IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 7, July 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

9940 572 462 | 6381 907 438 | ijircce@gmail.com | www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007094 |

Corona Virus Healthcare ChatBot

Rajat Agarwal , Jaya Gupta, Komaljyot Kaur, Dr. Ankush Mittal

Department of CSE, IIIT, Delhi, India

Department of IT, Coforge Ltd, India

Department of CSE, Graphic Era University Dehradun, India

Department of CSE, Raman Classes Roorkee, India

ABSTRACT: We aim to build a question answering system for the COVID-19 domain. Question Answering model provides the best answer for each selected paper. In this work, we specifically study FAQ retrieval for COVID-19, a contagious and fatal pandemic still evolving daily. As the spreading of the pandemic urged entire nations into a state of emergency, the medical and scientific communities found themselves under enormous pressure for the requirement of solutions to a problem still largely unknown to the public. Here the user will post a question related to COVID 19, and the system will respond with the best answer from the corpus. The answer will be text generated from the corpus. If the answer is below a certain threshold, then the system will not respond to the answer to such a question because we do not want
to give false information to the question for such a critical and crucial domain.

KEYWORDS: Question Answering, Chatbot, FAQ retrieval

I. INTRODUCTION

The outbreak of the COVID-19 pandemic has greatly affected humanity's lives, causing tremendous suffering and deaths at the international level. There is a high demand for reliable and up-to-date information on COVID-
19. On social media and office official public websites, there is a lot of information and misinformation about COVID-19, e.g., ingestion of aspirin or ibuprofen. The misinformation can affect people's lives, which motivates the need to answer questions like "Should I ingest disinfectants to treat COVID 19?" and "Can I use Aspirin with COVID?" Building a QA system designed to answer COVID-19-related questions automatically would greatly aid in effectively combating the current pandemic. As we have a large number of documents, for a given query, the search space is enormous, which leads to increased search time, and thus such a system can not be used in real time scenarios; therefore, we need to reduce the search space efficiently without compromising the quality of results. Creating a QA model specific to COVID-19 poses several challenges despite such progress. The first challenge is that there are no QA datasets oriented specifically to COVID 19. The second challenge is incorporating conventional biomedical text mining tools into existing QA models. In order to deal with this, we shift our focus to a critical feature used in biomedical text mining: biomedical named entities. The most prominent dataset used by researchers in this domain is CORD-19 by Wang et al. [1]. The Covid-19 Open Research Dataset (CORD-19) is a growing resource of scientific papers on Covid-19 and related coronavirus research. This dataset consists of a collection of research articles arranged by The White House and the Allen Institute of AI. It includes more than 80K articles with full text, containing the most recent studies conducted on the COVID 19 disease. CORD-19 aims to connect the machine learning community with biomedical domain experts and policymakers to identify effective treatments and management policies for Covid-19 so that the common person can benefit. The goal is to harness these diverse and complementary pools of expertise to discover relevant information from the literature quickly. As the data does not have question-

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007094 |

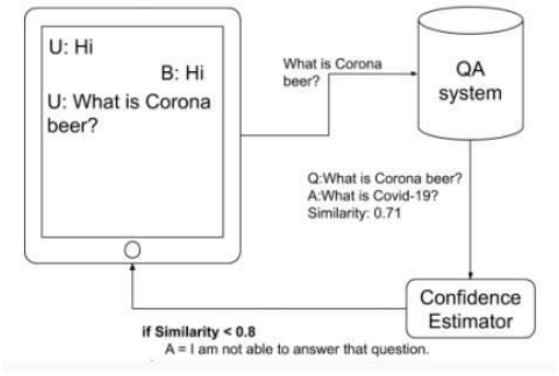

Fig 1. Question Answering System for the Covid-19 domain

answer pairs, we used which has questions and their corresponding answers. We compare the actual answer and predicted answer to get the score. For developing a question answering system for the COVID 19 domain. The user will post a question related to COVID-19, and the system will respond with the best answer from the corpus. The answer will be text generated from the corpus. If the answer is below a certain threshold, then the system will not respond to the answer to such a question because we do not want to give false information to the question for such a critical and crucial domain.

II. RELATED WORK

In the recent literature review, we have focused on understanding the QA system in any domain. For this purpose, we studied the QA system built on Amazon reviews by Gupta et al. [2]. The author has discussed segregating questions into relevant and irrelevant questions and working on relevant questions here. We findthis idea exciting and will try to incorporate it into our final model. Bao et al. [4] have prepared a medical chatbot system-based hybrid of knowledge graph and text similarity. They use Hierarchical bi-directional LSTMto find similar questions in the corpus. They claim that their approach outperforms SOTA [5] models like BERT[6].

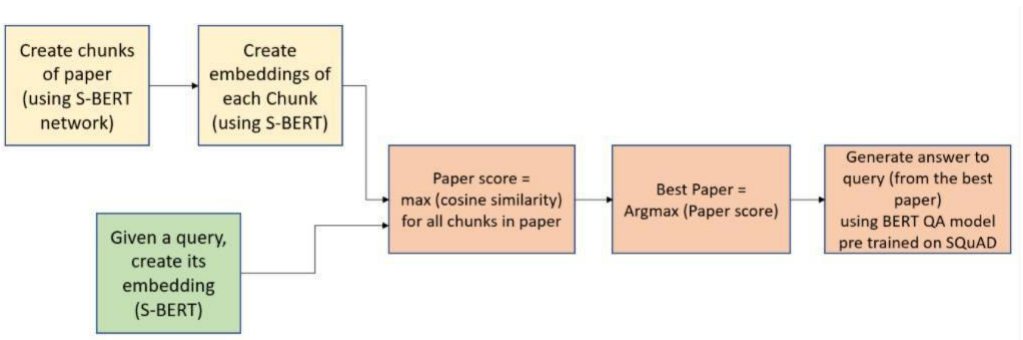

Fig 2. Baseline 1: An analysis of bert faq retrieval models for covid-19 infobot Sun et al [13]

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007094 |

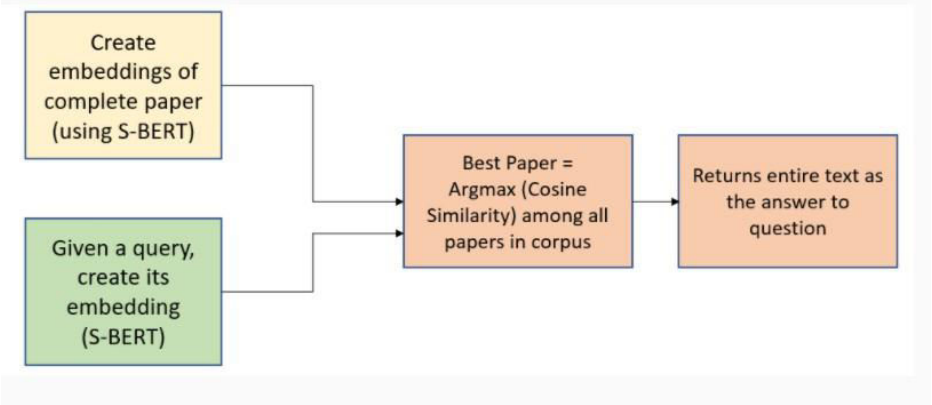

Fig 3. Baseline 2:RECORD, Muffo et al.[14]

One of the important features that many QA models do not Fig. 2. Baseline 1: An analysis of bert faq retrieval models for covid-19 infobot Sun et al [13]. Fig. 3. Baseline 2: RECORD, Muffo et al. [14] consider is the recency of information. Recency is an essential aspect as models that use up-to-date information would be able to provide more relevant results. Many QA models that use unstructured text are often given a document or a paragraph that contains an answer to each question. However, it is more realistic to retrieve a relevant document first and then find an answer rather than being provided the document. Furthermore, the dataset that we areusing in our work which is CORD-19 [1] is a collection of all research papers published in the domain of COVID-19. It is not in the form of QA pair; therefore, we need test data with questions and Answers to generatebaseline results. For this purpose, we are using the COUGH [3] dataset, which has several QA pairs. Sakata et al. [7] have generated query-question similarity and then searched for the relevant answer. They have worked ona different dataset where QA pairs are available. In Baseline-1 Sun et al [13], S-BERT is used to create embeddings of the complete paper. With a query given, its embeddings are created using S-BERT [12]. The best paper amongst all the papers in the corpus is selected using cosine similarity, and the entire paper is given as an answer to the question raised by a user. Baseline 2 Muffo et al. [14], is a two-step pipeline to extract answers from the CORD-19 dataset. The data is filtered according to the publishing year and a list of keywords identifying the COVID-19 virus. The first step is to extract a set of most semantically-correlated papers, where it is more likely to find an answer. All the chunks of the body texts in the dataset are embedded using the Sentence-BERT model. Once the subset of most relevant papers is selected, the second step is to extract an answer for each paper.

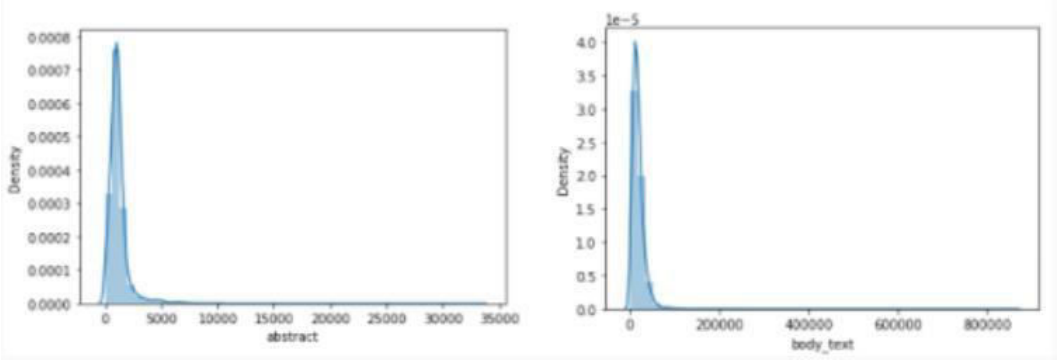

Fig 4. Data Preprocessing

- **Methodology**

We present a model that is Top2vec [8], which is an algorithm for topic model ling and semantic search. This algorithm automatically detects topics present in the text and generates jointly embedded topic, document, and word vectors. The Top2vec model does not require stop-word lists, stemming or lemmatization. The resulting topic vectors that the model generates are jointly embedded with the document and word vectors with distance between them representing semantic similarity. Our model demonstrates that it finds topics which are significantly more informative and representative of the corpus trained on than probabilistic generative models. The topic modelling used the following techniques - Latent

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007094 |

dirich let allocation, Probabilistic Latent Semantic Analysis, Non negative matrix factorization and Latent semantic allocation. The top2vec algorithm follows the specific procedure that is:-

1) Data Preprocessing:- Data with NULL Abstract or Body text is not considered for training the model. So basic text pre-processing steps are applied to clean the data like converting whole text to Lower case, Stopwords removal, Links removal etc. Average Abstract and Body length is 2.5k, and 10K characters, respectively. Fig.4 depicts the average abstract length and body length.

2) Create Semantic Embedding:- We need Embedding to extract topics, jointly document and word vectors where the distance between document vectors and word vectors represents the semantic association. In this technique semantically similar documents should be placed close or together in the embedding space and semantically dissimilar documents should be placed further. By using jointly embedded document and word vectors, we calculate topic vectors. Jointly embedded documents and word vectors are mainly created using Doc2vec [9].

3) Create Low Dimensional Document Embedding:- We know that in high dimensional space, document vectors are very sparse, so dimension reduction allows for dense clusters of documents to be found more efficiently and Fig. 5. Cluster the documents and assign prominent topics to each cluster Fig. 6. Finding Topic words accurately in the reduced space. For creating low dimensional document embedding, UMAP [10] technique is used. UMAP is a manifold learning technique for dimension reduction.

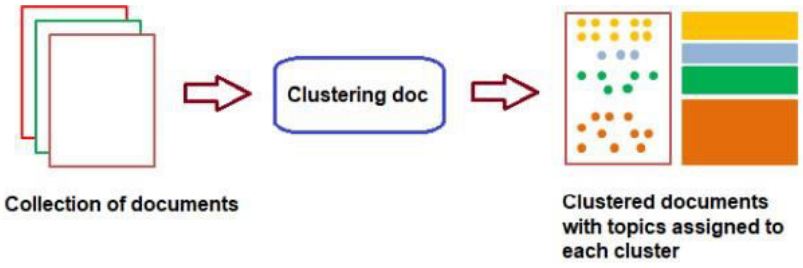

Fig. 5. Cluster the documents and assign prominent topics to each cluster

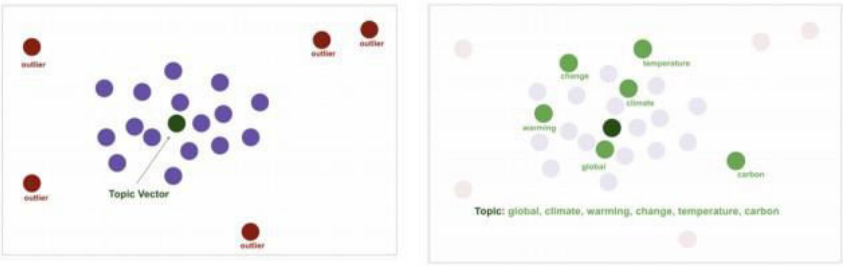

Fig. 6. Finding Topic words

4) Find Dense Clusters of Documents:- The purpose of finding dense clusters of documents is to find areas of highly similar documents in the semantic space. The HDBSCAN [11] clustering algorithm is used to find dense clusters for documents.

5) Calculate Topic Vectors:- There are several ways that the topic vector can be calculated from the document vectors. The simplest method is to calculate the centroid, i.e., the arithmetic mean of all the document vectors in the same dense cluster. The centroid is calculated for each set of document vectors that belong to a dense cluster, generating a topic vector for each set.

6) Find n-closest Word Vectors:- The closest word vectors become the topic words. In the semantic space, every point represents a topic best described semantically by its nearest word vectors.

Fig.7 shows the training stage of the model. By using a topic number, we will generate a word cloud in our methodology, and we are going to generate word clouds (Fig.8) for the top 5 most similar topics. The topics that are generated by our methodology are 78, and some of them are given below - Fig.9 shows the testing phase. In this stage, we give the question as input and extract the keywords from the question, Fig. 7. Training Fig. 8. Word cloud corresponding to cluster 1 and 2 out of total 78 clusters.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007094 |

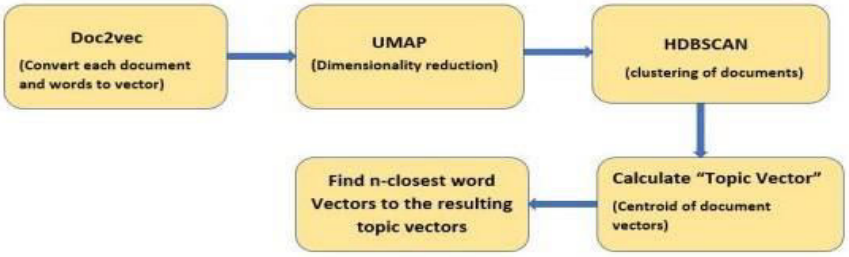

Fig. 7. Training

Fig. 8. Word cloud corresponding to cluster 1 and 2 out of total 78 clusters.

Fig. 9. Testing.

Fig. 9. Testing. then find clusters based on the related topics. After finding clusters, we find the nearest documents present in the clusters. In the end, the system returns the abstract of the best-searched document.

III. RESULTS AND EVALUATION

We have found all the metrics like BLEU score, precision, recall, and F1 score at baseline 1, baseline 2 and our proposed approach.

Performance Metrics:

1) BLEU SCORE: BLEU (bilingual evaluation understudy) is an algorithm for evaluating the quality of text machine-translated from one natural language to another.

2) Precision: Model precision score represents the model's ability to predict the positives out of its positive predictions correctly. The precision score is a valuable measure of the success of prediction when the classes are very imbalanced.

3) Recall: Recall is a Classification error metric. It evaluates the outcome of the classification algorithms for which the target/response value is a category.

4) F-Score: F-score is a machine learning model performance metric that gives equal weight to both the Precision and Recall for measuring its performance in terms of accuracy, making it an alternative to Accuracy metrics.

We can see that our results at the proposed approach are betterthan baseline 1 and baseline 2. The comparison of results is shown in the below table.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007094 |

Table 1 describes the performance of the different approaches.Comparison for the testing time:

	Baseline-1	Baseline-2	Proposed Approach
BLEU Score	0.0048	0.021	0.067
Precision	0.0038	0.060	0.617
Recall	0.3380	0.483	0.230
F1-Score	0.0075	0.012	0.257

Table 1:Evaluation Scores

	Baseline-1	Baseline-2	Proposed Approach
Time	3.214 sec	2.119 sec	1.645 sec

Table 2: Comparison for the testing time

Table 2 describes the performance of the different ap proaches in terms of time(sec).

We compare our proposed approach with baseline-1 and baseline-2 in terms of time(sec). We have found the test time for Baseline-1, Baseline-2, and our proposed approach, and on comparing the results, we got improvement concerning time i.e. our proposed model is faster than both the baselines.

Comparing Baseline-1 and Baseline-2 results, precision of Baseline-2 is higher as compared to Baseline-1 as in Baseline 1 the whole paper is returned as the answer to the question asked by the user. Moreover, theproposed model's precision is quite high compared to the Baseline-1 and the Baseline-2. Also, the F1-Score of the proposed model is high as compared to the previous two approaches. Search time which was a significant issue in the previous two approaches, has been significantly reduced with the help of the proposed approach.The evaluation metric reflects that users of the proposed method will need to read the minimal context of answers (i.e., a sentence) regardless of the correctness of the answer itself.

IV. CONCLUSION

Our model improves space and time and gives a real time performance, and the clustering algorithm automatically finds the best number of clusters, not parameter bounded. Furthermore, our results are Precision- oriented; false-positive is comparatively less and has a comparable or better score than the baseline models. We hope our system will be able to aid researchers in their search for knowledge and information [2] not only regarding COVID-19 but for future pandemics as well. Further, we plan to expand our work by trying BERT embeddings which store better semantic and contextual information.

REFERENCES

1. Q. Bao, L. Ni, and J. Liu. Hhh: an online medical chatbot system based on knowledge graph and hierarchical bi-directional attention. In Proceedings of the Australasian Computer Science Week Multiconference, pages 1– 10, 2020.
2. Gupta, M., Kulkarni, N., Chanda, R., Rayasam, A., Lipton, Z. C. (2019). Amazonqa: A review-based question answering task. arXiv preprintarXiv:1908.04364.
3. Li, Z., Wang, H., Zhang, X., Wu, T., Yang, X. (2020). Effects of space sizes on the dispersion of cough- generated droplets from a walking person. Physics of Fluids, 32(12), 121705.
4. Bao, Qiming, Lin Ni, and Jiamou Liu. "HHH: an online medical chatbot system based on knowledge graph and hierarchical bi-directional attention." Proceedings of the Australasian Computer Science Week Multiconference. 2020.
5. Abeywickrama, Dhaminda B., Nicola Bicocchi, and Franco Zambonelli. "SOTA: Towards a general model for self-adaptive systems." 2012 IEEE 21st International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises. IEEE, 2012.
6. de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., Nissim, M. (2019). Bertje: Adutch bert model. arXiv preprint arXiv:1912.09582.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007094 |

7. Sakata, Wataru, et al. ”FAQ retrieval using query-question similarity and BERT-based query-answer relevance.” Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019.
8. Angelov, D., 2020. Top2vec: Distributed representations of topics. arXiv preprint arXiv:2008.09470.
9. Lau, Jey Han, and Timothy Baldwin. ”An empirical evaluation of doc2vec with practical insights into document embedding generation.” arXiv preprint arXiv:1607.05368 (2016).
10. [10] S.Kannadhasan and R,Nagarajan, Performance Improvement of H-Shaped Antenna With Zener Diode for Textile Applications, The Journal of the Textile Institute, Taylor & Francis Group, DOI: 10.1080/00405000.2021.1944523
11. McInnes, L., Healy, J., Melville, J. (2020). UMAP: uniform manifold approximation and projection for dimension reduction.
12. McInnes, Leland, John Healy, and Steve Astels. ”hdbscan: Hierarchical density based clustering.” J. OpenSource Softw. 2, no. 11 (2017): 205.
13. Choi, Hyunjin, et al. ”Evaluation of bert and albert sentence embed ding performance on downstream nlptasks.” 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021.
14. Sun, S. and Sedoc, J., 2020. An analysis of bert faq retrieval models for covid-19 infobot.
15. Muffo, Matteo, Aldo Cocco, Mattia Messina, and Enrico Bertino. ”RECORD: a Question Answering tool for COVID-19.” (2020)

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details