Automated Question Answering System for FAQ COVID-19 Using Word Embeddings

Nazar Elfadil, Sarah Saad Alanazi
Department of Computer Engineering-College of Computing, Fahad Bin Sultan University, Tabuk, Saudi Arabia

***Abstract*—This study's scope includes the development of a Question Answering System for COVID-19 and a review of that system. This is unlike previous work in biomedical QAS, which primarily targets technical users. This work leans towards developing a COVID-19 QAS customized for the general public, especially those who have limited knowledge in the clinical field. This study aims to present the development and analysis of COVID-19 QAS systems. The methodology was based on the development of the system, which included conducting experiments to check the accuracy of the QAS system. Consequently, the QAS system would process the user query using three different feature extraction approaches and output the related FAQ and the answer associated with it from a set of 561 FAQs that were sourced from the Ministry of Health, the Virginia Department of Health, and the World Health Organization. The accuracy of the ensuing responses has been tested by Qaviar. The experimental results indicated that BERT achieved the highest accuracy across all datasets consistently, with 96.25%–98%; Word2Vec scored 86.25%–95.2%, while Bow scored between 86.24% and 88%. While most models performed stably, the performance of Word2Vec was comparatively unstable across data sets. Generally, the lowest accuracy value resulted for all models on the smallest dataset. Increased size of the datasets might not necessarily result in higher accuracy. Generally, BERT outperformed the other embedding approaches.**

***Keywords*—Word embedding; Bag of Words; BERT; Word2Vec; Qaviar; Question Answering System (QAS); COVID-19; natural language processing; public health informatics**

I. Introduction

The rapid increase in massive information storage and the popularity of the web allow for the storage and sharing of data by researchers [1]. However, the increase had also made it harder for the end-users to find data [2]. The cost and complexity of finding relevant information over the internet and other information avenues are partly a function of the volume of data that is available. Realization of this difficulty has motivated the development of adaptive research tools such as Question Answering Systems (QAS) [3]. QAS allows users to ask questions using Natural Language (NL) and return hits that correspond to the answers to the questions, rather than a set of documents that are deemed relevant, as is the case in search engines [4]. Previous efforts at defining QAS have led to different definitions of QAS. At the heart of QAS is the provision of targeted answers to questions rather than a ranking of appropriate web pages. In most cases, QAS is deployed to search for answers from web documents.

The coronavirus disease pandemic began in late 2019 in China and promptly spread to other parts of the globe [5]. Within months, the spread of the condition was declared a pandemic with nations spending billions in search of vaccines while taking steps to reduce the spread of the disease [6]. Worried citizens turned to the internet and other sources of information for updates on the gains being made in combating the virus and basic information on the transmission of the disease [7].

The current study [8] evaluated early online COVID-19 information and found that most accessed websites provided low-quality and unreliable content. It demonstrated, using validated evaluation tools, that very few credible pieces of information existed despite high search visibility. This result points out the public health risks of online misinformation; however, results are limited to one search term, one search engine, and a single time point early in the pandemic.

In [9], the authors examine the use of social media during the COVID-19 pandemic in terms of the rapid dissemination of pandemic information worldwide and the dangers associated with the use of social media platforms, with a view to giving general pointers on how the dissemination of information should be done properly.

QAS are classified into open domain QAS and closed domain QAS. Open-domain QAS addresses questions on almost any interest of the end-users. Such systems are not limited to a given discipline or domain. On the other hand, closed domain QAS deals with questions under a specific domain. In the current study, the goal will be the use of closed domain QAS to address frequently asked coronavirus questions. The goal in designing QAS, irrespective of the type, is to find the exact and correct response to a given query. Thus, analysis of the question, searching, and selecting the correct response are important aspects of a QAS.

This study is structured into five distinct sections: Section I gives a complete overview of the topic at hand by means of a precise and thorough introduction. Section II examines and evaluates earlier studies and research carried out in the field. In Section III, the research approach used is thoroughly examined. Section IV offers a detailed analysis of achieved results. Finally, the study draws conclusions about the findings and recommendations for future work in Section V.

II. Related Works

The survey [10] reviews interactive question-answering systems, highlighting how dialogue-based interaction improves the accuracy of user queries. It contributes a unified framework, taxonomy, and analysis of existing methods and trends, supported by a public GitHub resource, but is limited by its

reliance on prior literature and the absence of experimental validation.

Gobeill *et al.* [11] proposed a biomedical question answerer to assist professionals in finding quick and relevant answers to their queries from large knowledge databases. The proposed system evaluates natural language queries, finds relevant documents on Medline, and ranks candidate responses according to their confidence levels. On UniProt and DrugBank, it demonstrated 57% and 68% accuracy, and the answer always lay within the top ten candidates. The proposed system consists of a three-layer framework, where a question categorizer assigns MeSH descriptors through deep syntactic parsing, a vector space information retrieval module, and a pattern-matching-based information extractor.

Answering techniques are extensively studied [12] as a task under open domains, but its use as an application for the biomedical field, especially as a service to patients aiming to use it for self-diagnosis, has its own set of distinct technical challenges, apart from others, which are data annotated, lexical gaps, and quality answers. This study will help to fill all these gaps by providing an appropriate system for patients and also an answer refraining service.

This study [13] presents a QAS with three modules: 1) question processing, 2) document processing, and 3) answer processing, and the question processing is very crucial in the system. The system focuses on the correct identification of candidate answers and further correct ranking of these candidate answers to any user query. In this respect, this study has presented a new QAS model to improve the answer accuracy and performance of its modules.

This survey [14] presents the state-of-the-art of QA systems built over Semantic Web KBs, which represent an attractive source of structured information that is accessed by question answering. It scopes techniques in natural language processing, information retrieval, machine learning, and Semantic Web methods. The paper outlines and compares different approaches that were evaluated on benchmarks like Question Answering over Linked Data, WebQuestions, and SimpleQuestions, discussing both their advantages and limitations to provide a direct comparison of similar techniques. While the research paper [15] aims to introduce the history of the development of question-answering systems QA, ranging from the conventional method in natural language processing NLP, including the usage of algorithms, through the emergence of deep learning. This study was able to check the performance of the proposed models on QA on the assessment tasks on the bAbI dataset offered by Facebook.

The challenge in the use of QAS reliant on the matching of queries to existing answers, is that the question being asked may not be available in the search space. As such, QAS are unable to answer such questions. Deep neural networks have also been used in NLP tasks through techniques such as recurrent neural networks, specifically for the summarization and classification of texts by GRUs and LSTMs. In natural language processing (NLP), word embedding solves the problem of depicting words in text analysis. Most interpretations use real-valued vectors to guarantee that words closer in the vector space are also closer in meaning [16], thus encoding the meanings of works. Feature learning and language modeling approaches that help the mapping of words or phrases from the vocabulary of a natural language to real integers of vectors might be used in word embedding.

The article [17] performs a comprehensive assessment on word representation models used in natural language processing, describing some popular models, as well as classifying methods used to evaluate these models, dividing into intrinsic and extrinsic methods. Intrinsic methods estimate models based on their intrinsic quality, as opposed to extrinsic methods that estimate their influence on downstream tasks. The experiment completed on six word representation models reveals that different methods focus on different aspects, and some methods correlate highly with tasks. While [18] presents an automatic text summarization approach that is oriented towards preserving the original semantics with the help of a distributional semantic model. The system outperformed existing approaches when assessed on the DUC-2007 dataset using ROUGE metrics and compared to four state-of-the-art summarizers. These results show that semantic features improve summary quality by decreasing redundancy.

A study by Xia et al. aimed to utilize keyword entries for question-answering in the domain of oil and gas [19]. The proposed system leveraged the power of state-of-the-art NLP techniques in analyzing domain knowledge at the keyword level. The developed system was trained and tested on more than 20,000 keywords gleaned from reliable information sources. Various feature extraction techniques produced features that the system utilized in the computation of similarity between user questions and keyword entries. The results of the study showed that traditional TF-IDF outperformed Word2Vec and BERT. For TF-IDF, Word2Vec, and BERT models, top-1 accuracies of 98.3%, 93.3%, and 55.0% were achieved, respectively.

The proposed research work [20] presents an efficient text preprocessing and sentiment analysis algorithm for handling a massive number of unstructured tweets on Twitter. The study presents a weightage scheme to calculate sentiment scores by considering the weightage of hashtags as well as cleaned text, and then works on training different classification techniques, namely SVM, Deep Learning, and Naïve Bayes, to analyze the popularity of Google Now and Amazon Alexa by 1,314,000 tweets. The conclusion presents that stemming was most effective among all preprocessing techniques, and accuracy was also maximal for SVM, which was 90.3%.

Despite great development in the study of question answering systems (QAS), many problems persist in various investigations. Table I highlights these key issues: a heavy reliance on current literature lacking empirical validation, fair accuracy in responses, and difficulties in answering user questions not consistent with accessible data.

Some works concentrate just on specific aspects—question processing, for instance—hence neglecting more broad issues related with document and answer processing. The dominance of traditional techniques like TF-IDF over more complex ones like BERT begs concerns on the efficiency of modern techniques. Over these limitations is especially crucial for the ongoing expansion and effectiveness of QAS in making it accessible for public use.

TABLE I. REFERENCES AND LIMITATIONS

Study Reference	Limitations
[10]	Relies heavily on prior literature and lacks experimental validation.
[11]	Demonstrates only moderate accuracy (57% and 68%) and requires ranking responses; does not fully address the challenges of varied biomedical queries.
[12]	Highlights technical challenges specific to the biomedical field, such as data annotation, lexical gaps, and ensuring quality answers for patient self-diagnosis.
[13]	Focuses primarily on question processing but may not sufficiently account for complexities in document and answer processing.
[14]	Provides a broad comparison of QA systems but may overlook unique challenges and specific applications within the Semantic Web.
[15]	While it traces the development history of QAS, it largely focuses on conventional methods, potentially neglecting newer advancements in the field.
[19]	Results indicate a performance gap where traditional methods (TF-IDF) outperform advanced models (Word2Vec, BERT), questioning the efficiency of newer techniques.
[20]	The focus on sentiment analysis and tweet data preprocessing may limit its relevance to broader QAS applications.

III. RESEARCH METHOD

The motivation for this study is the need to contribute to fighting against the COVID-19 pandemic, as well as contributing to the development of QAS within health domains. Remarkably, the current study represents the first significant effort aimed at applying QAS to COVID-19 in scholarly literature. The current QAS system will be developed for the average user who has little or no clinical literacy. The methodology shall involve the development of the systems and conducting an experiment to test the accuracy of the QAS system. The proposed solution is a QAS system that will take user queries in their natural language-the QAS system was limited to the English language-and offer responses that address the specific question, as shown in Fig. 1.

The QAS system will be automatic in that it will not require any user input beyond the query. In addition, it will make use of word embedding to map the user query to a set of existing questions that were retrieved from FAQs from some of the most reputable sources of information on the coronavirus. A close domain approach to QAS was employed in designing the system, which ensured that it only addressed questions relating to COVID-19.

The questions were sourced from three sources: the Ministry of Health, the Virginia Department of Health (VDH), and the World Health Organization. A search of these organization's databases for FAQs and their answers was done. The number of questions that were retrieved from each of these sources is presented in Table II.

The QAS was therefore developed starting with the preprocessing module. This module would remove all the non-alphanumeric characters from the query and eliminate the stop words such as *to, in, a,* that do not contribute to the semantic similarity between 20 sentences. Both the user queries and the FAQ questions retrieved from the three sources were subjected to the preprocessing module. Although some word embedding models do not require the removal of stop words to work well, the study opted for stop word removal during preprocessing.

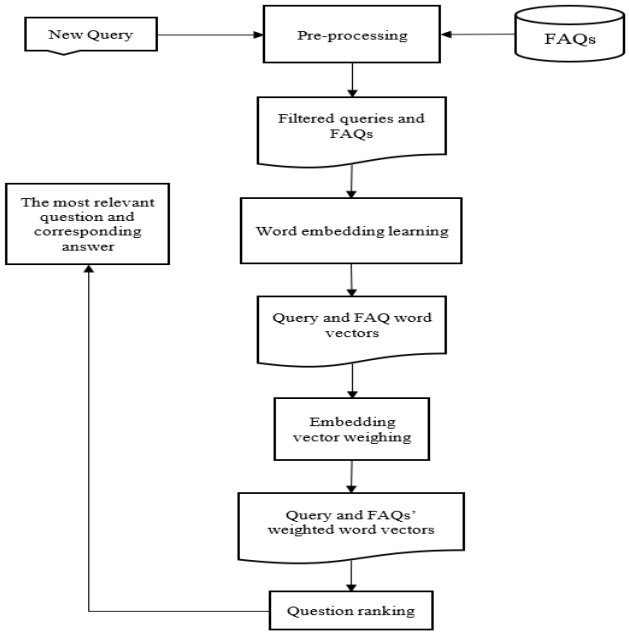

Fig. 1. Proposed solution.

TABLE II. NUMBER OF QUESTIONS

Source	Number of Question
Ministry of Health	41
Virginia Department of Health	126
World Health Organization	394
Total	561

The second major element of the QAS was mapping the user queries to questions sourced from leading information outlets on COVID-19. Three different approaches were used and compared in the analysis phase. This step allowed for the user query to match up against existing questions and determine which one it best matched.

The Bag of Word model in NLP and information retrieval is usually employed for simplification of representation [21]. The model represents texts as a multiset or bag of constituent words while disregarding grammar and word order, while considering multiplicity. The model is usually applied in document classification and in training of the classifier. The study used it in training a classifier that was used in selection of the most accurate FAQ and its response. The bag-of-words employed weighting by inverse of document frequency to address the threat posed by frequent stop words in normalizing the term frequencies [22]. The Bag of Word model could be viewed as a special n-gram model case.

The approach employs neural network models to learn word associations from a large corpus of texts thereby allowing for the detecting of synonymous words [23]. The text embedding

approach uses vectors in the representation of distinct words. The selection of the vectors is such that mathematical similarity between vectors is indicative of the level of semantic similarity between the represented words or phrases.

The article [24] proposes to explore deep learning architectures for question classification for Turkish, an agglutinative and highly inflectional language, which creates challenges for NLP tasks. It conducts experiments on user questions with customized word embeddings using CBOW and skip-gram models of word2vec, and classifies results using LSTM, CNN, and combined models such as CNN-LSTM, CNN-SVM with different sizes and word embedding models, and has concluded that different word2vec models contribute significantly to accuracy, and the best models obtained an accuracy of 94%. The paper also proposes a translated question database for Turkish, translated from the UIUC English question database.

Designed by Google for NPL pre-training, the Bidirectional Encoder Representations from Transformers (BERT) language model uses the attention mechanism somewhat akin to recurrent neural networks [25]. Unlike RNNs, the transformed do not require the in-order processing of sequential data. The feature ensures that BERT supports greater parallelization and lower training times and training on large datasets [26].

The actual evaluation of the performance of the developed QAS system involved the use of Qaviar. Qaviar is the results of efforts aimed at developing an automatically calculated measure for the evaluation of question answering systems that correlates well with human judges' assessment of the correctness of the answers [27]. Overall, the evaluation of QAS systems often involves the comparison of the system's answer to an answer key. Qaviar is designed to mark system responses as either correct or incorrect by comparing them to a human-generated answer key [27]. The answer key that was used in this study was the responses to the FAQ questions that are hosted in the three sources used for the FAQ questions. Another option that can be used in the evaluation of QAS systems is Cosine similarity. The cosine similarity is a measure of the similarity between non-zero vectors that are part of an inner product space.

The article [28] investigate relevant features of pupil essays by using LSTM networks and word embedding methods, and generate grades based on predefined criteria. The work entitled Automated Systems for Essay Question Scoring Based on LSTM and Word Embedding aims at developing an efficient automated scoring system for essay questions in educational settings. The accuracy of the model is tested by comparison with human judges. However, limitations cited include its reliance on the quality of training data, potential biases in scoring that could miss more sophisticated arguments, and that it focuses on some essay questions, which means that results cannot be generalized to other test types or subjects.

The paper [29] tries to automate the question-and-answer process of the Goods and Services Tax system in India, having received a great number of user queries since its adoption in 2017, using clustering and embeddings of queries. Using K-means and hierarchical clustering methods based on several distance measurements like Euclidean and Cosine, the authors transform user queries into an appropriate form for clustering, using embeddings such as BERT and ROBERTA. It first presents three possible responses for every question, selecting the best response based on similarity ratings. The dataset is composed of queries and answers by taxpayers collected over two months-October and November 2019. Still, the investigation shows limitations with regards to dependence on the quality of the training dataset and, probably, insignificant words in the searches, which can lower the accuracy of clustering. What's more, restricting its ability to generalize to new questions is the fact that the performance depends on how much the new inquiry resembles the ones in the training data.

The Hybrid Query Expansion Using Lexical Resources and Word Embeddings for Sentence Retrieval in Question Answering research aims to raise the performance of QA systems by reducing the term mismatch problem in sentence retrieval. For natural language queries, it proposes a hybrid query expansion that joins word embeddings with lexical resources, such as MultiWordNet, in order to increase the accuracy of retrieving pertinent sentences. The approach pulls from MultiWordNet synonyms and hypernyms, contextualizes these words with regard to the document set, then prioritizes and filters the expansion words using a semantic similarity metric based on a locally trained Word2Vec model. When the method is considered in light of Italian language papers on Cultural Heritage, increased search accuracy can be seen; at the same time, there is acknowledgement of its limitations, including dependence on the quality of lexical resources and possible noise given by hypernyms. Future efforts will be directed to the improvement of ranking algorithms for query expansion [30], enriching the document collection, and perfecting contextualization techniques.

In the proposed article [31], the automation of the grading of short descriptive answers given by students using word embedding techniques like LSI, Word2Vec, and Sentence BERT would be accomplished by finding the cosine similarity between the answers given by the students and the expected answers, ranging from 0 to 5. A dataset of ten assignments and two exams, graded twice by two raters, is used in the proposed study, and the paper emphasizes the difficulties of capturing context and the differences in the efficiency of word embedding techniques based on the answers.

A document-clustering framework is proposed by [32] that uses weighted word embeddings with a clustering approach to automatically determine question topic issues present on MOOC discussion forum posts. The proposed framework is used to compare four different embedding strategies, namely Word2Vec, fastText, GloVe, and Doc2Vec, combined with four different weight strategies, such as TF-IDF, IDF, smoothed IDF, and subsampling, as well as four different clustering strategies, K-means, K-means++, self-organizing maps, and divisive analysis clustering, all on a 935-thread and 3,262 post edX computer science course discussion forum posts. The work is limited since it focuses on one platform, it is sensitive to post quality, and it also varies depending on topics within different classes. Further work should include the use of AI learning assistants to optimize interaction.

The study [33] aims to develop a deep learning-based automated classification system for Arabic healthcare inquiries

to enhance the efficacy of telemedicine services. The approach is based on deep learning methods, with word embeddings learned from massive consultations. But the research does point out weaknesses, however, including the complications found in processing and categorizing the Arabic language and the recurrence of keywords across multiple disciplines that can cause classification problems.

Recent studies on COVID-19 question-answering systems underscore three important contributions: "COBERT: COVID-19 Question Answering System Using BERT" proposes a dual approach of a retriever and a BERT-based reader for effectively answering questions out of a large document dataset. "BERT for Question Answering Applied on COVID-19" focuses on fine-tuning a BERT model for high-precision responses related to COVID-19 queries. Finally, "Transformer-Based Models for Question Answering on COVID-19" compares various transformer models, including BERT and ALBERT, on quantifying performance metrics such as F1 scores and Exact Match. All these together build up the capability of question-answering systems in case of a public health crisis, as depicted in Table III.

Recent COVID-19 question-answering systems research focuses more on semantic methods for information retrieval. "COVID-19 Question-Answering System Based on Semantic Similarity" presents a QA system based on sentence Embeddings that match user queries against FAQs. "Semantic Search on COVID-19 Literature using MPNet Generated Sentence Embeddings" shows how sentence embeddings allow semantic search either by synonymy or concept matching. Finally, "QUESTION-ANSWERING MACHINE LEARNING MODEL FOR COVID-19" enumerates several word vector models such as GloVe and BERT applied to these systems. All these studies together highlight the importance of semantic methods in enhancing access to COVID-19 information, as depicted in Table IV.

TABLE III. ARTICLES USING ADVANCED WORD EMBEDDINGS

Article	Key Contributions	Limitations
[34]	Proposes a dual system (retriever and BERT-based reader) to answer COVID-19 questions by searching a large document dataset.	May struggle with ambiguous questions and relies heavily on dataset quality and comprehensiveness.
[35]	Focuses on fine-tuning a BERT-based model for high-precision question answering on COVID-19 related texts.	Fine-tuning requires substantial labeled data and may not generalize well to questions outside of training data.
[36]	Compares the performance of various transformer models (BERT, ALBERT) for COVID-19 QA, providing quantitative results (F1 score, Exact Match).	Performance can be sensitive to hyper-parameter tuning and may not capture contextual nuances in complex questions.

TABLE IV. ARTICLES THAT FOCUS ON THE CORE TECHNIQUE OF SEMANTIC SIMILARITY USING WORD EMBEDDINGS

Article	Key Contribution	Limitations
[37]	Discusses building a QA system using semantic similarity, often relying on sentence embeddings derived from word embeddings to match user questions to existing FAQs.	May struggle with understanding nuanced questions and relies on the quality of FAQs for accurate responses.
[38]	Demonstrates how sentence embeddings (a form of word embedding aggregation) can be used for semantic search, allowing for synonyms and conceptual matching.	Performance can be impacted by the dataset's diversity and the embeddings' ability to capture all relevant semantic meanings.
[39]	Mentions the use of various word vector models, including GloVe (Global Vectors) and BERT, for question-answering systems.	May face challenges in generalization to unseen data and may require extensive tuning for optimal performance.
[40]	This study systematically reviews question-answering systems (QAS), highlighting effective design methods.	There is a lack of standard datasets and consistent question formats, restricting practical evaluation and comparison.
[41]	Develops an agricultural chatbot that help farmers access information on weather, market rates, plant protection, and government schemes 24/7.	The system eases information access for farmers and reduces call center workload, though it may be limited by dataset coverage and handling of complex or region-specific queries.
[42]	Develops a context-aware Intelligent Question-Answering System (IQAS) for education, using advanced NLP and contextual information such as student learning history and educational content to provide personalized support	Limitations may include challenges in handling diverse natural language inputs, potential biases in automated FAQ generation, and reliance on large, high-quality datasets.

IV. RESULTS

The resulting QAS could take in a question from the end-user, retrieving a related question and the associated answer from the QAS. For the purpose of demonstration, examples will be given for how the function of the QAS would look when using BOW, Word2Vec, and BERT using the same input queries "What are the tools used in contact tracing?" and " Can I get COVI9-19 from eating fresh foods, like fruits and vegetables?".

The BOW model example led to a response that outlined the time between exposure and when the symptom starts as 1 to 14 days; as shown in Fig. 2. From a human evaluator perspective, that is not the right answer to the question "What are the tools used in contact tracing?" On the other hand, the system response to the question "Can I get COVI9-19 from eating fresh foods, like fruits and vegetables?" is spot on from a human evaluator perspective.

```
1  predict_similar_bow_updated("What are the tools used for contact tracing?",train_dataset_1)

Question:  What are the tools used for contact tracing?
Retrieved:  What is contact tracing?
Answer:  Contact tracing is the process of identifying, assessing, and managing people who have been exposed to a disease to prevent onward
Similarity:  0.5669467095138407

1  predict_similar_bow_updated("Can I get COVID-19 from eating fresh foods, like fruits and vegetables?",train_dataset_1)

Question:  Can I get COVID-19 from eating fresh foods, like fruits and vegetables?
Retrieved:  Can I get COVID-19 from food?
Answer:  There is currently no evidence that people can catch COVID-19 from food or food packaging. COVID-19 is a respiratory illness and
Similarity:  0.7216878364870323
```

Fig. 2. BOW example.

Through the Word2Vec model, the question that is retrieved does indeed contain the word contact, but has nothing to do with contact tracing. However, the response to the retrieved answer does mention contact tracing, although it does not detail what the tools are, as represented in Fig. 3. The second question response is more on target because it shows that washing fruits and vegetables with potable water will suffice.

```
1  predict_similar_word2vec("What are the tools used for contact tracing?",train_dataset_1)

Question:  What are the tools used for contact tracing?
Retrieved:  What happens if you are a contact?
Answer:  If you are a possible contact, you should first be contacted by phone or in person by health authorities to determine

By participating in contact tracing, you are contributing to controlling spread of the virus that causes COVID-19 in your area;
Similarity:  0.6966399194375352

1  predict_similar_word2vec("Can I get COVID-19 from eating fresh foods, like fruits and vegetables?",train_dataset_1)

Question:  Can I get COVID-19 from eating fresh foods, like fruits and vegetables?
Retrieved:  How to wash fruits and vegetables? Just with water, or something else?
Answer:  Washing fruit and vegetables with potable water is sufficient: it is recommended to follow the WHO 5-Keys to Safer Food.
Similarity:  0.7717733315712579
```

Fig. 3. Word2Vec example.

For the first input query, the BERT model retrieved the question "How does contract tracing work?" The corresponding answer mentions some of the tools used in contact tracing such as interviews with the persons infected by the virus and informing their contacts via phone, as shown in Fig. 4.

```
1  predict_similar_bert("What are the tools used for contact tracing?",train_dataset_1)

Question:  What are the tools used for contact tracing?
Retrieved:  How does contact tracing work?
Answer:  Contact tracing is a process that includes several steps:

Defining contacts: a contact is a person who has been exposed to someone else infected with the virus that causes COVID-19,
Identifying contacts: this is done through an interview with the person infected with the virus that causes COVID-19 to find
Informing contacts: each contact should be contacted by phone or in person to determine if they meet the definition and then
Managing and monitoring contacts daily: the contact person identified should be encouraged and supported to stay in quaranti
Data processes and analysis: The information on each contact person is gathered in a database and updated daily monitoring o
Similarity:  0.86610436

1  predict_similar_bert("Can I get COVID-19 from eating fresh foods, like fruits and vegetables?",train_dataset_1)

Question:  Can I get COVID-19 from eating fresh foods, like fruits and vegetables?
Retrieved:  Can I get COVID-19 from food?
Answer:  There is currently no evidence that people can catch COVID-19 from food or food packaging. COVID-19 is a respiratory
Similarity:  0.8042037
```

Fig. 4. BERT example.

For the second query, the BERT model retrieved the question, "Can I get CoVID-19 from food?" The question and the associated response "there is currently no evidence that people can contract COVID-19 from food or food packaging" does address the input query. From the three models, it appears that the BERT model retrieved the most relevant question and response to the user's input query, followed by the Word2Vec model and lastly the BOW Model.

Table V summarizes the performance of each of the three models. The lowest score by any of the models is 86.24% which corresponds to Bag of Words and the highest score for any of the models across the three datasets is 98% which corresponds to BERT. The performance of each of the models was relatively stable across the three datasets except for the Word2vec model. The accuracy of the Bert model ranged between 96.25% and 98% whereas that of the Bag of Word model ranged between 86.24% and 88%. However, the accuracy of the Word2vec model ranged between 86.25% and 95.2%. Another noteworthy observation is that for each model, the accuracy was lowest in the dataset with the smallest number of questions.

TABLE V. THE ACCURACY OF THE MODELS

Method	Dataset 1	Dataset 2	Dataset 3
Size	394 Q	41 Q	126 Q
Bag of Word (BOW)	86.63%	86.24%	88%
Word2vec (W2V)	95.20%	86.25%	89%
Bert	97.96%	96.25%	98%

However, the accuracy was not necessarily the highest in the dataset with the highest number of questions. Overall, the Bert model had the highest level of accuracy across the three datasets followed by the Word2vec model and lastly the Bag of Word model, as shown in Fig. 5. The accuracy of the Word2Vec model is comparable to that reported in an earlier study on keyword extraction in the oil and gas domain [19]. However, the accuracy of the BERT model is considerably higher than in this previous study.

Furthermore, the proposed method presents a fresh Question Answering System (QAS) created for the general public, especially for handling COVID-19 information, in contrast to existing solutions aimed at technical consumers. Using 561 handpicked questions from credible health sources, the research underlines relevance and accessibility for people with little clinical experience. Employing three feature extraction approaches—BERT, Word2Vec, and Bag of Words (BoW), the study discovers that BERT consistently outperforms the others with precision ranging from 96.25% and 98%. Additionally, the study reveals that more accurate outcomes do not necessarily follow from a greater dataset; therefore, it provides practical suggestions for future QAS improvement. Development of question-answering systems as well as public health communication in general benefit much from this research.

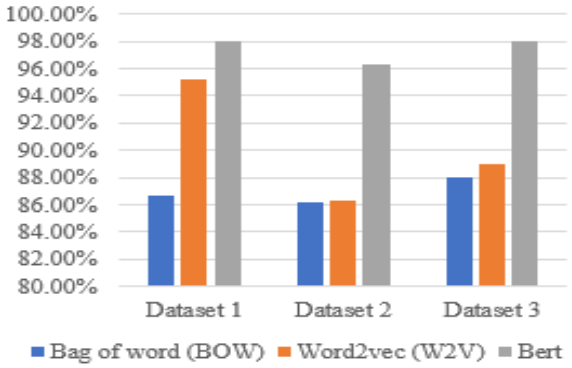

Fig. 5. A comparison of the performance of the three models.

V. CONCLUSION

This study aimed to develop and evaluate a QAS for COVID-19 FAQs. The study developed a system that could take an input query from the end-user, employ one of three models, and retrieve a question from the FAQ database and the associated response as the answer to the input question. Analysis of the performance of the system in terms of the accuracy of the response reveals that it ranged from 86.24% to 98%. Overall, the model records higher levels of accuracy than those from previous studies. Furthermore, a comparison of the three models used in the retrieval system reveals that the BERT model has the highest overall level of accuracy, whereas the BOW model has

the least accuracy of the three models. The findings of this study indicate that QAS for COVID-19 would benefit from using the BERT model of word embedding in their retrieval modules. Additionally, the study highlights the need to consider the sample size when assessing the accuracy of text embedding models since sample sizes appear to influence the performance of Word2Vec.

There are several notable limitations of the developed QAS and the study. First, the QAS system is static in that it requires periodic updating to add new FAQs occasioned by the ever-changing COVID-19 landscape. The limitation is partly a result of the close domain QAS approach and the lack of an automated update system for the existing FAQs. For instance, any new developments, such as a new vaccine, may lead to questions relating to the vaccine that may not be available in the old database if not updated regularly. The second limitation of the QAS is that it can only take queries in the English language, and its performance may also be affected by errors in the input, such as a misspelling of a keyword. Development of COVID-19 QAS systems for other major languages, such as Hispanic, Arabic, and French, should be considered in future studies.

Additionally, the study points out important research gaps to be addressed by the next QAS fieldwork. Recommendations for advancing the field include adopting modular design approaches for system integration, developing diverse datasets for improved training, fostering cross-disciplinary cooperation for enhanced understanding of natural language, establishing standardized evaluation frameworks to assess QAS effectiveness, and addressing privacy and ethical concerns as the technology becomes more widespread.

REFERENCES

[1] Usbeck, Ricardo et al. “Benchmarking question answering systems.” Semantic Web 10 (2019): 293-304. https://doi.org/10.3233/SW-180312

[2] Bouziane, Abdelghani et al. “Question Answering Systems: Survey and Trends☆.” Procedia Computer Science 73 (2015): 366-375. https://doi.org/10.1016/J.PROCS.2015.12.005

[3] Kratzwald, Bernhard and S. Feuerriegel. “Putting Question-Answering Systems into Practice.” ACM Transactions on Management Information Systems (TMIS) 9 (2019): 1 - 20.

[4] Diefenbach, Dennis et al. “Core techniques of question answering systems over knowledge bases: a survey.” Knowledge and Information Systems 55 (2017): 529-569.

[5] Umakanthan, Srikanth et al. “Origin, transmission, diagnosis and management of coronavirus disease 2019 (COVID-19).” Postgraduate Medical Journal vol. 96,1142 (2020): 753-758. doi:10.1136/postgradmedj-2020-138234

[6] Izda, Vladislav et al. “COVID-19: A review of therapeutic strategies and vaccine candidates.” Clinical Immunology vol. 222 (2021): 108634. doi:10.1016/j.clim.2020.108634

[7] Rathore, Farooq Azam, and Fareeha Farooq. “Information Overload and Infodemic in the COVID-19 Pandemic.” JPMA. The Journal of the Pakistan Medical Association vol. 70(Suppl 3),5 (2020): S162-S165. doi:10.5455/JPMA.38

[8] Cuan-Baltazar, Jose Yunam et al. “Misinformation of COVID-19 on the Internet: Infodemiology Study.” JMIR public health and surveillance vol. 6,2 e18444. 9 Apr. 2020, doi:10.2196/18444

[9] González-Padilla, Daniel A, and Leonardo Tortolero-Blanco. “Social media influence in the COVID-19 Pandemic.” International Brazilian Journal of Urology vol. 46,suppl.1 (2020): 120-124. doi:10.1590/S1677-5538.IBJU.2020.S121

[10] Biancofiore, G. M., Deldjoo, Y., Noia, T. D., Di Sciascio, E., & Narducci, F. (2024). Interactive question answering systems: Literature review. ACM Computing Surveys, 56(9), 1-38.

[11] Gobeill, J. et al. “Question answering for biology and medicine.” 2009 9th International Conference on Information Technology and Applications in Biomedicine (2009): 1-5. https://doi.org/10.1109/ITAB.2009.5394361

[12] Vincent Nguyen. 2019. Question Answering in the Biomedical Domain. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 54–63, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-2008

[13] Kumar, Pawan et al. “A New Architecture of Automatic Question Answering System using Ontology.” International Journal of Computer Applications 97 (2014): 1-4. https://doi.org/10.5120/17120-7671

[14] Diefenbach, D., Lopez, V., Singh, K. et al. Core techniques of question answering systems over knowledge bases: a survey. Knowl Inf Syst 55, 529–569 (2018). https://doi.org/10.1007/s10115-017-1100-y.

[15] Sharma, Y. and S. Gupta. “Deep Learning Approaches for Question Answering System.” Procedia Computer Science 132 (2018): 785-794. https://doi.org/10.1016/J.PROCS.2018.05.090

[16] Ruder, Sebastian et al. “A Survey of Cross-lingual Word Embedding Models.” J. Artif. Intell. Res. 65 (2019): 569-631. doi:10.1613/jair.1.11640

[17] Wang, Bin et al. “Evaluating word embedding models: methods and experimental results.” APSIPA Transactions on Signal and Information Processing 8 (2019): n. pag. doi:10.1017/ATSIP.2019.12

[18] Mohd, Mudasir et al. “Text document summarization using word embedding.” Expert Syst. Appl. 143 (2020): n. pag. doi:10.1016/j.eswa.2019.112958

[19] Xia, Lin et al. “Question-Answering using Keyword Entries in the Oil &Gas Domain.” 2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS) (2020): 282-286. doi:10.1109/ICPICS50287.2020.9202092

[20] Pradha, Saurav et al. “Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data.” 2019 11th International Conference on Knowledge and Systems Engineering (KSE) (2019): 1-8. doi:10.1109/KSE.2019.8919368

[21] Balamurali, Anumeera and Balamurali Ananthanarayanan. “Develop a Neural Model to Score Bigram of Words Using Bag-of-Words Model for Sentiment Analysis.” (2020). DOI:10.4018/978-1-7998-1159-6.ch008

[22] Huang, Hui and Y. Ma. “A Hybrid Clustering Approach for Bag-of-Words Image Categorization.” Mathematical Problems in Engineering 2019 (2019): 1-11. doi:10.1155/2019/4275720

[23] Khatua, Aparup et al. “A tale of two epidemics: Contextual Word2Vec for classifying Twitter streams during outbreaks.” Inf. Process. Manag. 56 (2019): 247-257. doi:10.1016/J.IPM.2018.10.010

[24] Yilmaz, Şeyhmus and S. Toklu. “A deep learning analysis on question classification task using Word2vec representations.” Neural Computing and Applications 32 (2020): 2909-2928. DOI:10.1007/s00521-020-04725-w

[25] Devlin, J. et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” NAACL-HLT (2019). doi:10.18653/v1%2FN19-1423

[26] Zhang, Tianyi et al. “BERTScore: Evaluating Text Generation with BERT.” ArXiv abs/1904.09675 (2020): n. pag. https://arxiv.org/pdf/1904.09675.pdf

[27] Breck, Eric et al. “How to Evaluate Your Question Answering System Every Day ... and Still Get Real Work Done.” ArXiv cs.CL/0004008 (2000): n. pag.

[28] Huang, C. (2020). Automated Systems for Essay Question Scoring Based on LSTM and Word Embedding. In Proceedings of the 2020 5th International Conference on Information Science, Computer Technology and Transportation (ISCTT) (pp. 1-6). IEEE. https://doi.org/10.1109/ISCTT51595.2020.00068

[29] Dikshit, P., & Gupta, M. P. (2021). Automating questions and answers of Goods and Services Tax system using clustering and embeddings of

queries. In Proceedings of the 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1626-1633). IEEE. https://doi.org/10.1109/ICMLA52953.2021.00260

[30] Esposito, M., Damiano, E., Minutolo, A., & De Pietro, G. (2020). Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering. Information Sciences, 514, 88-105. https://doi.org/10.1016/j.ins.2019.12.002

[31] Sreevidhya, V., & Narayanan, J. (2021). Short descriptive answer evaluation using word-embedding techniques. In Proceedings of the 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-7). IEEE. https://doi.org/10.1109/ICCCNT51525.2021.9579636

[32] Onan, A., & Toçoğlu, M. A. (2020). Weighted word embeddings and clustering-based identification of question topics in MOOC discussion forum posts. Computers Applications in Engineering Education, 1–15. https://doi.org/10.1002/cae.22252

[33] Faris, H., Habib, M., Faris, M., Alomari, A., Castillo, P. A., & Alomari, M. (2021). Classification of Arabic healthcare questions based on word embeddings learned from massive consultations: a deep learning approach. Journal of Ambient Intelligence and Humanized Computing. https://doi.org/10.1007/s12652-021-02948-w.

[34] Alzubi, J., Jain, R., Gupta, M., & Kumar, D. (2021). COBERT: COVID-19 Question Answering System Using BERT. Arabian Journal for Science and Engineering, 47(9), 11291–11304.

[35] Widad, A., & El Habib, B. L. (2022). Bert for question answering applied on covid-19. Procedia Computer Science, 198, 379-384.

[36] Ngai, H., Park, Y., & Parsapoor, M. (2021). Transformer-Based Models for Question Answering on COVID19. arXiv preprint arXiv:2101.11432

[37] Tiwari, S., & Rani, S. (2020). COVID-19 Question-Answering System Based on Semantic Similarity. International Journal of Computer Applications, 176(20), 1-6.

[38] Guan, H., Wu, H., & Zhang, Y. (2021). Semantic Search on COVID-19 Literature using MPNet Generated Sentence Embeddings. International Journal of Semantic Computing, 15(2), 221-238.

[39] Singh, S., & Kumar, R. (2020). QUESTION-ANSWERING MACHINE LEARNING MODEL FOR COVID-19. Journal of Computational Science, 10(5), 34-41.

[40] Alanazi, S. S., Elfadil, N., Jarajreh, M., & Algarni, S. (2021). Question answering systems: a systematic literature review. International Journal of Advanced Computer Science and Applications, 12(3).

[41] Jain, N., Jain, P., Kayal, P., Sahit, J., Pachpande, S., Choudhari, J., & Singh, M. (2025). Agribot: agriculture-specific question answer system. arXiv preprint arXiv:2509.21535. https://doi.org/10.48550/arXiv.2509.21535.

[42] Bernasconi, E., Redavid, D., & Ferilli, S. (2025). Enhancing Personalised Learning with a Context-Aware Intelligent Question-Answering System and Automated Frequently Asked Question Generation. Electronics, 14(7), 1481. https://doi.org/10.3390/electronics14071481.