A Qualitative Evaluation of Language Models on Automatic Question-Answering for COVID-19

David Oniani
Mayo Clinic
Kern Center for the Science of Health Care Delivery
Rochester, MN, USA
oniani.david@mayo.edu

Yanshan Wang
Mayo Clinic
Division of Digital Health Sciences
Rochester, MN, USA
wang.yanshan@mayo.edu

ABSTRACT

COVID-19 (2019 Novel Coronavirus) has resulted in an ongoing pandemic and as of 26 July 2020, has caused more than 15.7 million cases and over 640,000 deaths. The highly dynamic and rapidly evolving situation with COVID-19 has made it difficult to access accurate, on-demand information regarding the disease. Online communities, forums, and social media provide potential venues to search for relevant questions and answers, or post questions and seek answers from other members. However, due to the nature of such sites, there are always a limited number of relevant questions and responses to search from, and posted questions are rarely answered immediately. With the advancements in the field of natural language processing, particularly in the domain of language models, it has become possible to design chatbots that can automatically answer consumer questions. However, such models are rarely applied and evaluated in the healthcare domain, to meet the information needs with accurate and up-to-date healthcare data. In this paper, we propose to apply a language model for automatically answering questions related to COVID-19 and qualitatively evaluate the generated responses. We utilized the GPT-2 language model and applied transfer learning to retrain it on the COVID-19 Open Research Dataset (CORD-19) corpus. In order to improve the quality of the generated responses, we applied 4 different approaches, namely tf-idf (Term Frequency - Inverse Document Frequency), Bidirectional Encoder Representations from Transformers (BERT), Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT), and Universal Sentence Encoder (USE) to filter and retain relevant sentences in the responses. In the performance evaluation step, we asked two medical experts to rate the responses. We found that BERT and BioBERT, on average, outperform both tf-idf and USE in relevance-based sentence filtering tasks. Additionally, based on the chatbot, we created a user-friendly interactive web application to be hosted online and made its source code available free of charge to anyone interested in running it locally, online, or just for experimental purposes. Overall, our work has yielded significant results in both designing a chatbot that produces high-quality responses to COVID-19-related questions and comparing several embedding generation techniques.

KEYWORDS

covid-19, cord-19, gpt-2, bert, biobert, use, tf-idf, dataset, nlp, ai, semantic similarity

ACM Reference Format:
David Oniani and Yanshan Wang. 2020. A Qualitative Evaluation of Language Models on Automatic Question-Answering for COVID-19. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '20), September 21–24, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3388440.3412413

1 INTRODUCTION

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1]. As of 26 July 2020, more than 15.7 million cases have been recorded, resulting in over 640,000 deaths [2]. The sudden global outbreak of COVID-19 sent millions of people into quarantined, due to the social distancing measures. Additionally, the COVID-19 pandemic caused a historic rise in mental health problems, such as depression, post-traumatic stress disorder, and suicide, due to the state-wide quarantine. People are isolated and stressed, and may develop long-term psychological consequences, beyond the quarantine period [3] [4] [5]. Therefore, most of the time, people rely on online and web-based resources for getting news and updates concerning COVID-19. Given that many web sources currently do not contain accurate information about the pandemic and misinformation campaigns are running rampant [6], it is critically important that people and patients receive accurate, up-to-date, and useful information regarding COVID-19. Online communities, forums, and social media provide potential venues to search for relevant questions and answers, or post questions and seek answers from other members. However, due to the nature of such sites, there are a limited number of relevant questions and responses to search from, and posted questions are rarely answered immediately. To address these issues, we propose to develop a chatbot enhanced by neural language models that is able to automatically answer questions related to COVID-19 through conversational interactions.

A conversational chatbot is a software which is able to conduct a conversation via text and/or other means. There are different taxonomies for the type of conversational chatbot. Based on how the natural language conversations are generated, there are two main categories: script chatbot and intelligent chatbot. The entire

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

BCB '20, September 21–24, 2020, Virtual Event, USA

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-7964-9/20/09…$15.00
https://doi.org/10.1145/3388440.3412413

interaction in a script chatbot is based on a pre-determined model that determines what the chatbot can and cannot do. The "script" is usually a decision tree that is manually crafted by domain experts to determine which specific path to take given a response to one question task. It is usually very labor-intensive and nongeneralizable to develop conversation decision trees. The intelligent chatbot is built using Artificial Intelligence (AI) and Natural Language Processing (NLP) techniques that automatically generate natural language on the back end. With the advancements in AI and NLP, the functionality and the performance of modern chatbots have been dramatically improved. However, these techniques are rarely applied and evaluated in the healthcare domain to meet the information needs with accurate, up-to-date, and interactive healthcare information.

The outbreak of COVID-19 has motivated us to develop a chatbot with advanced NLP techniques and evaluate the approach in automatically answering questions related to COVID-19. To the best of our knowledge, this is the first study of such kind. Our contributions are:

- We applied and compared the performance of four embedding generation approaches, namely tf-idf (Term Frequency - Inverse Document Frequency) [7], Bidirectional Encoder Representations from Transformers (BERT) [8], BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) [9], and Universal Sentence Encoder (USE) [10] for refining the automatically generated answers.
- We proposed a qualitative evaluation guideline for automatic question-answering for COVID-19.
- We assessed the performance of the proposed "hybrid" approach for automatic question-answering for COVID-19.
- We built a web-based chatbot using the language models that facilitate question-answering for users.

This paper is organized as follows. We will proceed by discussing the related work and the efforts in Section 2. Section 3 will be dedicated to materials and Section 4 the proposed approach. We will report the chatbot evaluation strategy and the experimental results in Sections 5 and 6, respectively. Finally, we will also discuss a web-based chatbot with the proposed model and future work in Section 7, and conclude the work in Section 8.

2 RELATED WORK

Recent neural language models of dialogue generation offer great promise for generating responses for intelligent chatbots. The LSTM (Long Short-Term Memory) sequence-to-sequence (seq2seq) model is one type of neural generation model that maximizes the probability of generating a response given the previous dialogue turn [11] [12] [13]. XLNet uses a context of the word for predicting the next word where the context word is constrained to two directions (backward or forward) [14]. Self-Attentive Associative Memory (SAM) is a technique where two memories are wired into a single sequential model capable of both memorization and relational reasoning [15].

In the GPT-2 domain, Lee and Hsiang [16] have fine-tuned GPT-2 for generating patent claims. Klein and Nabi [17] have applied GPT-2 in conjunction with BERT for automatic question generation purposes. Zhang, Sun, et al. developed a large and tunable neural conversational model DialoGPT using GPT-2 [18]. Lee, Shu et al. developed RecipeGPT for automatic generation of cooking recipes by fine-tuning GPT-2 on a large cooking recipe dataset [19]. We are unaware of the work which applied GPT-2 model for transfer learning purposes on CORD-19.

In regard to the work related to comparing pretrained AI models, Jin et al. made some efforts to conduct probing experiments and compare BERT, ELMo [20], and BioBERT. Sharma and Daniel [21] compared the performance of BERT networks to that of FLAIR [22].

In the general AI-based chatbot domain, Serbal et al. [23] have applied deep reinforcement learning for building a conversational AI chatbot. Adiwardana et al. [24] have developed a multi-turn open-domain chatbot trained end-to-end on data mined social media conversations. Yin et al. [25] have developed a deep learning based chatbot for psychological therapy purposes.

Semantic similarity of texts, on the other hand, has been studied for a long time and recent breakthroughs allowed for development of new models such as BERT, BioBERT, and Universal Sentence Encoder (USE). Today, one of the state-of-the art conversational AI models is GPT-2. GPT-2 is a pretrained model, so we have applied transfer learning utilizing CORD-19 for retraining purposes. The resulting chatbot gave irregularly long responses that would not be typical of a human. We have therefore decided to further filter the responses by applying embedding generation algorithms and models such as tf-idf, BERT, BioBERT, and USE and then using semantic similarity approaches such as cosine similarity and inner product. In other words, we first let a human ask a question and make GPT-2 come up with an answer. We then further process the response with additional filters and ultimately, apply an embedding generation model for finding the sentences that are most relevant to the question.

Cosine similarity is one of the most commonly used approaches in calculating semantic similarity of texts. Therefore, it is naturally employed in NLP tasks. Many NLP applications need to compute the semantic similarity between two short texts. Its flexibility allows one to apply it under virtually any settings, as long as documents can be represented as vectors. Besides, finding cosine similarity is usually not a time-consuming task and can be done really quickly. Therefore, it is also commonly used for benchmarking purposes [26].

Our study has produced a chatbot that is both performant and extensible. Additional layers of filters have shown success in classifying sentences. The chatbot is also able to be retrained and readjusted to new data, in case there are new discoveries and scientific achievements related to COVID-19. Furthermore, chatbot responses have been annotated by medical experts and the results were consistent across the annotators.

3 MATERIALS

The White House Office of Science and Technology Policy alongside with the coalition of leading research groups has released a COVID-19 machine readable dataset - COVID-19 Open Research Dataset (CORD-19) [27]. It consisted of over 128,000 scholarly articles regarding COVID-19, SARS-CoV-2, and related coronaviruses, including over 59,000 with full text, and called researchers globally to develop text and data mining tools for finding answers to the

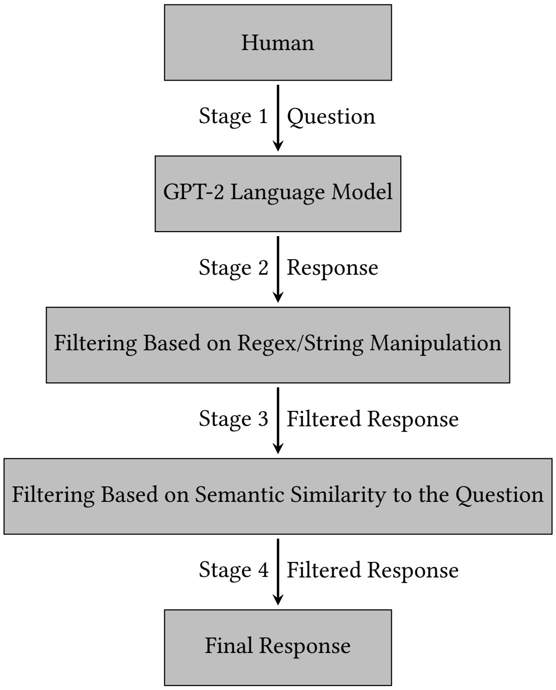

Figure 1: Workflow for Response Generation.

questions within this content in support of the ongoing COVID-19 response efforts worldwide [28].

We used CORD-19 to train a language model that would automatically answer questions related to COVID-19. The chatbot would not only help improve information acquisition, but also serve as a knowledge base for COVID-19. We harvested the data from the initial *commercial use subset* of CORD-19, containing 9000 scholarly articles in the form of JSON files. We extracted the abstract and the main body of the article from every JSON file, combined them together, and used as a corpus for retraining the unsupervised language model.

4 METHODS

We applied a hybrid approach for generating responses: GPT-2 was used to generate answers to questions, then an additional filtering step was applied for pruning the irrelevant sentences from the answers, and subsequently, semantic similarity methods were employed to retain the sentences that are most semantically similar to the question. Such hybrid approach to the response generation produced high quality responses to COVID-19-related questions. Figure 1 illustrates the pipeline of the proposed approach.

4.1 GPT-2 Language Model

GPT-2 has a Transformer-based [29] architecture which, in many ways, is similar to Open AI GPT model[30][31].

There are a total of 4 different GPT-2 models that were released by OpenAI: 124 million (124M), 355 million (355M), 774 million (774M), and 1.5 billion (1.5B) parameters [32] models. While the model with 1.5 billion parameters showed the best results in the original paper [30], in our experiments, we found that it was difficult to fine-tune and use for the transfer learning purposes. Additionally, the training was unbearably slow, even if run on TPUs (Tensor Processing Unit) provided by Google Colaboratory [33] which we used as our training ground.

We therefore utilized 774M model and ran transfer learning for 2500 iterations with the batch size of 8. After 2000 iterations, the loss was not decreasing so we let the language model train for the additional 500 iterations and stopped the training. The batch size of 8 was chosen due to the memory limitations of Google Colaboratory. As for the optimizer, we used Adam [34] and set the learning rate of 0.0001 (`1e-4`).

Adam is an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments [34]. It is highly memory-efficient and has shown good results in retraining our chatbot. We have also tried SGD [35], yet Adam has shown the better performance and hence, we have released the Adam-based retrained model.

The original GPT-2 was written in tensorflow [36] and this is the version we used. That said, for retraining purposes, we applied the TPU-trainable version of the GPT-2 [37].

As for the hardware, Google Colaboratory provided us with cloud TPUs and training capabilities. It came 25 GB RAM and since we connected the Colab to Google Drive [33], we had enough storage to do transfer learning.

The link for downloading the model is available on our GitHub page [38].

4.2 Filtering Based on Regex/String Manipulation

The GPT-2 responses are usually very lengthy and for the most part, the answer is not relevant to the question. To prune the responses generated from GPT-2, we first chunked the answer into the list of sentences using Python’s built-in module for dealing with regular expressions (`re` [39]) and then for each answer in the list of answers, performed the following regex/string operations:

(1) Eliminated redundant spaces
(2) Eliminated extra punctuation marks (specifically, “.”, “!”, and “?”)
(3) Removed redundant parentheses and square brackets
(4) Further split the sentence into separate sentences if it contained a period (“.”)

Steps 2 and 4, once again, employed `re` module while for steps 1 and 4, just the built-in string operations were sufficient (hence, no built-in or external module was used).

These operations have significantly improved the quality of the answer and allowed us directly passing them to the pretrained models for generating embeddings.

4.3 Filtering Based on Semantic Similarity to the Question

Semantic similarity is a metric that quantifies the degree to which two texts or text documents are similar to each other. The two approaches we have used include cosine similarity and inner product. The difference between the two is that cosine similarity pays attention to only the angle between the vectors, while the inner product cares about both the angle and the magnitude. That said, if one has a normalized data, both approaches are nearly equivalent.

To put each sentence in a vector representation, we tested and applied four different approaches for generating embeddings:

- tf-idf [40]: a simple, tf-idf based embedding-generation method.
- BERT: Large model, uncased (24 layers and 340M parameters) [41].
- BioBERT: Large v1.1 (+PubMed 1M) model based on BERT-large Cased (custom 30k vocabulary) [42].
- USE: Version 3, large [43].

In all cases, the similar strategy was applied for filtering sentences. The following equation defines the embedding generation process:

$$E = emb(S) \tag{1}$$

where $S = (s_1, s_2, \dots s_n, q)$ and denotes the list of sentences obtained by performing the split based on regex/string manipulation plus the question, *emb* denotes one of the embedding generation approaches (i.e., tf-idf, BERT, BioBERT, or USE), and q denotes the question being asked. E denotes the embeddings generated for the list of sentences. In other words, the list of sentences is passed to the embedding generation function and the feature matrix is returned.

Once the embeddings were generated, we have applied cosine similarity and inner product for ranking the sentences. Inner product was used only with USE since its embeddings were approximately normalized. The following equation shows the ranking process:

$$M = (sim(e_1, e_q), sim(e_2, e_q) \dots sim(e_n, e_q)) \tag{2}$$

where $e_1, e_2, \dots e_n, e_q \in E$ and *sim* is a function that finds the cosine similarity or the inner product between two embeddings. For every sentence embedding $e_1, e_2, \dots e_n$, the approach calculates the cosine similarity to the question embedding e_q. M is a similarity matrix where every column denotes the similarity vector for a particular sentence. The similarity vector contains the similarity values of the particular sentence with respect to all the sentences (including itself). The last column/similarity vector in the matrix has the similarity measures with respect to the question embedding.

Finally, we ranked the sentences based on their similarity scores, selected top five most similar sentences using Python's numpy [44] package, and concatenated them into a single, unified answer.

4.4 Example Responses

We present example responses in the form of a human-bot conversation.

The dialogue shows a consistent pattern in generating responses. Response to the first question regarding the vaccines and therapeutics states that "several drugs have been developed against COVID-19." Furthermore, it says that "drugs that inhibit pro-inflammatory responses (e.g., IL-6) have the potential to inhibit COVID-19 infection and viral replication in nonhuman primates." Regarding the response to the second question, concerning the risk factors, it addressed the question directly by stating that "a large number of travel is seen as one of the major causes of COVID-19 spread." As

HUMAN: What do we know about vaccines and therapeutics of COVID-19?

BERT BOT: Because of their critical role in viral pathogenesis and host defense against infections, several drugs have been developed against COVID-19 With this in mind, we present evidence that drugs that inhibit pro-inflammatory responses (e.g., IL-6) have the potential to inhibit COVID-19 infection and viral replication in nonhuman primates after administration up to 12 days following infection Because of their critical role in viral pathogenesis and host defense against infections, several drugs have been developed against COVID-19

USE BOT: The introduction of such a component could increase vaccine efficacy against these infections The idea of a vaccine against viral diseases has long been debated, and vaccination has been largely controlled by the development of vaccines that are effective at the time of application The introduction of such a component could increase vaccine efficacy against these infections

HUMAN: What do we know about COVID-19 risk factors?

BERT BOT: However, COVID-19 is now a threat to global health security According to the latest World Health Organization report, a large number of travel is seen as one of the major causes of COVID-19 spread However, COVID-19 is now a threat to global health security

TF-IDF BOT: We also know that vaccination is very effective at limiting the risk of PEDV infection, especially in young children We also know that vaccination is very effective at limiting the risk of PEDV infection, especially in young children In this sense, it is of the utmost importance to know the temperature variation with temperature

HUMAN: What is known about transmission, incubation, and environmental stability of COVID-19?

BIOBERT BOT: For example, a number of reports have suggested that animal and human hosts may have provided the necessary environment for the transmission of COVID-19 in people and poultry In fact, all known animal hosts that have been associated with animal-to-human transmission of COVID-19 are wild birds, including chickens, geese, mule deer, herons, and turkey In fact, all known animal hosts that have been associated with animal-to-human transmission of COVID-19 are wild birds, including chickens, geese, mule deer, herons, and turkey

TF-IDF BOT: The case of COVID-19 has demonstrated the ability to be managed rapidly and effectively during a crisis or emergency The case of COVID-19 has demonstrated the ability to be managed rapidly and effectively during a crisis or emergency The case of COVID-19 has illustrated an important, but under-appreciated, capability and the opportunity for collaboration between global health, the global private sector, and other global institutions

Figure 2: Example Responses in the Form of Human-Bot Conversation.

for the third question, about transmission, incubation, and environmental stability of COVID-19, it has mentioned that "large number of travel is seen as one of the major causes of COVID-19 spread" and additionally, talks about "animal-to-human transmission." In all cases, sentences were highly readable and understandable. That said, in some cases, the same sentences were repeated due to how the hybrid approach was implemented. This can be avoided, which we discuss in the section 7.

Table 1: Testing questions from CORD-19.

Number	Question
#1	Are there geographic variations in the mortality rate of COVID-19?
#2	What is known about transmission, incubation, and environmental stability of COVID-19?
#3	Is there any evidence to suggest geographic based virus mutations of COVID-19?
#4	Are there geographic variations in the rate of COVID-19 spread?
#5	What do we know about virus genetics, origin, and evolution of COVID-19?
#6	What has been published about ethical and social science considerations of COVID-19?
#7	What has been published about medical care of COVID-19?
#8	What do we know about diagnostics and surveillance of COVID-19?
#9	What do we know about COVID-19 risk factors?
#10	What has been published about information sharing and inter-sectoral collaboration of COVID-19?
#11	What do we know about vaccines and therapeutics of COVID-19?
#12	What do we know about non-pharmaceutical interventions of COVID-19?

5 QUESTIONS AND EVALUATION

In order to evaluate the performance of the proposed approaches as well as the overall performance of the chatbot, it is crucial to have a question dataset that both are frequently asked and related to COVID-19. For this purpose, we decided to use 12 questions from the Kaggle's COVID-19 Open Research Dataset Challenge (CORD-19) [45]. Most of the questions included the term "COVID-19" but others did not, in which case we appended the term to the end of the question. Table 1 presents all 12 questions.

For each of the 12 questions, we generated five different answers by applying the four different embedding generation techniques, resulting in a total of 240 answers. Therefore, the response for every question was generated exactly 5 times using the same technique. This ensured a fair and consistent distribution of both the questions and the approaches across the dataset. We made all of the answers publicly available on GitHub [46]. We then asked two experienced medical experts to evaluate the quality of these responses by assigning different relevance scores according to the categories in Table 2. Having 5 categories allowed for a flexibility and diversity of opinions/judgements as well as a broad range of scores that ultimately gave us a better way to evaluate our approaches. The evaluation was done primarily by averaging the scores for a particular approach.

Table 2: Five Rating Categories.

Category	Description	Point(s)
Relevant	The answer partially or fully answers the question and/or makes clear attempts to do so and is related to the question	5
Well-formed	the answer makes a logical sense and is somewhat related to both the question and COVID-19, yet it does not (partially or fully) answer the question	4
Informative	The answer is not related to the question, but provides some information about COVID-19 and makes a logical sense	3
Acceptable	The answer makes some logical sense and is weakly related to the question or COVID-19, but is mostly difficult to understand	2
Poor	the answer is totally unrelated to the question or COVID-19 and/or does not make a logical sense	1

Our annotation process had two phases. In the first phase, we let the annotators evaluate the test subset of the responses generated by the language model. The test subset was comprised of 20 questions. We then computed the IAA (Inner Annotator Agreement) which was approximately equal to 0.389. Due to having 5 categories, we used Pearson correlation coefficient for computing the IAA (as opposed to Cohen's Kappa, etc). Low correlation value led us to having a meeting with both annotators where we discussed why the they had different scores on particular responses to questions. Finally, both annotators reached an agreement and gave the same scores for every question in the test subset of 20. Once the agreement was reached, we then let the annotators evaluate the remaining 220 questions. Note that we evaluated our model based on the 240 responses and included initial subset, where both annotators agreed on the judgement. This was done for the sake of fairness and consistency.

Table 3: Average Scores of Embedding Generation Approaches Across the Annotators.

Approach	A1	A2	Overall
tf-idf	3.967	3.8	3.884
BERT	4.167	4.283	**4.225**
BioBERT	4.133	4.067	4.100
USE	3.683	4.083	3.883

6 EMPIRICAL RESULTS

6.1 Performance by Approach.

Table 3 lists the evaluation results of different approaches. It shows the approach, the average scores based on the approach for each annotator, and the overall average across the annotators. The first annotator rated BERT as the best approach with the average score of 4.167. BioBERT shows slightly worse performance with a score of 4.133 than BERT. The tf-idf approach performs well with a score of 3.967, yet it could not outperform either BERT or BioBERT. USE has the worst performance out of all embedding generation techniques with the score of 3.683 out of 5. The second annotator, similarly, gave the highest average score to BERT (4.283). USE was the second best with the score of 4.083 followed by BioBERT with approximately the same score of 4.067. The tf-idf approach has yielded the worst results, rated 3.8.

In general, the results are consistent between two annotators with an inner annotator agreement score of 0.521, which was calculated using the Pearson correlation. Models from the BERT family showed the best performance in automatically answering COVID-19 questions, with BERT slightly outperforming BioBERT (4.225 vs. 4.100 - average scores) being the best. The tf-idf approach and USE show roughly similar performance (3.884 vs. 3883)., yet inferior to BERT and BioBERT. All four approaches, on average, can be considered to be in the "well-formed" category with BERT and BioBERT being close to the "Relevant" category. The overall average was 4.023 (Well-formed).

6.2 Performance by Question

Table 4 shows the average scores for annotators A1 and A2, the overall average, and the difference based on the question asked to the language model. From the table, it is clear that the proposed approach had the best results for responses to questions # 11, # 3, and # 9 with average relevance scores of 4.700, 4.700, and 4.550, respectively After conducting the one-tailed one sample t-test, taking the differences as the values, we conclude that the differences for the scores (0.200, 0.300, and 0.200 respectively) are not significant with p-values being approximately 0.682, 0.221, and 0.682 respectively (all being above our threshold value of 0.05). This means that both annotators made consistent judgement.

It should be noted that all 3 of these questions seem to be rather short in length. The responses to question # 7, on the other hand, have the worst average score and interestingly, both annotators gave the same score of 2.667. That said, the question is also one of the shortest questions in length. Therefore, the length does not seem to always have a correlation with the score (hence, the performance).

Table 4: Question-Based Average Scores of Embedding Generation Approaches Across the Annotators.

Question	A1	A2	Average	Difference
# 1	4.200	3.450	3.825	0.750 (A1)
# 2	4.350	4.100	4.225	0.250 (A1)
# 3	4.550	4.850	**4.700**	0.300 (A2)
# 4	4.100	4.150	4.125	0.050 (A2)
# 5	3.600	3.950	3.775	0.350 (A2)
# 6	4.100	4.250	4.175	0.150 (A2)
# 7	2.650	2.650	2.650	0.000 (NA)
# 8	3.850	4.450	4.150	0.000 (A2)
# 9	4.650	4.450	4.550	0.200 (A1)
# 10	3.412	3.706	3.559	0.294 (A2)
# 11	4.600	4.800	**4.700**	0.200 (A2)
# 12	3.700	3.850	3.775	0.150 (A2)

To further analyze why responses to question # 7 had the lowest average score, we determined whether the terms of the question are present in the dataset. Terms "inter-sectoral collaboration" and "information sharing" were both present in CORD-19. Therefore, it is likely that the issue stems from the model itself and not the dataset.

According to the scores, we also found that the terms in the question had some correlation with the score. For example, the questions that featured words strongly linked to COVID-19, such as virus in question # 3, vaccine in question # 11, and risk in question # 9, had higher average response scores than those that did not (e.g., question # 10).

7 DISCUSSION

The project had several limitations. First, due to hardware constraints and the difficulty of fine-tuning, we have not used the larger 1.5B GPT-2 model that could potentially yield better results in generating responses. Second, the question pool was also limited and comprised of 12 questions. Additionally, we have tried only four specific embedding generation approaches, which might not be a fair representation of all such techniques in the domains of AI and NLP.

In order to make the language model more accessible to the general audience for automating the response generation, we built a web-based chatbot using the trained GPT-2 with options of tf-idf, BERT, BioBERT, and USE approaches. Please find the released code on our GitHub: . The application is powered by Python's Flask [47] package and gives a simple and user-friendly interface for the interactive communication with the chatbot. Please note that the health information generated by the chatbot is for general research purposes only. It is not a diagnostic tool, nor is it a substitute for medical advice or treatment for specific conditions.

Although our work has demonstrated the feasibility of using language models for automatically answering COVID-19 questions, much can be done in further research. First, we would like to explore why certain questions had higher scores than others. Secondly, other approaches for generating embeddings for sentences, such as BioWordVec [48], could potentially improve the performance of the chatbot and can be another avenue for exploration. From the human-bot dialogue (Figure 2), it is clear that GPT-2 could generate duplicate sentences which might be irrelevant to the question. In that case, the same sentence might be repeated in the final answer. One could incorporate an additional simple step of eliminating duplicate sentences which could potentially improve the quality of the answers. Adding an additional, third layer of filtering can also be tested to see whether it improves the quality of the responses. One could also change the evaluation process and let experts give k answers manually and then use the metric (ROUGE [49], BELU [50], etc.) to evaluate the results quantitatively. The evaluation could further be improved by considering both experts and non-experts. Moreover, since the chatbots are generally used for conversations and the evaluation is done as question answering task, one could adjust the chatbot for conversations by techniques such as incorporating the history of the conversation. Additionally, the GPT-2 model can always be further retrained on a new corpus, which could potentially improve the result. The 1.5B GPT-2 model could also be applied for retraining purposes. Finally, given that a much larger GPT-3 [51] language model was recently released, we believe that it is feasible for the chatbot with the capable hardware to explore the realm of possibilities with this model which could also evolve into an interesting future work.

8 CONCLUSION

In this paper, we applied the GPT-2 language model to automatically answer questions related to COVID-19, and quantitatively evaluate the proposed approach. To refine the responses generated by GPT-2, we compared four different embedding generation techniques, namely tf-idf, BERT, BioBERT, and USE. We utilized the collected corpus from the CORD-19 task to pretrain the GPT-2 model, and evaluated the automatically generated answers on twelve questions from the CORD-19. The results were evaluated by two medical experts. In general, the results were consistent between two annotators. The empirical results show that BERT achieved the best performance in automatically answering COVID-19 questions. We also built a web-based chatbot using the trained GPT-2 model and opensoured the code.

ACKNOWLEDGMENTS

This work was supported by NIH grant R01LM11934, the Mayo Clinic Center for Health Equity and Community Engagement Research Award, and the Mayo Clinic Office of Patient Education. The funders had no role in the design of the study, or collection, analysis, and interpretation of data and in preparation of the manuscript. The views presented in this report are not necessarily representative of the funder's views and belong solely to the authors.

REFERENCES

[1] Ghinai Isaac, McPherson Tristan, Hunter Jennifer, and Kirking Hannah et al. [n. d.] Retrieved 04/04/2020 from https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30607-3/fulltext.

[2] World Health Organization (WHO). 2020. Coronavirus disease 2019 (covid-19). Retrieved 07/26/2020 from https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200726-covid-19-sitrep-188.pdf?sfvrsn=f177c3fa_2.

[3] Rajkumar RP. 2020. Covid-19 and mental health: a review of the existing literature. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7151415/.

[4] Xiao Chunfeng. 2020. A novel approach of consultation on 2019 novel coronavirus (covid-19)-related psychological and mental problems: structured letter therapy. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7047000/.

[5] EA Holmes, RC O'Connor, and VH Perry et al. 2020. Multidisciplinary research priorities for the covid-19 pandemic: a call for action for mental health science. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7159850/.

[6] Gautam Kishore Shahi, Anne Dirkson, and Tim A. Majchrzak. 2020. An exploratory study of covid-19 misinformation on twitter. eprint: arXiv:2005.05710.

[7] Salton G., Wong. A, and Yang C. S. [n. d.] A vector space model for automatic indexing. https://dl.acm.org/doi/pdf/10.1145/361219.361220.

[8] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: on the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2.*

[9] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, (September 2019). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz682. https://doi.org/10.1093/bioinformatics/btz682.

[10] Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. eprint: arXiv:1803.11175.

[11] Hochreiter Sepp and Schmidhuber Jürgen. 1997. Long short-term memory. https://www.bioinf.jku.at/publications/older/2604.pdf.

[12] Giovanni Gennaro, Amedeo Buonanno, Antonio Girolamo, Armando Ospedale, and Francesco Palmieri. 2020. Intent classification in question-answering using lstm architectures, (January 2020). https://arxiv.org/pdf/2001.09330.pdf.

[13] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. https://arxiv.org/pdf/1409.3215.pdf.

[14] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: generalized autoregressive pretraining for language understanding. https://arxiv.org/pdf/1906.08237.pdf.

[15] Hung Le, Truyen Tran, and Svetha Venkatesh. 2020. Self-attentive associative memory, (February 2020).

[16] Jieh-Sheng Lee and Jieh Hsiang. 2019. Patent claim generation by fine-tuning openai gpt-2. eprint: arXiv:1907.02052.

[17] Tassilo Klein and Moin Nabi. 2019. Learning to answer by learning to ask: getting the best of gpt-2 and bert worlds. eprint: arXiv:1911.02365.

[18] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: large-scale generative pre-training for conversational response generation. arXiv: 1911.00536 `[cs.CL]`.

[19] Helena Lee, Ke Shu, Palakorn Achananuparp, Philips Kokoh Prasetyo, Yue Liu, Ee-Peng Lim, and Lav Varshney. 2020. Recipegpt: generative pre-training based cooking recipe generation and evaluation system, (March 2020).

[20] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. eprint: arXiv: 1802.05365.

[21] Shreyas Sharma and Ron Daniel Jr. 2019. Bioflair: pretrained pooled contextualized embeddings for biomedical sequence labeling tasks. eprint: arXiv:1908.05760.

[22] Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, 1638–1649.

[23] Iulian V. Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, Sai Rajeswar, Alexandre de Brebisson, Jose M. R. Sotelo, Dendi Suhubdy, Vincent Michalski, Alexandre Nguyen, Joelle Pineau, and Yoshua Bengio. 2018. A deep reinforcement learning chatbot (short version). eprint: arXiv:1801.06700.

[24] Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. eprint: arXiv: 2001.09977.

[25] Junjie Yin, Zixun Chen, Kelai Zhou, and Chongyuan Yu. 2019. A deep learning based chatbot for campus psychological therapy. eprint: arXiv:1910.06707.

[26] Vitalii Zhelezniak, Aleksandar Savkov, April Shen, and Nils Y. Hammerla. 2019. Correlation coefficients and semantic textual similarity. eprint: arXiv:1905.07790.

[27] 2020. Covid-19 open research dataset (cord-19). https://pages.semanticscholar.org/coronavirus-research.

[28] The White House Office of Science and Technology Policy. 2020. Call to action to the tech community on new machine readable covid-19 dataset. https://www.whitehouse.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need, 5998–6008. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors. http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

[30] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

[31] Alec Radford and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

[32] [n. d.] Gpt-2 model card. https://github.com/oniani/covid-19-chatbot.

[33] Google. [n. d.] Google colaboratory. https://colab.research.google.com/.

[34] Diederik P. Kingma and Jimmy Ba. 2014. Adam: a method for stochastic optimization. eprint: arXiv:1412.6980.

[35] Kiefer J. Kingma and Wolfowitz J. 1952. Stochastic estimation of the maximum of a regression function. https://projecteuclid.org/euclid.aoms/1177729392.

[36] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. (2015). http://tensorflow.org/.

[37] Shawn Presser. [n. d.] https://github.com/shawwn/gpt-2/.

[38] Oniani David and Wang Yanshan. [n. d.] Github: covid-19 chatbot. https://github.com/oniani/covid-19-chatbot.

[39] [n. d.] Re — regular expression operations. https://docs.python.org/3/library/re.html.

[40] [n. d.] Scikit-learn: tfidfvectorizer. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html#sklearn.feature_extraction.text.TfidfVectorizer.

[41] GitHub: Google Research. [n. d.] https://github.com/google-research/bert.

[42] [n. d.] Github: dmis laboratory - korea university. https://github.com/dmis-lab/biobert.

[43] [n. d.] Tensorflow hub: universal sentence encoder (version 3, large). https://tfhub.dev/google/universal-sentence-encoder-large/3.

[44] [n. d.] Numpy. https://numpy.org/.

[45] Kaggle (An AI challenge with AI2, CZI, MSR, Georgetown, and NIH & The White House). 2020. Covid-19 open research dataset challenge (cord-19). https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks.

[46] 2020. Annotation results and generated answers on github. https://github.com/oniani/covid-19-chatbot/tree/master/annotations.

[47] [n. d.] Welcome to flask – flask documentation. https://flask.palletsprojects.com.

[48] Zhang Yijia, Chen Qingyu, Yang Zhihao, Lin Hongfei, and Lu Zhiyong. 2019. Biowordvec, improving biomedical word embeddings with subword information and mesh, (May 2019). https://www.nature.com/articles/s41597-019-0055-0.

[49] Lin Chin-Yew. 2004. Rouge: a package for automatic evaluation of summaries. https://www.aclweb.org/anthology/W04-1013.pdf.

[50] Papineni Kishore, Roukos Salim, Ward Todd, and Zhu Wei-Jing. 2002. Bleu: a method for automatic evaluation of machine translation. https://www.aclweb.org/anthology/P02-1040.pdf.

[51] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. arXiv: 2005.14165 `[cs.CL]`.