking	I know
6	Sleep	Treatment	To defecate	Disorder	And feeling
7	People	Depression	More	Speech	When
8	Symptoms	Doctor	4 times	Home	I sleep
9	Depression	Treatment	Per day	Things	I have become
10	Heart	I have	Especially	Personal	Marriage
11	I have	Physician	Work	Eating	Always
12	Condition	Fear	And I suffer	Sleep	Knowledge
13	Problem	Doctor	Sitting	I suffer	People
14	And lack of	Best	Chair	Stress	Situation
15	Concentration	Problem	My stomach	Compulsive	Regret

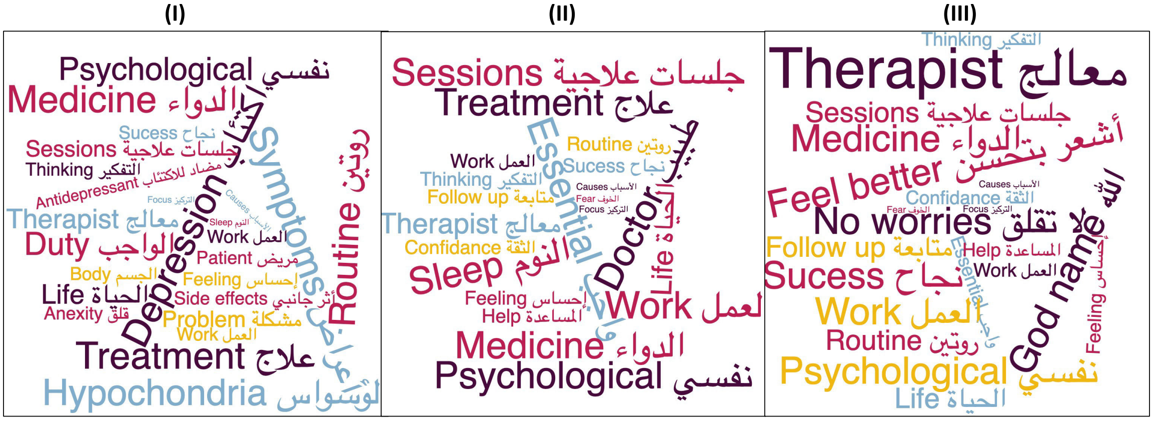


FIGURE 7. Frequently mentioned words in the answers according to different answer strategies where is (I) answers providing only information strategy, (II) answers providing only direct guidance strategy, and (III) answers providing information and emotional support strategy.

TABLE 5. Average answer time measured in number of days and average word frequencies in answers across different levels of question types.

Question types	Avg-answer-time (days)	Avg-word-counts
Diagnosis (A)	2.61	22.33
Treatment (B)	11.73	31.37
Diagnosis & Treatment (A, B)	11.06	47.30
Diagnosis & Epidemiology (A, D)	0.30	49.07
Diagnosis & Healthy Lifestyle (A, E)	11.00	27.3
Treatment & Epidemiology (B, D)	2.30	35.61
Diagnosis, Treatment & Healthy Lifestyle (A, B, E)	2.20	20.90

Table 6 also highlights the top three labels that caused the most confusion among the annotators. Diagnosis emerges as the label with the highest confusion percentage, accounting for 35.31% of the cases, followed by Treatment at 22.65%. Additionally, Healthy lifestyle contributes to the confusion, with percentages of 13.06%. These findings provide valuable insights into the areas of contention and uncertainty within our annotation study, emphasizing the specific labels that require further attention and clarification to enhance the overall study of annotation. The subjectivity of mental health questions can contribute to confusion between diagnosis and treatment labels. For instance, a question like "Is this depression, and how can I solve this problem?" might seem diagnostic, but it could also be seeking clarification on symptoms. By understanding these potential reasons for confusion, we can develop targeted solutions to improve the clarity and consistency of the classification schema in future iterations of the MentalQA dataset.

IV. DISCUSSION

In this study, our primary aim was to develop a comprehensive question-answering mental health dataset in Arabic by utilizing posts gathered from an online platform dedicated to mental health. This dataset holds significant value due to the scarcity of research in the field of Arabic language mental health as discussed in [28]. To facilitate effective data analysis, we devised an annotation scheme that categorized the posts into six distinct question types and three answer strategies. We conducted an annotation study to validate the reliability of our dataset, which revealed substantial agreement among the annotators. This outcome not only confirmed the high quality of the data but also demonstrated the suitability and applicability of the developed annotation scheme within the context of mental health.

TABLE 6. The analysis of disagreement.

Criteria	Stat.
#Samples-Fully Disagreement (Q)	6
#Samples-Fully Disagreement (A)	0
Top 3 Labels That Cause the Most Confusion	%
Diagnosis	35.31%
Treatment	22.65%
Healthy lifestyle	13.06%

Additionally, we performed extensive analyses to gather evidence on the potential and benefits of our MentalQA dataset. The Results section of this paper presents the findings from various analyses conducted. These analyses included correlations between question types and answer strategies, an examination of the top-18 words used, an exploration of patient demographics, an analysis of sentiment trends, and an investigation of answering behavior. These analyses provide valuable insights into patients' concerns when discussing their mental health issues and how healthcare professionals typically respond to their inquiries.

Furthermore, we conducted an analysis of patient demographics to gain insights into their concerns and priorities, providing valuable information on their specific needs and interests. The higher frequency of treatment-related inquiries from gender and age groups suggests a shared interest in understanding and exploring treatment options. However, variations in the emphasis on diagnosis indicate nuanced differences in their information-seeking behaviors. Understanding these trends enables healthcare providers to tailor their services and communication strategies to better address the specific needs and preferences of different patient groups, ultimately enhancing the quality of care and patient satisfaction.

Moreover, we conducted analysis on the top-18 words in the dataset. This analysis can help us to capture the collective voice of patients, providing a deeper understanding of the language patterns and preferences that shape their online interactions. For instance, patients used certain words when expressing their mental health concerns on treatment- and diagnosis-related inquiries, which is aligned with the findings of [33]. Leveraging this knowledge can aid in developing monitoring tools that offer appropriate recommendations to patients and customize content to better meet their needs and interests across various question types. Also, we performed sentiment analysis to understand the emotional dynamics within patient-doctor interactions [33], [34]. The prevalence of negative sentiments expressed by patients emphasizes the need for empathetic and supportive healthcare practices. Likewise, the predominantly neutral responses from doctors highlight their professionalism while acknowledging occasional displays of negative emotions when warranted by the patient's well-being. These findings contribute to a deeper understanding of the emotional aspects involved in healthcare communication and can inform strategies for improving patient experiences and overall care.

Lastly, we conducted an analysis of answering behavior, which yielded valuable insights into the average response time and word count observed in doctors' replies to patient inquiries. For example, treatment-related questions required more time for doctors to respond compared to diagnosis-related questions. Furthermore, questions containing multiple types tended to be longer, as doctors needed to address multiple inquiries conveyed in the patients' posts.

A. IMPLICATIONS

MentalQA has the potential to revolutionize mental healthcare access and communication for Arabic-speaking communities. The dataset provides a valuable foundation for developing several tools and resources. For instance, MentalQA can be used to train Large Language Models (LLMs) and other AI models to understand user intent and deliver appropriate responses in Arabic. This paves the way for the development of chatbots or virtual assistants capable of offering initial mental health support, answering basic questions, and directing users to resources.

Insights from MentalQA can inform the development of text mining tools for mental health research. The analysis of question types, corresponding answer strategies, and the identification of sentiment and emotional patterns within the dataset can be valuable for researchers exploring trends and user behavior in mental health communication.

In addition, insights from MentalQA can inform the development of communication training materials for healthcare professionals. The analysis of question types and corresponding answer strategies can help providers tailor their communication to address patients' specific needs.

Furthermore, MentalQA holds immense potential for advancing Question Answering (QA) systems specifically tailored for mental health support in Arabic. The question-and-answer format of MentalQA closely aligns with the functionalities of these systems. By leveraging MentalQA for training, QA systems can be equipped to understand user queries related to mental health, provide informative responses, and potentially connect users with relevant resources.

While the MentalQA dataset focused on a smaller volume of data, it is important to recognize that the methodologies and procedures outlined in this study are not limited solely to this specific dataset. The steps detailed in this research for data collection and annotation can be readily applied to handle larger datasets. As a result, researchers can adopt the data collection and annotation scheme presented here to effectively manage and process significantly larger datasets, thereby facilitating broader research endeavors in the field.

B. LIMITATIONS

We used textual data sourced from a consumer health platform to provide a valuable resource for identifying question types and answer strategies within a question-answering platform. However, it is important to note that the user-generated nature of the data introduces limitations to the generalizability of our findings. For instance, not all individuals experiencing mental health issues may have access to the internet or may be unable to express their concerns through such platforms. Consequently, researchers and practitioners utilizing our work should exercise caution in interpreting and applying the results.

For this study, we conducted an annotation study on a total of 500 Q&A posts, including both question types and answer strategies, resulting in a total of 1000 annotations. The limited availability of resources prevented us from annotating the entire dataset. However, as part of our future work, we plan to annotate the remaining Q&A posts. The current focus of MentalQA is on Arabic-language interactions. However, we believe the dataset's annotation process and insights can hold promise for future exploration in other languages. Further research is necessary to explore the generalizability of the dataset's findings and its potential adaptation for other languages. This could involve investigating transfer learning techniques or developing similar datasets in other languages following the design principles established for MentalQA.

The current classification schema for question types and answer strategies in MentalQA offers a valuable framework for annotating Arabic mental health conversations. However, it has limitations in capturing the richness and complexity of cultural nuances within Arabic-speaking countries. These variations can significantly influence how mental health concerns are expressed and how help is sought. Future research can address this limitation by incorporating annotations that capture cultural aspects of the questions and answers. This could involve labeling the emotional expression style, help-seeking preference, or indirect phrasing used in questions.

C. ETHICAL CONSIDERATIONS

To ensure the ethical integrity of our research, we implemented rigorous protocols for data collection from the online platform. Our primary focus was to safeguard the privacy and security of personal data, adhering to strict ethical guidelines. We took extensive measures to anonymize the data, removing any personally identifiable information, and implemented robust data security measures. We are confident that our research will not have any negative ethical implications. However, it is important to note that our analyses may provide valuable insights into the nature of patient inquiries and doctors' responses, allowing for a better understanding of healthcare interactions and potentially leading to improvements in patient care and support. In addition, it's crucial to emphasize that MentalQA can be designed to be assistive and not a replacement for healthcare professionals. While AI models to be trained on this dataset can learn from real-world interactions, they cannot independently diagnose or treat mental health conditions.

V. CONCLUSION

We have created a novel Arabic mental health dataset. This dataset comprises interactions, which include questions posed by patients and corresponding answers provided by professional doctors. This two-way communication aspect adds immense value to understanding how to effectively support individuals affected by mental health disorders. To ensure the quality and reliability of the dataset, we conducted an annotation study following a well-defined annotation schema and used a quality control process. We also discussed the results of the annotation study and presented statistics outlining the distribution of categories within our dataset. We further included extensive analyses that demonstrate the potential and benefits of our data.

In the future, we aim to expand the annotation study to cover the entirety of the data, including the annotation of both questions and answers. In order to annotate the entire dataset, there are two potential approaches that can be employed. The first approach, as demonstrated in this study, involves human annotation, where individuals manually label the data. Alternatively, machine learning techniques, specifically weakly-supervised and semi-supervised learning, can be utilized to achieve the task. These techniques leverage both labeled and unlabeled data to enhance the annotation process. Our future goal is to leverage the created dataset to design powerful text mining tools that can make a significant impact in the field of mental health.

ACKNOWLEDGMENT

The authors would like to thank the use of ChatGPT, an AI chatbot developed by OpenAI, for supplementing their own writing, and not to replace them. They verified the accuracy and relevance of the AI-generated text before incorporating it into their manuscript.

REFERENCES

[1] Accessed: Apr. 13, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-disorders

[2] M. Zolezzi, M. Alamri, S. Shaar, and D. Rainkie, "Stigma associated with mental illness and its treatment in the Arab culture: A systematic review," *Int. J. Social Psychiatry*, vol. 64, no. 6, pp. 597–609, Sep. 2018.

[3] A. M. Kilbourne, K. Beck, B. Spaeth-Rublee, P. Ramanuj, R. W. O'Brien, N. Tomoyasu, and H. A. Pincus, "Measuring and improving the quality of mental health care: A global perspective," *World Psychiatry*, vol. 17, no. 1, pp. 30–38, Feb. 2018.

[4] J. A. Hoffmann, M. M. Attridge, M. S. Carroll, N.-J.-E. Simon, A. F. Beck, and E. R. Alpern, "Association of youth suicides and county-level mental health professional shortage areas in the U.S.," *JAMA Pediatrics*, vol. 177, no. 1, p. 71, Jan. 2023.

[5] I. Petersen, A. Bhana, L. R. Fairall, O. Selohilwe, T. Kathree, E. C. Baron, S. D. Rathod, and C. Lund, "Evaluation of a collaborative care model for integrated primary care of common mental disorders comorbid with chronic conditions in South Africa," *BMC Psychiatry*, vol. 19, no. 1, pp. 1–11, Dec. 2019.

[6] H. Al Shamsi, A. G. Almutairi, S. Al Mashrafi, and T. Al Kalbani, "Implications of language barriers for healthcare: A systematic review," *Oman Med. J.*, vol. 35, no. 2, p. e122, Mar. 2020.

[7] H. Alhuzali, T. Zhang, and S. Ananiadou, "Emotions and topics expressed on Twitter during the COVID-19 pandemic in the United Kingdom: Comparative geolocation and text mining analysis," *J. Med. Internet Res.*, vol. 24, no. 10, Oct. 2022, Art. no. e40323.

[8] S. Tutun, M. E. Johnson, A. Ahmed, A. Albizri, S. Irgil, I. Yesilkaya, E. N. Ucar, T. Sengun, and A. Harfouche, "An AI-based decision support system for predicting mental health disorders," *Inf. Syst. Frontiers*, vol. 25, no. 3, pp. 1261–1276, Jun. 2023.

[9] A. C. van Heerden, J. R. Pozuelo, and B. A. Kohrt, "Global mental health services and the impact of artificial intelligence–powered large language models," *JAMA Psychiatry*, vol. 80, no. 7, p. 662, Jul. 2023.

[10] T. Zhang, K. Yang, H. Alhuzali, B. Liu, and S. Ananiadou, "PHQ-aware depressive symptoms identification with similarity contrastive learning on social media," *Inf. Process. Manage.*, vol. 60, no. 5, Sep. 2023, Art. no. 103417.

[11] A. Le Glaz, Y. Haralambous, D.-H. Kim-Dufor, P. Lenca, R. Billot, T. C. Ryan, J. Marsh, J. DeVylder, M. Walter, S. Berrouiguet, and C. Lemey, "Machine learning and natural language processing in mental health: Systematic review," *J. Med. Internet Res.*, vol. 23, no. 5, May 2021, Art. no. e15708.

[12] H. Alhuzali, T. Zhang, and S. Ananiadou, "Predicting sign of depression via using frozen pre-trained models and random forest classifier," in *Proc. CLEF (Work. Notes)*, 2021, pp. 888–896.

[13] S. Graham, C. Depp, E. E. Lee, C. Nebeker, X. Tu, H.-C. Kim, and D. V. Jeste, "Artificial intelligence for mental health and mental illnesses: An overview," *Current Psychiatry Rep.*, vol. 21, no. 11, p. 116, Nov. 2019.

[14] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, and W. Zhu, "Depression detection via harvesting social media: A multimodal dictionary learning solution," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Melbourne, VIC, Australia. CA, USA: International Joint Conferences on Artificial Intelligence Organization, Aug. 2017, pp. 3838–3844.

[15] E. Turcan and K. Mckeown, "Dreaddit: A Reddit dataset for stress analysis in social media," in *Proc. 10th Int. Workshop Health Text Mining Inf. Anal. (LOUHI)*, 2019, pp. 97–107.

[16] A. Rastogi, Q. Liu, and E. Cambria, "Stress detection from social media articles: New dataset benchmark and analytical study," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2022, pp. 1–8.

[17] M. Garg, M. Gaur, R. Goswami, and S. Sohn, "LoST: A mental health dataset of low self-esteem in Reddit posts," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2023, pp. 3854–3859.

[18] S. Ghosh, A. Ekbal, and P. Bhattacharyya, "Cease, a corpus of emotion annotated suicide notes in english," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 1618–1626.

[19] T. Atapattu, M. Herath, C. Elvitigala, P. de Zoysa, K. Gunawardana, M. Thilakaratne, K. de Zoysa, and K. Falkner, "Emoment: An emotion annotated mental health corpus from two South Asian countries," in *Proc. 29th Int. Conf. Comput. Linguistics*, 2022, pp. 6991–7001.

[20] J. Chaturvedi, S. Velupillai, R. Stewart, and A. Roberts, "Identifying mentions of pain in mental health records text: A natural language processing approach," 2023, *arXiv:2304.01240*.

[21] M. Garg, C. Saxena, S. Saha, V. Krishnan, R. Joshi, and V. Mago, "Cams: An annotated corpus for causal analysis of mental health issues in social media posts," in *Proc. 13th Lang. Resour. Eval. Conf.*, 2022, pp. 6387–6396.

[22] M. K. Kabir, M. Islam, A. N. B. Kabir, A. Haque, and M. K. Rhaman, "Detection of depression severity using Bengali social media posts on mental health: Study using natural language processing techniques," *JMIR Formative Res.*, vol. 6, no. 9, Sep. 2022, Art. no. e36118.

[23] H. Sun, Z. Lin, C. Zheng, S. Liu, and M. Huang, "PsyQA: A Chinese dataset for generating long counseling text for mental health support," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021.

[24] A. Alasmari, L. Kudryashov, S. Yadav, H. Lee, and D. Demner-Fushman, "CHQ- SocioEmo: Identifying social and emotional support needs in consumer-health questions," *Sci. Data*, vol. 10, no. 1, p. 329, May 2023.

[25] S. H. Aldhafer and M. Yakhlef, "Depression detection in Arabic tweets using deep learning," in *Proc. 6th Int. Conf. Inf. Technol., Inf. Syst. Electr. Eng. (ICITISEE)*, Dec. 2022, pp. 1–6.

[26] N. Al-Musallam and M. Al-Abdullatif, "Depression detection through identifying depressive Arabic tweets from Saudi Arabia: Machine learning approach," in *Proc. 5th Nat. Conf. Saudi Comput. Colleges (NCCC)*, Dec. 2022, pp. 11–18.

[27] A. Al-Laith and M. Alenezi, "Monitoring people's emotions and symptoms from Arabic tweets during the COVID-19 pandemic," *Information*, vol. 12, no. 2, p. 86, Feb. 2021.

[28] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural language processing applied to mental illness detection: A narrative review," *NPJ Digit. Med.*, vol. 5, no. 1, pp. 1–13, Apr. 2022.

[29] H. Guo, X. Na, and J. Li, "Qcorp: An annotated classification corpus of Chinese health questions," *BMC Med. Inform. Decis. Making*, vol. 18, no. S1, pp. 39–47, Mar. 2018.

[30] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, Nov. 1971.

[31] K. Denecke and Y. Deng, "Sentiment analysis in medical settings: New opportunities and challenges," *Artif. Intell. Med.*, vol. 64, no. 1, pp. 17–27, May 2015.

[32] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, "The interplay of variant, size, and task type in Arabic pre-trained language models," in *Proc. 6th Arabic Natural Lang. Process. Workshop*, 2021, pp. 92–104.

[33] C. Li, J. Fu, J. Lai, L. Sun, C. Zhou, W. Li, B. Jian, S. Deng, Y. Zhang, Z. Guo, Y. Liu, Y. Zhou, S. Xie, M. Hou, R. Wang, Q. Chen, and Y. Wu, "Construction of an emotional lexicon of patients with breast cancer: Development and sentiment analysis," *J. Med. Internet Res.*, vol. 25, Sep. 2023, Art. no. e44897.

[34] H. Khanpour and C. Caragea, "Fine-grained emotion detection in health-related online posts," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 1160–1166.

HASSAN ALHUZALI received the M.S. degree in information science from Indiana University-Bloomington, USA, in 2016, and the Ph.D. degree in computer science from The University of Manchester, U.K., in 2022. Prior to that, he was an Associate Researcher with The University of Manchester. Following the completion of the master's degree, he embarked on a period as a Visiting Student with the Positive Psychology Center, UPENN, USA, and UBC, CA, USA. Currently, he is an Assistant Professor with Umm Al-Qura University, SA, USA. His ongoing research interests include natural language processing, affective computing, and mental health.

ASHWAG ALASMARI received the Ph.D. degree from the University of Maryland, Baltimore County, USA, in 2021. She is currently an Assistant Professor with the Department of Computer Science, King Khalid University, Abha, Saudi Arabia. Her research interests include medical question answering, health consumer informatics, human information interaction, natural language processing, and deep learning. Prior to her current position, she held research positions with the Johns Hopkins University School of Medicine and U.S. National Library of Medicine (NLM), National Institutes of Health (NIH).

HAMAD ALSALEH is currently a Faculty Member with the College of Computer and Information Science, Department of Information Systems, King Saud University. His expertise encompasses two key areas: natural language processing (NLP) and human–computer interaction (HCI). With a focus on NLP, his explores language understanding and generation, while their work in HCI delves into improving user interaction with technology. Particularly, he passionate about addressing socio-cultural problems.