Computers, Materials & Continua
DOI:10.32604/cmc.2022.020358
Review

Tech Science Press

Mining the Chatbot Brain to Improve COVID-19 Bot Response Accuracy

Mukhtar Ghaleb[1,*], Yahya Almurtadha[2], Fahad Algarni[3], Monir Abdullah[3], Emad Felemban[4], Ali M. Alsharafi[3], Mohamed Othman[5] and Khaled Ghilan[6]

[1]Department of Information Systems, College of Sciences and Arts, University of Bisha, Al Namas, 67392, Saudi Arabia
[2]Department of Computer Science, College of Computing and Information Technology, University of Tabuk, Tabuk, 71491, Saudi Arabia
[3]Department of Computer Science, College of Computing and Information Technology, University of Bisha, Bisha, 61922, Saudi Arabia
[4]Computer Engineering Department, Umm Al-Qura University, Makkah, 21955, Saudi Arabia
[5]Department of Communication Technology and Network, Universiti Putra Malaysia, Selangor, 43400, Malaysia
[6]Faculty of Public Health and Tropical Medicine, Jazan University, Jazan, 45142, Saudi Arabia
[*]Corresponding Author: Mukhtar Ghaleb. Email: mghaleb@ub.edu.sa
Received: 20 May 2021; Accepted: 21 June 2021

Abstract: People often communicate with auto-answering tools such as conversational agents due to their 24/7 availability and unbiased responses. However, chatbots are normally designed for specific purposes and areas of experience and cannot answer questions outside their scope. Chatbots employ Natural Language Understanding (NLU) to infer their responses. There is a need for a chatbot that can learn from inquiries and expand its area of experience with time. This chatbot must be able to build profiles representing intended topics in a similar way to the human brain for fast retrieval. This study proposes a methodology to enhance a chatbot's brain functionality by clustering available knowledge bases on sets of related themes and building representative profiles. We used a COVID-19 information dataset to evaluate the proposed methodology. The pandemic has been accompanied by an "infodemic" of fake news. The chatbot was evaluated by a medical doctor and a public trial of 308 real users. Evaluations were obtained and statistically analyzed to measure effectiveness, efficiency, and satisfaction as described by the ISO9214 standard. The proposed COVID-19 chatbot system relieves doctors from answering questions. Chatbots provide an example of the use of technology to handle an infodemic.

Keywords: Machine learning; text classification; e-health chatbot; COVID-19 awareness; natural language understanding

1 Introduction

Artificial Intelligence (AI) enables machines to act independently and intelligently without prior programming. AI learns from continuous interaction with the environment and users. It is of great interest to develop smart conversation agents (chatbots) that interact intelligently with users. Some chatbots interact with appliances and other devices.

The characteristics of smart chatbot systems include the following:

- Its ability to interact simultaneously with multiple users reduces the need for service employees;
- It can work continuously, 24/7;
- Intelligent conversation systems are psychologically unaffected by customers;
- It significantly reduces expenses.

Conversation agents are either dialogue systems or chatbots. Dialogue systems perform specific functions such as making flight reservations [1]. Chatbots mimic human conversation [2]. While English language dialogue systems have been proven stable and accurate, 97% of current dialogue systems are unsuitable for use with Arabic [3].

We propose a methodology to mimic the human brain by grouping related topics for ease of classification, fast retrieval, and increased accuracy of chatbots. Specifically, we propose a COVID-19 chatbot.

Coronaviruses (COVID-19) infect animals and humans and may severely affect health. The coronavirus detected in December 2019 caused a worldwide pandemic [4–6].

Many countries in the Middle East have begun awareness-raising campaigns focusing on prevention rather than treatment, including tips for dealing with COVID-19 and preventing its spread, fighting rumors around it, emphasizing hand-washing, remaining at home, avoiding crowds, practicing social-distancing, and identifying symptoms [7]. The Saudi Ministry of Health has broadcast more than three billion educational text messages in 24 languages [8]. Hassounah et al. [9] highlighted how the Kingdom of Saudi Arabia (KSA) used digital technology in the early stages of the pandemic, and praised the use of chatbots in both the United States and Singapore.

Governments around the world desire to stop the spread of COVID-19. Increasing awareness of pandemic effects is a high priority. This study investigates the development of a chatbot to respond to coronavirus inquiries and share information and advice to help reduce the spread of the virus. This can efficiently increase awareness, for the following reasons:

- Majority of people like using recent technologies;
- Chatbots reduce anxiety and stress by combating the infodemic of fake news [10,11], as a single, credible source of information from organizations such as the World Health Organization (WHO) and Kingdom of Saudi arabia (KSA) Ministry of Health;
- Chatbots are available around the clock. Their information is updated quickly and authoritatively. They can interact with thousands of people simultaneously at a low cost;
- Chatbots reduce demands on healthcare practitioners;
- The proposed system can significantly reduce the cost of healthcare awareness services;
- A chatbot merely requires an internet connection, which makes it convenient, efficient, and fast.

The remainder of this paper is organized as follows. Section 2 presents a literature review covering chatbots, particularly in the medical area. Section 3 describes the methodology and implementation of the proposed chatbot. Section 4 presents an evaluation. Section 5 provides concluding remarks and suggestions for future work.

2 Related Work

People communicate with each other primarily through conversation. Intelligent conversation agents communicate in this way. Moreover, the recent intelligent conversation agents have

convinced the users engaging with the chatbot that the conversational chatbot agent has human-like attributes [12]. Such systems have developed in various areas. For example, Boné et al. [13] developed a Portuguese-speaking chatbot for use in disasters.

Researchers have experimented with chatbots in areas such as education, health, and business. Their potential in education has been analyzed [14,15]. Labeeb [16] introduced an intelligent conversational agent to enhance course teaching and allied learning outcomes. Chien et al. [17] proved that students could collaborate with chatbots for better design solutions. Fryer et al. [18] investigated why chatbots are not yet a powerful tool for language learning. It was argued that the chatbot as a learning tool to improve teaching is still in its infancy [19].

Black et al. [20] systematically reviewed the impact of e-health on health care quality and safety. To develop a conceptual model helps health professionals to adopt it in their areas [21]. Several researchers investigated the use of emerging technologies to improve the health sector. Uohara et al. [22] summarized how chatbot technologies provide the means for triage and to supply care at scale. Technology was found sufficient to persuade patients to modify their behaviors. Van Gemert-Pijnen et al. [23] explained technology can be used as a persuasive approach in the health field. During a pandemic, persuasive technology is vital to convince the public to follow precautions.

Researchers have discussed AI techniques to fight infodemics, both directly and indirectly. Twitter is a significant foundation for infodemiology research [24]. Bahja et al. [25] discussed the importance for policymakers to use social media to identify concerns. Alomari et al. [26] proposed a tool using unsupervised latent Dirichlet allocation (LDA) machine learning to inspect Twitter data in Arabic to identify government pandemic measures and public concerns. Another effort employed a chatbot for news dissemination [27]. Battineni et al. [28] discussed the use of a chatbot to assist patients living in remote areas by encouraging preventive measures, providing updates, and reducing psychological harm triggered by isolation and fear.

Applications and evaluation measures of health-related chatbots were reviewed [29]. Bibault et al. [30] compared chatbots and physicians at delivering information to breast cancer patients. A medical chatbot used AI to diagnose disease before doctor visits [31]. An overview was provided on the use of conversational agents in clinical phycology [32]. A study examined the responses of four commonly used conversational apps to mental health questions [33]. A research model was developed to explain the adoption of conversational agents for disease diagnosis [34]. The willingness to interact with intelligent health chatbots was studied [35].

A study reviewed the role of AI to provide information to prevent COVID-19 infection [36]. Jamshidi et al. [37] extracted reactions to fight the virus through AI. Shen et al. [38] analyzed over 200 articles on robotic systems and concluded that the pandemic would fuel the growth of the robotics industry. A study argued that chatbots could provide needed information updates and lessen the psychological damage caused by fear and isolation [39].

Tanoue et al. [40] adopted a chatbot for the mental health of family, friends, and coworkers of COVID-19 patients in Japan. We found a significant development of chatbots for screening. A chatbot was employed to screen health employees for COVID-19 infection possibility by answering several questions [41]. Martin et al. [42] found that Symptoma, a symptom-to-disease digital health assistant, could identify COVID-19 with 96.32% accuracy. Dennis et al. [43] studied how people react to COVID-19 screening chatbots, and identified a need to convince users that the chatbot can provide the same response as a human.

3 Building the Chatbot Brain

We completed two research objectives. The first objective was to propose an architecture for a chatbot with a human-like brain profile to improve its response accuracy. The second was to develop a chatbot with a credible knowledge base from World Health Organization (WHO) and the Kingdom of Saudi Arabia (KSA) Ministry of Health. Fig. 1 illustrates the phases of the proposed chatbot.

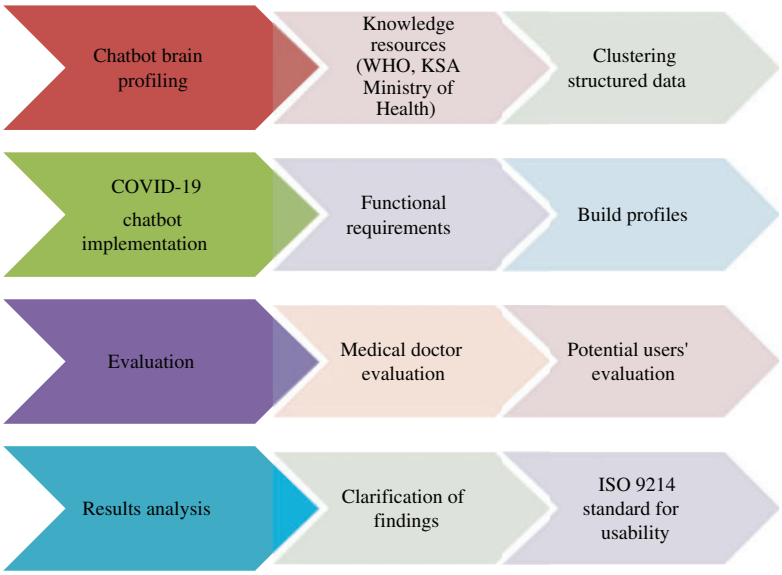

Figure 1: Phases of COVID-19 awareness chatbot development

Problem Definition: We sought to understand the research objectives and requirements from an intelligent software development process and public health concerns, and to formulate a community service problem definition. The chatbot was developed to decrease the burden on healthcare workers.

Planning: We first determined what people thought of chatbot applications and whether they would accept and trust a chatbot during a pandemic. We had to consider the user's motivations and capabilities, with the aim to promote pandemic prevention and behavioral change.

3.1 Goal A: Building Chatbot Brain Profiles

During data gathering, we retrieved COVID-19 health information from the official websites of the WHO and KSA Ministry of Health (Fig. 2).

A medical doctor helped us to order relevant information about each topic of the chatbot repository, as shown in Fig. 3. Data preparation considered activities required to populate the chatbot knowledge base. Tasks included data preprocessing, categorization (prevention, symptoms, and awareness), and selection, and a design suitable for natural language understanding (intents and entities).

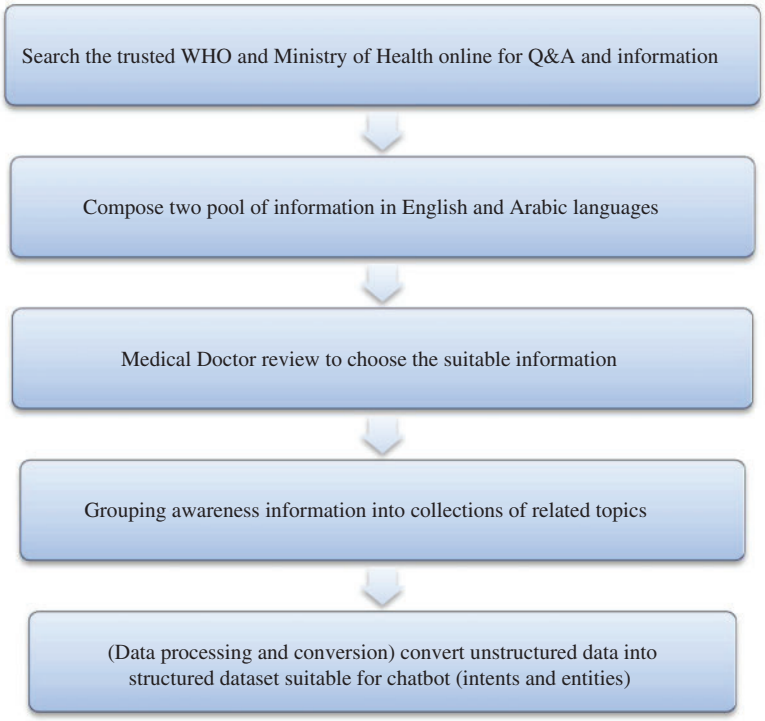

Figure 2: Data collection phase

During the development of the chatbot, we found that to retrieve the appropriate answer to a question could be quite difficult, as questions are short sentences. After tokenization and removal of stop-words, only a few words were left to be manipulated and processed so as to understand the context and find the answer. Although we tried different similarity methods to find the best match to the inquiry, the accuracy of the answer was somewhat questionable, and there was large classification error. We used a methodology and structure to guarantee better accuracy. Fig. 4 illustrates the methodology, which can be explained as follows.

A knowledge base prepared at the previous step was preprocessed and converted to a structured format suitable for the chatbot inference engine. The development team transformed this to entities (e.g., places, objects) and intents (what the human should obtain as a response). A doctor clustered the accepted dataset into groups of questions with similar intents, keywords, and phrases (terms).

The output is a number of clusters, each containing a group of questions and related answers with unique IDs. Clustering is a good way to reduce the calculation of the similarity of questions to specific clusters (profiles), instead of to all the questions in the dataset. Each cluster is associated with a list of terms from the cluster profile representing that cluster and distinguishing it from other clusters. These terms are generated by tokenizing the questions in the cluster. The stop-words were removed to keep only the meaningful words and reduce the possibility of counting words with the same meaning (e.g., 'liked' and 'liking'). Then the remaining tokens were stemmed using a porter algorithm to find the root word (e.g., 'like'). The frequency of stemmed words was counted. Finally, these terms were weight with TF-IDF [44].

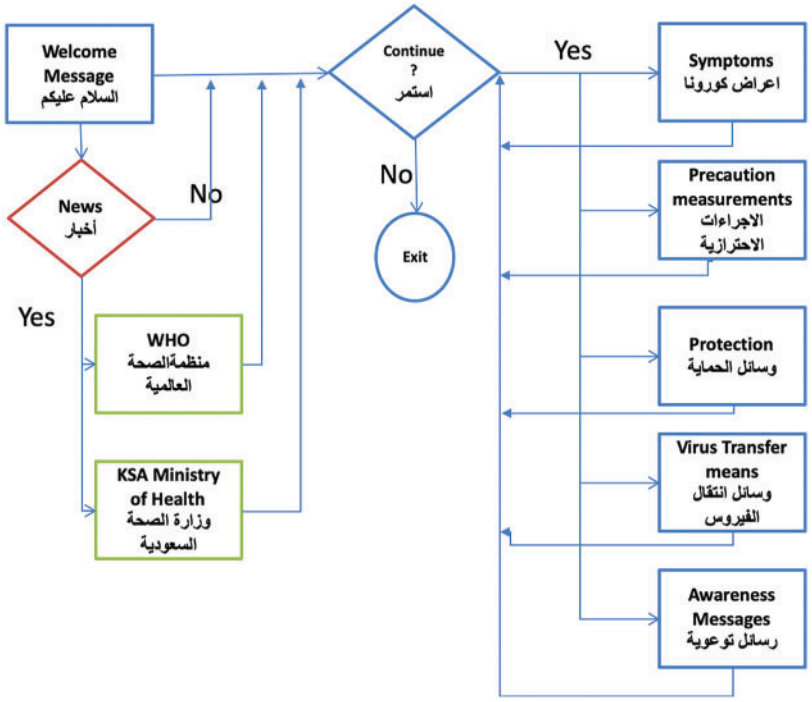

Figure 3: Flowchart of COVID-19 chatbot knowledge base

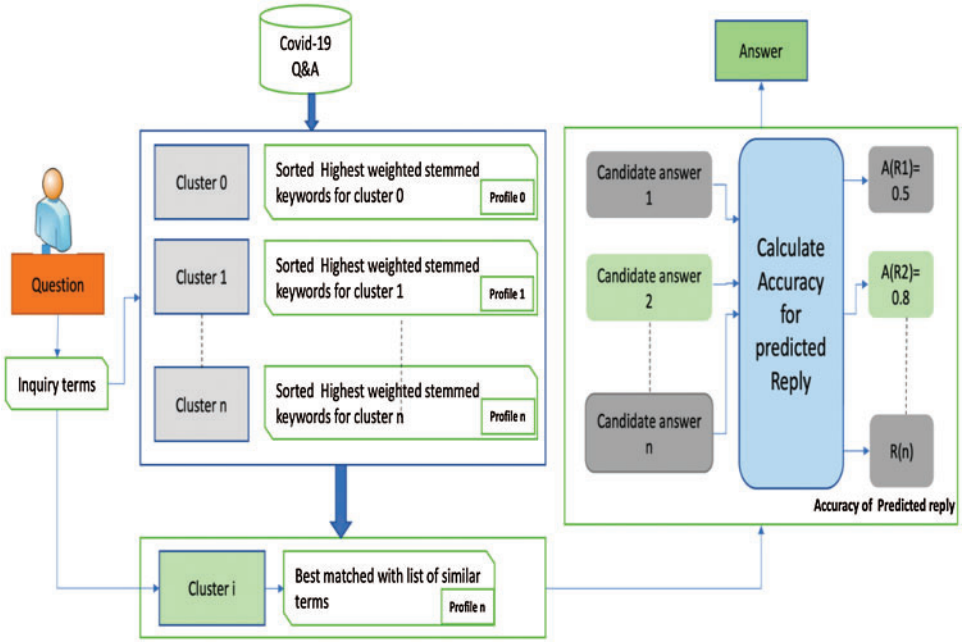

Figure 4: Building brain profiles for better similarity prediction

n tokens to form question set Q = {T_1, T_2,..., T_n}. TF-IDF calculates the frequency of occurrence of i in question set Q, and $W_{i,j}$ is the calculated weight for term i in question j.

$$tf(t,q) = \frac{f_{t,q}}{\sum_{t^- \in q} f_{t,q}} \tag{1}$$

$$idf(t,Q) = \log\left(\frac{N}{count(t \in T,\ q \in Q)}\right) \tag{2}$$

$$tf - idf(t,q,Q) = tf(t,q).\ idf(t,Q) \tag{3}$$

The weight for each term is calculated as

$$WW_{i,j} = \left(0.5 + 0.5\frac{f_{t,q}}{max_t f_{t,q}}\right) \cdot log\frac{N}{n_t} \tag{4}$$

where N is the total number of questions in the dataset to be clustered to groups of profiles. The stemmed and weighted keywords are sorted in ascending order based on the calculated weight using Eq. (4) to form the list of keywords that best represents the cluster. When a user asks a question, the similarity engine calculates the similarity of the question to the profiles and finds the most related profile. The similarity engine calculates the similarity to the questions in the selected profile to match it with the most similar question using similarity Eq. (5) and retrieves the answer.

Applying Natural Language Processing: NLP techniques are applied to build a chatbot brain to comprehend requests and respond accordingly. NLP can be categorized into natural language understanding (NLU) and natural language generation (NLG) [45]. Fig. 5 illustrates the main steps of getting the human question, digitizing it, understanding it, and finding the most suitable answer.

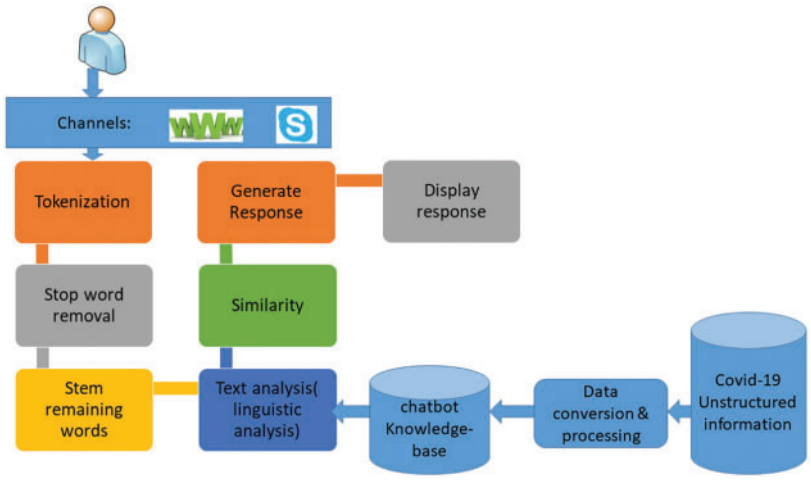

Figure 5: Receiving and responding to inquiries

Structuring input data: An NLP chatbot follows several steps to transform a human inquiry into structured data it can understand and choose the correct response. NLP then breaks down the investigation into tokens that can be processed and analyzed to extract meanings and relations.

For better similarity matching later, Arabic stop-words (such as في، من، إلى، الان، الذي، انا، انت،...) are removed. Remaining words are stemmed to their roots to assure that no two terms of the same purpose are extracted. Linguistic analysis is then applied to derive meanings. Finally, the chatbot pursues entity classes like COVID-19 symptoms, preventions, and awareness, which helps the Named Entity Recognition function to recognize the entities in the question to match it with the related answer.

A similarity measure, as explained by Algorithm 1, is applied to weigh and choose the highest practical intent for the purpose of finding the suitable answer for this inquiry. Similarity can be measured by Jaccard similarity or Levenshtein distance. Jaccard similarity is used to calculate the similarity of an inquiry (*Inq*) and a generated list of related intents and entities (*Res*). It is computed as

$$J(Inq, Res) = \frac{|Inq \cap Res|}{|Inq \cup Res|} = \frac{|Inq \cap Res|}{|Inq| + |Res| - |Inq \cap Res|} \tag{5}$$

where 0 ≤ J (*Inq, Res*) ≤ 1. The numerator in Eq. (5) is the intersection of elements in both statements, and the denominator is the total number of items across them. We assume that similarity scores greater than 65% are equivalent. Since several intents might be semi-related as an answer, the response with the highest calculated score is chosen.

The Levenshtein distance measures the difference between statement texts. The Levenshtein distance between two strings *Inq* and *Res* (of length |*Inq*| and |*Res*|, respectively) is given by

$$lev(Inq, Res) = \begin{cases} |Inq| & if\,|Res| = 0, \\ |Res| & if\,|Inq| = 0, \\ lev(tail(Inq), tail(Res) & if\ Inq[0] = Res[0] \\ 1 + min \begin{cases} lev(tail(Inq), Res) \\ lev(Inq, tail(Res)) \\ lev(tail(inq),\ tail(Res)) \end{cases} & otherwise \end{cases} \tag{6}$$

where the tail of some string *S* is a string of all but the first character of *S*, and *S[n]* is the n^{th} character of string *S*, starting with character 0.

Algorithm 1: Build-and-Select-best-Response.

Input: user-inquiry (*q*)
Output: chatbot appropriate response (*r*)

1. Convert the user's inquiry into structured data suitable for response generation process
2. begin
3. While *q* is not empty do
4. Convert inquiry sentence to lower-case
5. Convert inquiry *q* into set of words
6. extract_words(*q*→*w*[ ])
7. Remove-Stopwords(*w*[ ] - Stop-List[ ])
8. //Divide the remaining-extracted words into tokens linguistically meaningful
9. Tokenize(*w*[ ] → *t*[ ])
10. Token-stemming-using-snowballStemmer (*t*[ ] → *s*[ ])
11. stem_words *s* = [ ]

(Continued)

12. for t in words:
13. ss= SnowballStemmer_stemmer.stem(t)
14. stem_words.append(ss)
15. Named-Entity-Recognition E to recognize facts such as symptoms, prevention-measures, places
16. Recognize intents-List N
17. Calculate-Intents-priority to choose N with highest score to match the inquiry
18. Convert selected Intent into set of words with frequencies
19. Function Build-Responses-list (R)
20. Calculate-similarity[response]
21. $C \leftarrow$ Extract candidate answers related to R
22. for ($w \in C$) do
23. Weight answers(intents)
24. $C \leftarrow$ Sort intent-response
25. $S \leftarrow$Choose the intent-response with highest confidence
26. Return r

3.2 Goal B: An Awareness COVID-19 Chatbot

Our second goal is to help KSA authorities to increase awareness of the COVID-19 pandemic through the chatbot. The structured knowledge base for the chatbot was constructed from information from the WHO and KSA Ministry of Health. Fig. 6 shows the chatbot's architecture. The user asks a question, which the system forwards to the chatbot interface [46] as the system's back end. We used Chatterbot (https://chatterbot.readthedocs.io), a Python library, to develop the chatbot.

The NLP engine identifies intents and entities from an inquiry, and a list of candidate responses is generated. The response with the highest weight is sent back to the user as the response. Chat history is saved in a MongoDB database [47]. The front end was developed with Python and Flask, a Python framework used to develop Web applications. We used a PyCharm integrated development environment (IDE) for Python programming, and RapidAPI to search an API with updated COVID-19 information. A webhook connected the chatbot interface to the Python/Flask [48] framework. MongoDB Atlas was used to save inquiries and answers. Fig. 7 shows sample inquiries and responses in Arabic, with English translations.

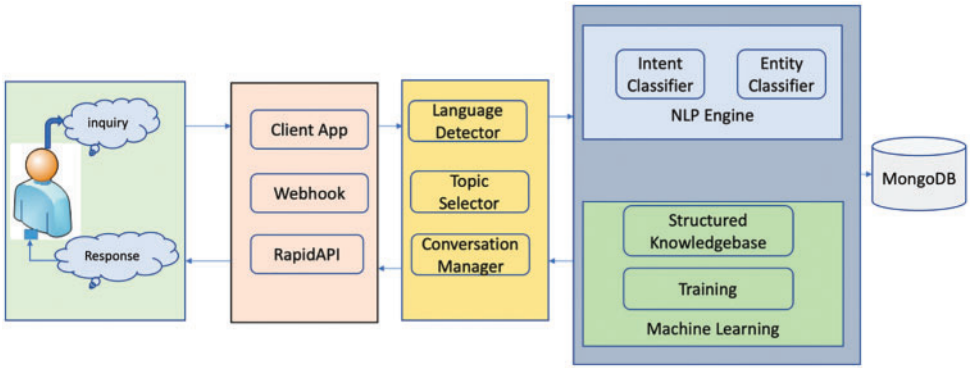

Figure 6: Chatbot architecture

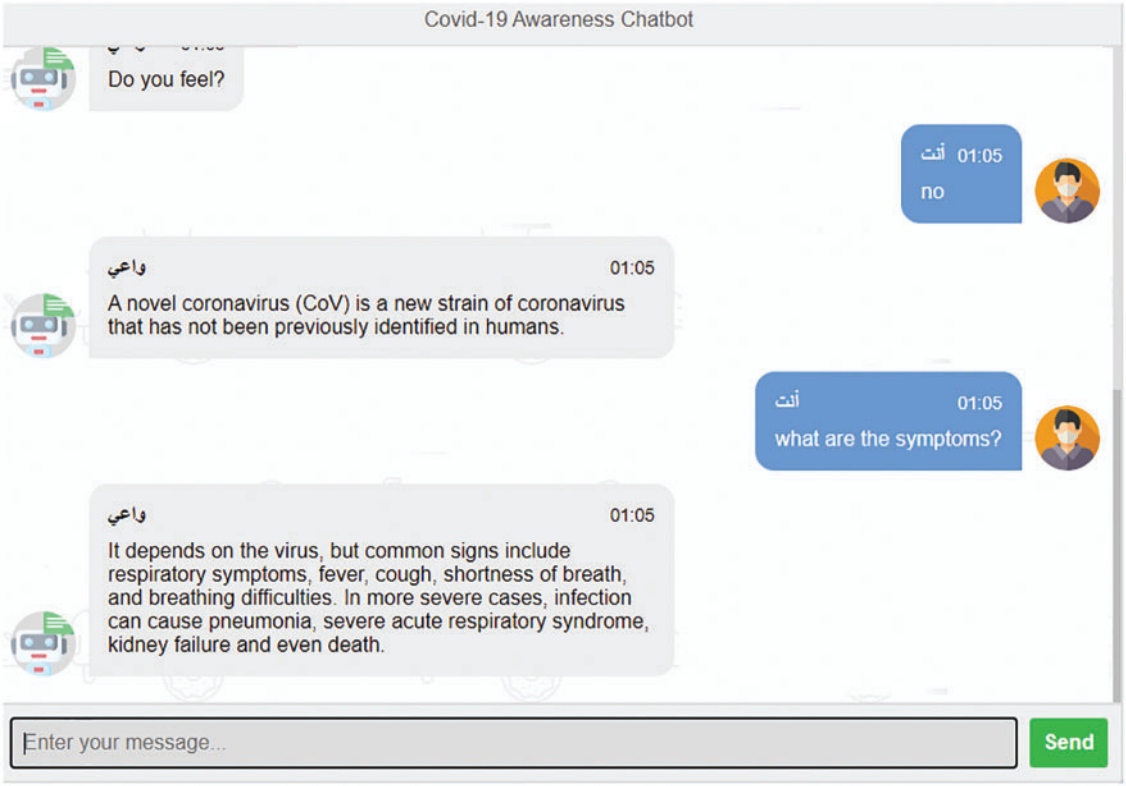

Figure 7: Sample of user inquiries and responses

4 Experimental Evaluation

We evaluated the ability of the chatbot to understand inquiries, and to accurately respond to them in a timely manner. We performed two experiments to evaluate the effectiveness of (1) clustering questions and answers into groups of related topics and contexts; and (2) the proposed chatbot.

4.1 First Experiment

We validated the proposed architecture for improving the similarity calculation and matching of questions to answers. Different classification algorithms were used to evaluate the performance of the proposed architecture using different volumes of questions in each round, such as 100, 200, 300, 400, 500, and 600 questions.

4.1.1 COVID-QA Dataset

To test the accuracy of the classification of answers to questions, we used the COVID-QA dataset on Kaggle (https://www.kaggle.com/xhlulu/COVIDqa) with over 800 paired questions and answers retrieved from FAQs of the Centers for Disease Control and Prevention (CDC) and WHO, which are available in eight languages. All pairs were cleaned with regex, labeled with metadata, converted to tables, and stored in CSV files.

4.1.2 Evaluation Results

The dataset was split into training (70%) and testing (30%) sets. We measured performance by Accuracy, Precision, Recall, and F1:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{10}$$

where TP = number of correctly predicted positive, TF = number of correctly predicted negative, FP = number of falsely predicted positive, and FN = number of falsely predicted negative.

The classification was trained on set QA = $(q_1, a_1),(q_2, a_2),\ldots,(q_n, a_n)$ of question (q_i)-answer (a_i) pairs. We used k-nearest neighbors (KNN) and Naïve Bayes classification algorithms on the RapidMiner platform [49]. When applying k-means, terms were distributed among the five generated clusters, as shown in Fig. 8. The centroids of terms in the clusters are depicted in Fig. 9, while Fig. 10 shows a heat map [50] of individual values in clusters. Terms are clearly allocated to exactly one cluster, which enhances the accuracy of matching questions and answers.

```
Cluster 0: 251 items
Cluster 1: 670 items
Cluster 2: 232 items
Cluster 3: 7329 items
Cluster 4: 147 items
Total number of items: 8629
```

Figure 8: Term distribution in each cluster

The KNN algorithm [51] finds the nearest neighbor of a new instance of the dataset by calculating the distance to the nearest neighbor in the n-dimensional space. Naïve Bayes [52] is a computationally inexpensive classifier normally used for text categorization. Tab. 1 lists the top 10 terms sorted by their calculated weights of importance that represent and distinguish each profile. For instance, profile 0 is made up of the coronavirus definition, updates and causes, while profile 2 is made up of coronavirus tests and whether the results are positive or negative. Tab. 2 displays the accuracy evaluation metrics depending on the number of questions, from which we can see that the accuracy increased with the number of questions. This is due to the adding of the correct terms representing each profile which increases the opportunity for the new question to be broken into words that match them with the correct group of answer in the profile.

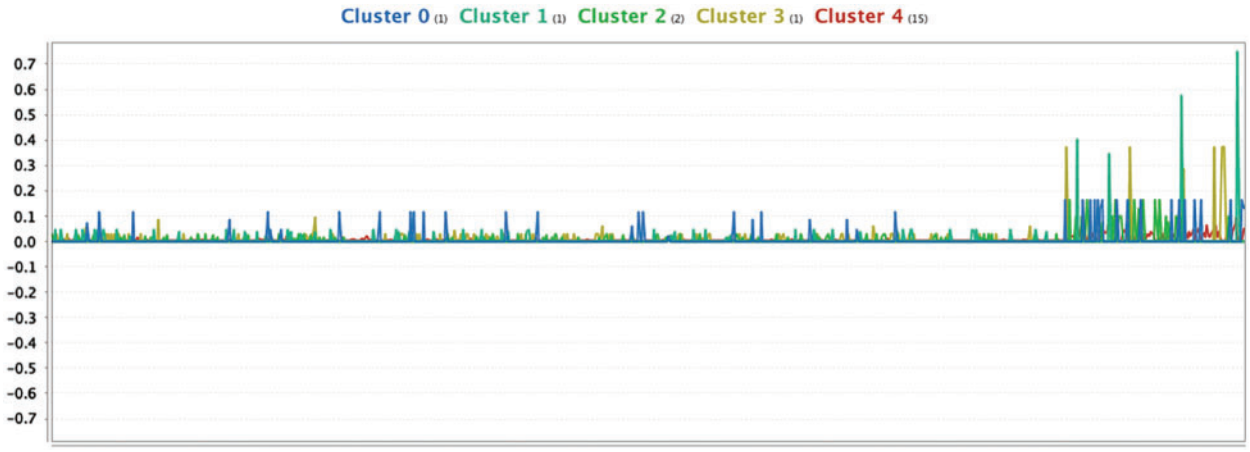

Figure 9: K-means centroid chart

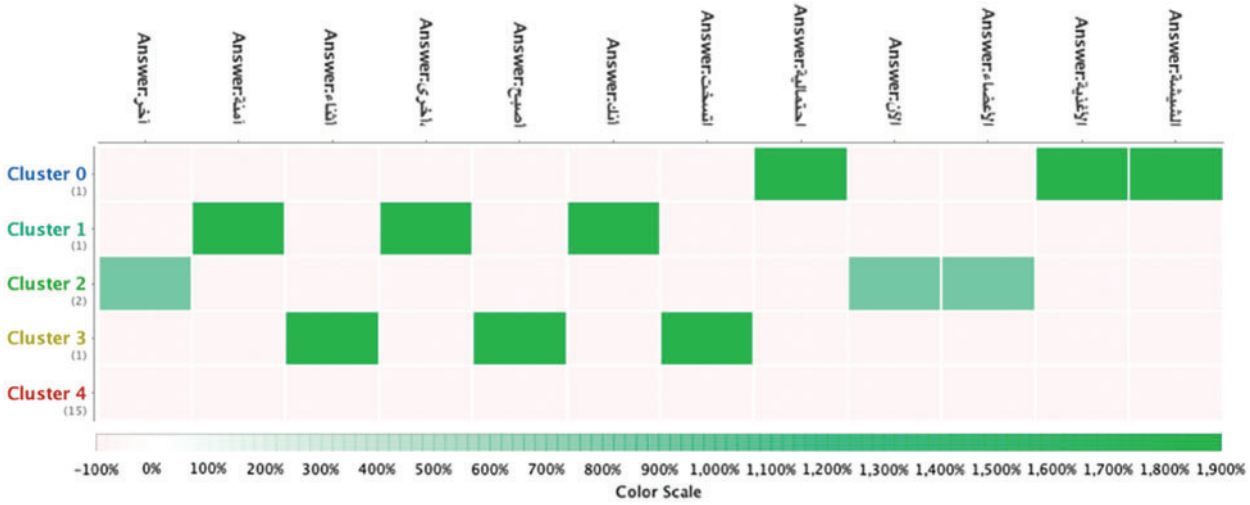

Figure 10: K-means heat map for some terms

Table 1: Accuracy evaluation metrics

Profile 0		Profile 1		Profile 2		Profile 3		Profile 4	
Term	Weight	Term	Weight	Term	Weight	Term	Weight	Term	Weight
Viruses	0.0719001	Coronavirus	0.0444876	Virus	0.1365204	WWW	0.0042182	Size	0.0189964
Vaccine	0.0433971	Question	0.0327215	Test	0.0474834	Article	0.0037754	Mutations	0.0185899
March	0.0431289	People	0.0317176	Positive	0.0246366	Italy	0.0035626	Claim	0.0181024
Bleach	0.0351639	Days	0.0288056	Covid	0.0211047	Viral	0.0034026	NSP	0.0174763
Update	0.0325141	China	0.0250509	Negative	0.0195566	Immune	0.0031156	Human	0.0173833
Surfaces	0.0304114	Answer	0.0243003	Sensitivity	0.0195245	Org	0.0030745	Sequence	0.0165536
Water	0.0269988	COVID	0.0231053	Specificity	0.0179726	Com	0.0030745	Protein	0.0156211
Bats	0.0239080	Cases	0.0200131	Viruses	0.0163624	Pandemic	0.0030736	Diameter	0.0155753
Cold	0.0230316	References	0.0186644	Infected	0.0150874	Cells	0.0029813	Lopinavir	0.0152514
Coronavirus	0.0221041	Infected	0.0178177	Tests	0.0132564	Pathogens	0.0029650	Origin	0.0145441

Table 2: Accuracy evaluation metrics

	KNN					Naïve Bayes				
Number of questions	Accuracy	Precision	Recall	Classification error	F1	Accuracy	Precision	Recall	Classification error	F1
100	79.4	78.2	79.4	20.5	78.79	80.4	78.5	80.04	19.5	79.26
200	82.05	80.55	82.02	17.9	81.278	83.56	80.94	82.82	16.44	81.869
300	82.05	80.55	82.02	17.9	81.27	83.56	80.94	82.82	16.44	81.86
400	83.12	86.71	83.12	11.88	84.877	82.53	80.47	81.99	17.47	81.222
500	84.82	83.2	83	14	83.099	84.82	83.04	84.65	15.18	83.837
600	87.1	85.7	87.1	12.8	86.39	88.12	86.7	88.12	11.8	87.40

4.2 Second Experiment

The developed chatbot was evaluated thoroughly. All steps in its construction were reviewed, particularly as related to the knowledge base. Most chatbots present options which lead to further layers of options, depending on the user's response. However, our chatbot was designed for open conversations without menus, options, or directions from the system. This makes accuracy more difficult for the following reasons:

- There are different human expressions for the same inquiry;
- There are various dialects;
- Not all user inputs can be predefined; hence, the chatbot must respond to unanticipated questions.

In addition to testing the chatbot with potential users, we asked medical professionals, academics, and students to use the chatbot and answer several questions regarding their level of experience, awareness, satisfaction, and recommendations. User feedback was reviewed by a medical doctor and statistics expert to evaluate the chatbot's efficiency and efficacy.

We employed four evaluation methods, based on (1) in-house; (2) experts; (3) real users; and (4) ISO 9214 standard of usability (effectiveness, efficiency, and satisfaction) [53].

4.2.1 In-House Evaluation

Training and testing the chatbot during the development is done by interacting with the chatbot and then retrieving the saved history of all the questions asked and inquiries made and what the system has responded to. We determined the percentage of correct answers. Knowing the questions with wrong answers helped us reclassify some questions, anticipate new questioning methods, and redefine intents and entities. We also learned of inquiries that we had not considered.

4.2.2 Expert Evaluation

Expert evaluation can determine whether chatbot responses are suitable or natural [53,54]. We fetched the conversation history of users and chatbots during testing. A medical doctor determined whether the chatbot's answers to questions were correct and appropriate. Based on this, we calculated the precision, as shown in Tab. 3.

The doctor explained some reasons behind the wrong answers. Some users asked strange and irrelevant questions such as "هل انا شجره؟" ("Am I a tree?") or "ماذا يحدث؟" ("What is happening?").

Table 3: Precision and accuracy of expert doctor evaluation

No. of sessions investigated	796 sessions
No. of sessions with incorrect answers to some questions	81 sessions
Precision	81.5%
Accuracy	89.82%

Some questions, like "اشرح لنا تجارب الدول في مكافحه الوباء" ("Explain other countries' experience in fighting the pandemic"), raised the need for more sophisticated responses. Some users asked questions to test the chatbot's ability to reply. A default answer was prepared for such questions: "فضلا اعد صياغه السؤال" ("Kindly ask relevant questions").

4.2.3 Real Users' Evaluation

We aimed to assess the following: (1) the effectiveness of the chatbot for real users; (2) the role of the chatbot to increase users' awareness; and (3) users' level of satisfaction. To do this, we tested the following research hypotheses (RHs) (Fig. 11):

H1: the chatbot's effective and accurate responses to inquiries leads to user satisfaction. This RH investigated the effectiveness of the ISO 9214 standard of usability for chatbot evaluation.
H2: Using the chatbot positively and significantly increases users' awareness. This RH investigated the efficiency of the ISO 9214 standard of usability for chatbot evaluation.
H3: Users' satisfaction of using the chatbot is significantly mediated by their awareness. This RH investigated the satisfaction metric of the ISO 9214 standard of usability for chatbot evaluation.

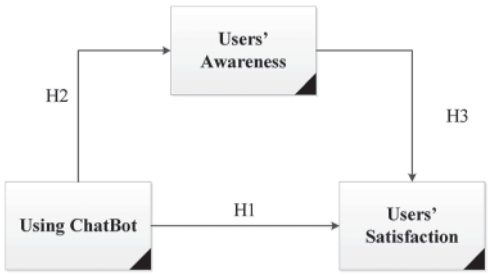

Figure 11: Empirical research model

We solicited users through WhatsApp. A Google Forms questionnaire was distributed to determine their awareness and satisfaction. The three-part questionnaire measured: (1) knowledge of using a chatbot system; (2) awareness created by using the chatbot system; and (3) user satisfaction with the chatbot's functionality, effectiveness, response precision, and speed of response. The targeted population was 35 million citizens residing in Saudi Arabia. The sample calculated using Morgan's table [55] for sampling size was calculated as 385. After one month, 308 responses were received, for a response rate of 80%.

Statistical Analysis Some major variables in the statistical analysis are shown in Tab. 4. Females accounted for 51.6% of respondents, students for 51.9%; 54.5% were single, 48.4% held a graduate degree, and 56.8% were 15–30 years old. We tested the variance of using chatbot program between Male and Female using Independent Sample t-test, as the data showed a normal distribution (P-value = 0.000 for both Kolmogorov–Smirnov and Shapiro–Wilk tests [56]). The results indicate a significant difference between the groups (t = −6.357, P-value = 0.000), where the mean of female chatbot use was more than that of males. Cronbach's alpha was 0.857, indicating that all constructs exhibited internal reliability. Tab. 5 shows the mean and standard deviation of each construct.

Table 4: Hypothesis variable descriptions

Variable	Description
CHB_USE	Ease of use
Awareness	Awareness as a method to fight the infodemic
Satisfaction	Feeling of acceptance and fulfillment of need for accurate information.

Table 5: Mean and standard deviation for major variables

Variable	Mean	Std. Deviation
CHB_USE	0.85110	4.1989
Awareness	0.62018	4.2558
Satisfaction	0.79174	4.2873

Correlation Results Tab. 6 presents correlations among the three major constructs, proving a significant relationship between the three constructs at the 0.01 level (2-tailed). This paves the way for further investigation of the effects between variables.

Table 6: Correlations between major constructs

Variables	CHB_USE	Awareness	Satisfaction
CHB_USE	1		
Awareness	0.567**	1	
Satisfaction	0.799**	0.649**	1

Hypotheses Results The results of hypothesis tests are shown in Fig. 12 and summarized in Tab. 7. According to the results found, there was a significant effect of chatbot program using and responding to users' inquiries on users' satisfaction at the 0.01 level (B = 0.799, P-value = 0.000). Moreover, the correlation presented in Tab. 8 between the chatbot program using and

users' satisfaction supports the direct relationship with the 79.9% correlation found between both constructs. Therefore, the first hypothesis is supported.

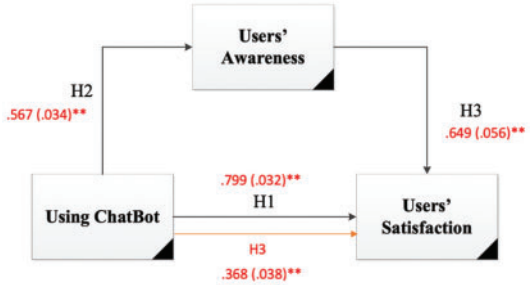

Figure 12: Estimated research model

Table 7: Results of direct and indirect relationship analysis

Hypothesis	Direct-beta coefficient	Indirect-beta coefficient	Mediation type observed
H1: Chatbot-effectiveness → Satisfaction	0.799 (0.032)**	—	—
H2: Chatbot-Using → increasing awareness	0.567 (0.034)**	—	—
H3: Chatbot-Using → Satisfaction (through increasing awareness)	—	0.368 (0.038)**	Full Med

Tests showed that use of the chatbot had a significant effect on user awareness at the 0.01 level (B = 0.567, P-value = 0.000). The correlation of 0.567 between both constructs indicates that the percentage of the relationship between chatbot program using ad users' awareness of 56.7% is supported by the direct relationship found. Hence the second hypothesis is supported.

The third hypothesis supposes a mediation effect of user awareness between chatbot use and user satisfaction. Results of a Sobel test indicate a significant mediation effect of users' awareness on the relationship between using the chatbot program and user satisfaction at the 0.01 level (B = 0.368, p-value = 0.000). Therefore, the third hypothesis is supported.

4.3 ISO 9214 Standard for Usability

As mentioned above, we adopted the ISO 9214 standard to support the chatbot evaluation. This standard is based on effectiveness, efficiency, and satisfaction. Effectiveness concerns the chatbot's ability to fulfill its intended purpose. Efficiency concerns the ability to perform tasks without wasting resources. Satisfaction concerns users' feelings that they get what they need. Of the 308 survey responses, 94% supported the high impact of using technology to promote health awareness, and 83.4% supported the use of the chatbot as a new awareness system that was better than emails and text messages. While 37.5% of respondents had used a chatbot, only 22.5% had

tried a smart system to learn about the coronavirus. Tab. 8 shows the statistical distribution of users' responses. It can be seen that the proposed chatbot effectively answered their inquiries, with 77% highly satisfied with the chatbot. Some 72% of the responses expressed that the chatbot had increased their awareness of COVID-19, 51% were very satisfied, and more than 31% were satisfied using the chatbot. Finally, 78% of users indicated that they would recommend the chatbot to others.

Table 8: ISO 9214 standard for usability evaluation summary

ISO 9214 Standard	Achieved objectives
Effectiveness (77%)	Functionality achieved via: • Credible information from trusted sources • High satisfaction with language and expressions used to answer inquiries • Simple language • Easy access Capability of smooth user interaction
Efficiency and awareness (72%)	Reliability of increasing coronavirus awareness Efficient and timely response
Satisfaction (82%)	Accessibility: satisfaction with ease of dealing with chat Quality of information Recommending that others use the chatbot Interactivity: Satisfaction with use of smart chat Guarantee of user privacy, since no identification or registration is required

5 Conclusions and Future Work

The COVID-19 pandemic has created an urgent need for knowledge. Smart chatbots can serve as a trusted knowledge base for three reasons. They raise awareness and encourage precautionary measures. They enable health professionals to focus on patients. They counteract the viral spread of fake news.

The proposed chatbot uses NLU to comprehend inquiries and infer responses. A profiling methodology for the knowledge base enhances similarity matching. The proposed chatbot was evaluated while it was built, by a medical doctor to test the accuracy of answers, and by 308 real users. Evaluation results and statistical analyses confirmed its effectiveness, efficiency, and user satisfaction.

For future work, we will consider adding features such as a voice assistant, especially for visually impaired users.

Funding Statement: The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia, for funding this research work (Project Number UB-2-1442).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. H. Al-Ajmi and N. Al-Twairesh, "Building an arabic flight booking dialogue system using a hybrid rule-based and data driven approach," *IEEE Access*, vol. 9, pp. 7043–7053, 2021.

[2] R. Klabunde, "Daniel jurafsky/James H. martin: Speech and language processing. an introduction to natural language processing, computational linguistics, and speech recognition," *Zeitschrift fur Sprachwiss*, vol. 21, no. 1, pp. 134–135, 2002.

[3] M. Hijjawi and Y. Elsheikh, "Arabic language challenges in text based conversational agents compared to the English language," *International Journal of Computer Science and Information Technology*, vol. 7, no. 3, pp. 1–13, 2015.

[4] World Health Organization, "Corona virus pandemic," 2020. [Online]. Available: https://www.who.int/home.

[5] UNICEF, "Coronavirus disease (COVID-19) information center," 2020. [Online]. Available: https://www.unicef.org/coronavirus/covid-19.

[6] Médecins Sans Frontières (MSF), "Our response to the coronavirus COVID-19 pandemic," 2020. [Online]. Available: https://www.msf.org/covid-19-depth.

[7] Saudi Press Agency, "Saudi health broadcasts two billion awareness messages about Corona," 2020. [Online]. Available: https://www.saudi24news.com/2020/04/saudi-health-broadcasts-two-billion-awareness-messages-about-corona-2.html.

[8] Ministry of Health, "MOH sends out over 3 billion educational text messages on novel coronavirus," 2020. [Online]. Available: https://www.moh.gov.sa/en/Ministry/MediaCenter/News/Pages/News-2020-04-16-006.aspx.

[9] M. Hassounah, H. Raheel and M. Alhefzi, "Digital response during the COVID-19 pandemic in Saudi Arabia," *Journal of Medical Internet Research*, vol. 22, no. 9, pp. e19338, 2020.

[10] F. A. Rathore and F. Farooq, "Information overload and infodemic in the COVID-19 pandemic," *Journal of Pakistan Medical Association*, vol. 70, no. 5, pp. 162–165, 2020.

[11] G. Eysenbach, "How to fight an infodemic: The four pillars of infodemic management," *Journal of Medical Internet Research*, vol. 22, no. 6, pp. e21820, 2020.

[12] S. Brahnam and A. De Angeli, "Gender affordances of conversational agents," *Interacting with Computers*, vol. 24, no. 3, pp. 139–153, 2012.

[13] J. Boné, J. C. Ferreira, R. Ribeiro and G. Cadete, "Disbot: A Portuguese disaster support dynamic knowledge chatbot," *Applied Sciences*, vol. 10, no. 24, pp. 9082–9101, 2020.

[14] R. Winkler and M. Soellner, "Unleashing the potential of chatbots in education: A state-of-the-art analysis," in *Academy of Management Annual Meeting (AOM)*, Chicago, USA, pp. 1–40, 2018.

[15] S. Roos, "*Chatbots in Education: A Passing Trend or a Valuable Pedagogical Tool?*," M.S. Thesis, Department of Media and Informatics, Uppsala University, Uppsala, Sweden, 2018.

[16] Y. Almurtadha, "LABEEB: Intelligent conversational agent approach to enhance course teaching and allied learning outcomes attainment," *Journal of Applied Computer Science and Mathematics*, vol. 13, no. 1, pp. 9–12, 2019.

[17] Y. H. Chien and C. K. Yao, "Development of an AI userbot for engineering design education using an intent and flow combined framework," *Applied Sciemces*, vol. 10, no. 22, pp. 7970–7984, 2020.

[18] L. K. Fryer, K. Nakao and A. Thompson, "Chatbot learning partners: Connecting learning experiences, interest and competence," *Computers in Human Behavior*, vol. 93, pp. 279–289, 2019.

[19] D. E. Gonda, J. Luo, Y. L. Wong and C. U. Lei, "Evaluation of developing educational chatbots based on the seven principles for good teaching," in *Proc. of 2018 IEEE Int. Conf. on Teaching, Assessment, and Learning for Engineering, TALE 2018*, Wollongong, NSW, Australia, pp. 446–453, 2018.

[20] A. D. Black, J. Car, C. Pagliari, C. Anandan, K. Cresswell *et al.*, "The impact of eHealth on the quality and safety of health care: A systematic overview," *PLOS Medicine*, vol. 8, no. 1, pp. e1000387–e1000403, 2011.

[21] T. Shaw, D. McGregor, M. Brunner, M. Keep, A. Janssen *et al.*, "What is eHealth (6)? development of a conceptual model for ehealth: Qualitative study with key informants," *Journal of Medical Internet Research*, vol. 19, no. 10, pp. e324–e336, 2020.

[22] M. Y. Uohara, J. N. Weinstein and D. C. Rhew, "The essential role of technology in the public health battle against COVID-19," *Population Health Management*, vol. 23, no. 5, pp. 361–367, 2020.

[23] L. J. E. W. C. van Gemert-Pijnen, S. M. Kelders, N. Beerlage-de Jong and H. Oinas-Kukkonen, "Persuasive health technology," in *eHealth Research, Theory and Development: A Multi-Disciplinary Approach*, 1st ed., New York: Routledge, Taylor & Francis, 2018.

[24] J. Xue, J. Chen, R. Hu, C. Chen, C. Zheng *et al.*, "Twitter discussions and emotions about the COVID-19 pandemic: Machine learning approach," *Journal of Medical Internet Research*, vol. 22, no. 11, pp. e20550–e20564, 2020.

[25] M. Bahja, R. Hammad and M. Amin Kuhail, "Capturing public concerns about coronavirus using arabic tweets: An NLP-driven approach," in *Proceedings-2020 IEEE/ACM 13th Int. Conf. on Utility and Cloud Computing*, Leicester, UK, pp. 310–315, 2020.

[26] E. Alomari, I. Katib, A. Albeshri and R. Mehmood, "COVID-19: Detecting government pandemic measures and public concerns from twitter arabic data using distributed machine learning," *International Journal of Environmental. Research and Public Health*, vol. 18, no. 1, pp. 282–316, 2021.

[27] T. A. Maniou and A. Veglis, "Employing a chatbot for news dissemination during crisis: Design, implementation and evaluation," *Future Internet*, vol. 12, no. 7, pp. 109–123, 2020.

[28] G. Battineni, N. Chintalapudi and F. Amenta, "AI chatbot design during an epidemic like the novel coronavirus," *Healthcare*, vol. 8, no. 2, pp. 154–162, 2020.

[29] L. Laranjo, A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen *et al.*, "Conversational agents in healthcare: A systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 9, pp. 1248–1258, 2018.

[30] J. E. Bibault, B. Chaix, A. Guillemassé, S. Cousin, A. Escande *et al.*, "A chatbot versus physicians to provide information for patients with breast cancer: Blind, randomized controlled noninferiority trial," *Journal of Medical Internet Research*, vol. 21, no. 11, pp. e15787–e15793, 2019.

[31] S. Divya, V. Indumathi, S. Ishwarya, M. Priyasankari and S. Kalpana Devi, "A self diagnosis medical chatbot using artificial intelligence," *Journal of Web Development and Web Designing*, vol. 3, no. 10, pp. 1–7, 2018.

[32] S. Provoost, H. M. Lau, J. Ruwaard and H. Riper, "Embodied conversational agents in clinical psychology: A scoping review," *Journal of Medical Internet Research*, vol. 19, no. 5, pp. e151–e169, 2017.

[33] A. S. Miner, A. Milstein, S. Schueller, R. Hegde, C. Mangurian *et al.*, "Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health," *JAMA Internal Medicine*, vol. 176, no. 5, pp. 619–625, 2016.

[34] S. Laumer, C. Maier and G. Fabian, "Chatbot acceptance in healthcare: Explaining user adoption of conversational agents for disease diagnosis," in *27th European Conf. on Information Systems-Information Systems for a Sharing Society, ECIS*, Stockholm & Uppsala, Sweden, pp. 1–18, 2019.

[35] T. Nadarzynski, O. Miles, A. Cowie and D. Ridge, "Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study," *Digital Health*, vol. 5, pp. 1–12, 2019.

[36] R. Vaishya, M. Javaid, I. H. Khan and A. Haleem, "Artificial intelligence (AI) applications for COVID-19 pandemic," *Diabetes Metabolic Syndrome: Clinical. Research and Reviews*, vol. 14, no. 4, pp. 337–339, 2020.

[37] M. Jamshidi, A. Lalbakhsh, J. Talla, Z. Peroutka, F. Hadjilooei *et al.*, "Artificial intelligence and COVID-19: Deep learning approaches for diagnosis and treatment," *IEEE Access*, vol. 8, pp. 109581–109595, 2020.

[38] Y. Shen, D. Guo, F. Long, L. A. Mateos, H. Ding *et al.*, "Robots under COVID-19 pandemic: A comprehensive survey," *IEEE Access*, vol. 9, pp. 1590–1615, 2021.

[39] A. S. Miner, L. Laranjo and A. B. Kocaballi, "Chatbots in the fight against the COVID-19 pandemic," *npj Digital Medicine*, vol. 3, no. 65, pp. 1–4, 2020.

[40] Y. Tanoue, S. Nomura, D. Yoneoka, T. Kawashima, A. Eguchi *et al.*, "Mental health of family, friends, and co-workers of COVID-19 patients in Japan," *Psychiatry Research*, vol. 291, pp. 113067–1133069, 2020.

[41] T. J. Judson, A. Y. Odisho, J. J. Young, O. Bigazzi, D. Steuer *et al.*, "Implementation of a digital chatbot to screen health system employees during the COVID-19 pandemic," *Journal of the American Medical Informatics Association*, vol. 27, no. 9, pp. 1450–1455, 2020.

[42] A. Martin, J. Nateqi, S. Gruarin, N. Munsch, I. Abdarahmane *et al.*, "An artificial intelligence-based first-line defence against COVID-19: Digitally screening citizens for risks via a chatbot," *Scientific Reports*, vol. 10, no. 1, pp. 1–7, 2020.

[43] A. R. Dennis, A. Kim, M. Rahimi and S. Ayabakan, "User reactions to COVID-19 screening chatbots from reputable providers," *Journal of the American Medical Informatics Association*, vol. 27, no. 11, pp. 1727–1731, 2020.

[44] Z. Zhou, J. Qin, X. Xiang, Y. Tan, Q. Liu *et al.*, "News text topic clustering optimized method based on TF-iDF algorithm on spark," *Computers, Materials & Continua*, vol. 62, no. 1, pp. 217–231, 2020.

[45] S. Y. Yoo and O. R. Jeong, "EP-Bot: Empathetic chatbot using auto-growing knowledge graph," *Computer, Materials & Continua*, vol. 67, no. 3, pp. 2807–2817, 2021.

[46] A. Singh, K. Ramasubramanian and S. Shivam, "Introduction to Microsoft Bot, RASA, and Google Dialogflow," in *Building an Enterprise Chatbot*, Berkeley, CA: Apress, pp. 281–302, 2019.

[47] MongoDB, "*MongoDB Architecture Guide*," *MongoDB White Pap*, USA: MongoDB Inc., pp. 1–17, 2016.

[48] M. Grinberg, "*Flask web development*," 2nd ed., CA, USA: O'Reilly Media, Inc., 2018.

[49] R. K. Markus Hofmann, "*RapidMiner: Data Mining use Cases and Business Analytics Application*," 1st ed., Florida, USA: Chapman and Hall/CRC Press, 2014.

[50] A. Kassambara, "*Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*," 1st ed., CA, USA: CreateSpace Independent Publishing Platform, 2017.

[51] J. Wang and X. Li, "An improved KNN algorithm for text classification," in *2010 Int. Conf. on Information, Networking and Automation*, Kunming, China, pp. V2-436–V2-439, 2010.

[52] S. Ruan, H. Li, C. Li and K. Song, "Class-specific deep feature weighting for naïve Bayes text classifiers," *IEEE Access*, vol. 8, pp. 20151–20159, 2020.

[53] J. Casas, M. O. Tricot, O. Abou Khaled, E. Mugellini and P. Cudré-Mauroux, "Trends & methods in chatbot evaluation," in *ICMI 2020 Companion-Companion Publication of the 2020 Int. Conf. on Multimodal Interaction*, Virtual Event Netherlands, pp. 280–286, 2020.

[54] W. Maroengsit, T. Piyakulpinyo, K. Phonyiam, S. Pongnumkul, P. Chaovalit *et al.*, "A survey on evaluation methods for chatbots," in *Proc. of the 2019 7th Int. Conf. on Information and Education Technology*, Aizu-Wakamatsu Japan, pp. 111–119, 2019.

[55] R. V. Krejcie and D. W. Morgan, "Determining sample size for research activities," *Educational and Psychological Measurement*, vol. 30, no. 3, pp. 607–610, 1970.

[56] Razali, N. Mohd and Y. B. Wah, "Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests," *Journal of Statistical Modeling and Analytics*, vol. 2, no. 1, pp. 21–33, 2011.