COVID-19 Semantic Search Engine Using Sentence-Transformer Models

Anagha Jose(✉) and Sandhya Harikumar

Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India
anaghajose13@gmail.com, sandhyaharikumar@am.amrita.edu

Abstract. With the onset of COVID-19, enormous research papers are being published with unprecedented information. It is impractical for the stake holders in medical domain to keep in pace with the new knowledge being generated by reading the entire research papers and articles in order to keep pace with new information. In this work, a semantic search engine is proposed that utilises different sentence transformer models such as BERT, DistilBERT, RoBERTa, ALBERT and DistilRoBERTa for semantic retrieval of information based on the query provided by the user. These models begin by collecting COVID-19-related research papers and are used as an input to the pre-trained sentence transformer models. The collected research papers are then converted into embedded paragraphs, and the input query is sent to the same model, which in turn delivers the embedded query. The model uses cosine similarity to compare both embedded paragraphs and the embedded query. Consequently, it returns the top K most similar paragraphs, together with their paper ID, title, abstract, and abstract summary. The bidirectional nature of the sentence transformer models allows them to read text sequences from both directions, making the text sequence more meaningful. Using these models, COVID-19 semantic search engine has been developed and deployed for efficient query processing. The similarity score for each model was computed by averaging the top 100 query scores. As a result, the RoBERTa model is faster, generates a higher score of similarity, and consumes less runtime.

Keywords: COVID-19 · BERT · RoBERT · ALBERT · DISTILRoBERT · DistilBERT · Pandemic · NLP

1 Introduction

There has been accelerated growth of research papers released during the ongoing COVID-19 pandemic. In this paper, a semantic search engine is developed that optimises the contents of various research articles. To begin, the researchers collect these COVID19-related research publications and then feed them into the pre-trained BERT model along with the input query. This model transforms the research papers into embedded paragraphs and query into embedded query. The

I. Raman et al. (Eds.): ICC3 2021, CCIS 1631, pp. 189–200, 2022.
https://doi.org/10.1007/978-3-031-15556-7_14

conversion to the embedded is beneficial for semantic search and information retrieval because it allows for more precise extraction of answers that are appropriate to the question. Using cosine similarity, the model compares both the embedded paragraphs and the embedded query, then delivers the top K most similar paragraphs, together with their paper id, title, abstract, and abstract summary.

In this work, the machine-learning topic modelling methodology is used to analyse text data and create cluster terms for a collection of documents. The researchers also employed sentence transformers such as BERT, DistilBERT, Roberta, ALBERT, and DistilRoBERTa in this research to select the best model based on their score and the maximum time each model took to process. The model uses bidirectional and self-attention techniques which result in better accuracy. A sentence transformer-based semantic search engine is more efficient in text classification and for more accurate results. The sentence transformer models handle a large corpus of data. This semantic search engine is useful for health care providers and other COVID-19 workers who need to stay up to date with COVID-19 related information.

The BERT [2] is a transformer encoder stack which is pretrained. A self-attention and a feed-forward network are included in each encoder layer. The input from the encoder is passed through a self-attention layer, which then passes the output from the self-attention layer to a feed-forward neural network, which finally passes it on to the next encoder. The COVID-19 dataset [11] includes over 500,000 research papers, nearly 100,000 of which contain full text with regard to coronavirus. The BERT model takes data in a specific format as input. The model [2] employs a special token called [CLS], [SEP], [MASK] for the input formatting, Each sentence begins with a [CLS] token and ends with a [SEP] token. The framework consists of two steps: the Masked Language Model (MLM) and Next Sentence Prediction (NSP). MLM attempts to identify the actual value of a masked word, whereas the NSP determines whether sentence B is a continuation of sentence A.

In the proposed system, COVID-19 related research articles and query from users are collected and fed into sentence transformer models like BERT, DistilBERT, Roberta, ALBERT, and DistilRoBERTa. These models convert research papers and query into embedded paragraphs and embedded query. Cosine similarity has been used to compare embedded paragraphs and embedded query. Consequently, the model returns the top K similar paragraphs with their paper id, title, abstract, abstract summary. Further, various sentence transformer models are evaluated in order to determine the optimal semantic search engine. For experimentation, each model's similarity score is determined by averaging the top 100 query scores. It is observed that the RoBERTa model is faster, produces a higher similarity score, and requires less runtime as compared to other models.

1.1 Main Contributions

- Proposed a semantic search engine to retrieve most relevant articles from COVID-19 research papers based on sentence transformer models.
- Efficient query processing strategy chosen based on comparative analysis of five models namely BERT, DistilBERT, Roberta, ALBERT, and Distil-RoBERTa in the context of semantic information retrieval.

2 Related Works

The researchers collected the COVID-19 related dataset from the Kaggle website. This dataset [11] includes over 500,000 academic papers, across over 100,000 containing actual transcripts related to the corona virus. This dataset enables the world's artificial intelligence research community to utilise text and information extraction technologies to answer queries about the information contained inside and across it, with the goal of advancing ongoing COVID-19 response activities worldwide. This AI approach will help to create new insights to combat this infectious disease.

Manish Pate et al. [9] has built a semantic-search engine that can search for queries and rate content from most meaningful to least meaningful by utilising neural networks and BERT embeddings. In difficult queries for a given set of documents, the results demonstrate an enhancement over one existing search engine. The similarity score is calculated using a neural network while Kassim J. M. [3] proposed a semantic-based search engine that includes ontology creation, ontology crawler, application server, information retrieval, and query processor, but the authors do not use ontology to retain word structure and generate database information structures.

The Xiaoyu GuoJing et al. [6] created a model of semantic search, built on a recurrent neural network (RNN). In this model, each sentence is first broken down into individual words, and then these words are converted into word embeddings using GloVe vectors (global vectors for vector representation). It aids in the creation of vector representations of words. The GloVe extracts the semantic relationship between words from the co-occurrence matrix. After that, the embedded vectors are fed into the recurrent neural network (RNN). The RNN's output is then sent to the attention layer. Then it goes to the output layer, where it assists in determining the final output.

Priyanka C Nair et al. [13] addressed the survey of various jobs conducted on discharge summaries and the studied technologies. The discharge report contains detailed information about the patient, including his or her medical history, symptoms, investigations, therapy, and medicines. While the discharge summary is structured in general, it is not structured in a way that it can be processed by clinical systems. Several natural language processing (NLP) and machine learning algorithms were employed to extract various important bits of information from discharge summaries.

Remya R.K. Menon et al. [14] developed a predictive model-based strategy for constructing semantically oriented topic representations from a document

collection. To begin, generate two matrices from the updated topic model: a matrix of document-topics and a matrix of term-topics. The collection of documents and the reconstructed documents are 85% identical. This may be shown by examining the reduced document-term matrix that was created from the two matrices. In topic models, the concept of themes is inferred from the regular appearance of concepts. It might be argued that the terms may not be semantically connected on the issue, even though they appear to be.

Akhil dev et al. [15] proposed a model for document data structure maintenance that makes use of a range of deep learning techniques. The same argument may be made for nearly all techniques. They can all be described by reference to their vector similarity. The research under consideration serves to improve the accuracy of documents by assessing alternate methods for maintaining document structure. Using the Doc2Vec model and applying TF-IDF produced better results than using the Tfidf model.

Paluru Asritha et al. [16] demonstrated the effectiveness of Intelligent Text Mining. Cyberbullying is described as the act of harassing, defaming, abusing, or threatening another person through the use of electronic communication. Censorship on social media has been increasingly important in recent years, and Twitter, Facebook, and Instagram are all at risk of cyberbullying. This can be somewhat minimised by isolating such frightening messages or remarks. The technique of sentiment analysis is used to determine whether a text is positive, negative, or neutral. It contributes to the establishment of a sentence's emotional tone. This paper describes a hybrid classifier technique for categorising these frightening messages, which distinguishes between good and negative assessments. According to the experimental results, the classifier is 89.36% accurate on the considered dataset.

The purpose of the feature extraction and processing system [17] is to extract information from english news stories and present it rationally to the user. Crawl and store the details of a preset set of websites. In between the lines of their publications, news organisations conceal a lot of information. Extracting data and arranging it in such a way that inferences may be drawn is crucial in analytics. The programme extracts identifiable elements such as the location, person, or organisation referenced in the news, as well as the headline and key terms for each article.

The document semantic representation based on matrix decomposition [18] addresses the critical issue of semantic text categorization for data acquisition in a data implementation. Often, search queries on documents look for pertinent information. Standard feature extraction techniques do not convey relevance and instead concentrate on phrase similarity for query processing. The difficulties inherent in semantic information in documents are in identifying critical aspects. The majority of strategies for detecting significant characteristics change the original data into another space. This results in a sparse matrix that is computationally inefficient. This approach identifies significant documents and terms in order to optimise data aggregation. Experiments on five sets of data demonstrate the effectiveness of this strategy.

3 Proposed Method

The architecture of the COVID-19 semantic search engine is shown in Fig. 1. The number of articles and research papers on the fight against the corona virus has increased in recent months. Health professionals find it difficult to follow the new information by reading all of these articles and research papers about the corona virus. This model aids in the discovery of answers to the query. By collecting the corona virus dataset from kaggle [11], they are converted into embedded paragraphs and the input query is sent to the same model, which in turn delivers the embedded query using the pretrained model's Robustly Optimized BERT Pretraining Approach (RoBERTa) [7]. This model uses cosine similarity to compare both embedded paragraphs and the embedded query. As a result, it returns the top K most similar paragraphs, together with their paper ID, title, abstract, and abstract summary.

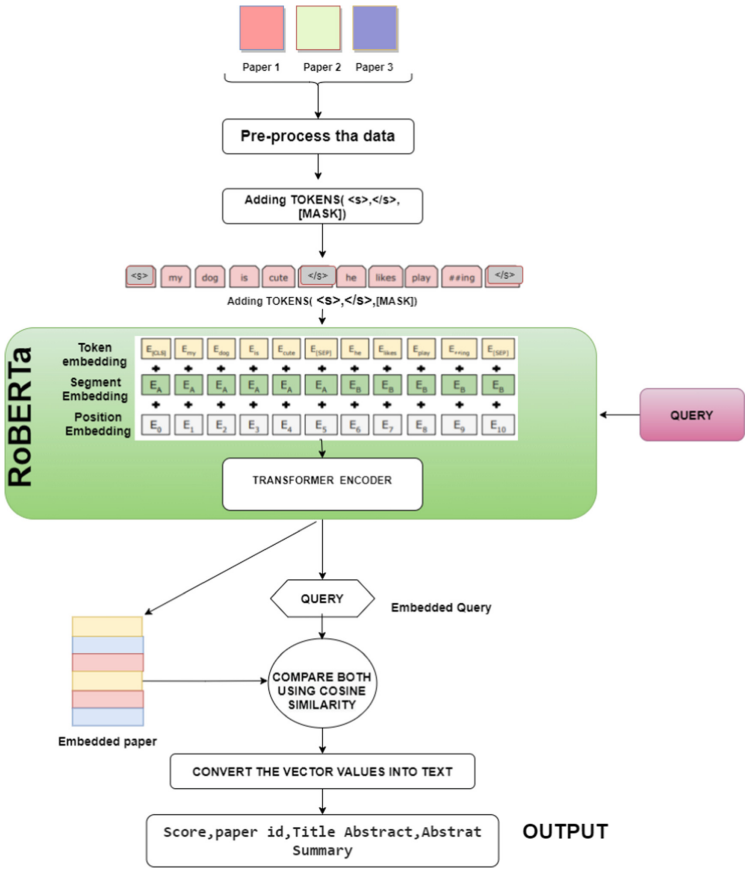

Fig. 1. Diagrammatic representation of the proposed approach

3.1 RoBERTa

RoBERTa [7] is an improved BERT Pretraining Approach. RoBERTa exceeds BERT by making the following changes:

- The training data used by RoBERTa is drawn from the large-scale text corpus (16G), news (76G), open-source recreation of the WebText corpus (38G), and articles (31G) databases.
- For the MLM aim, BERT masks training samples once, whereas RoBERTa replicates training sets 10 times and masks it differently.

Facebook has developed and published this model [7]. This pre-trained word representation language model can process long texts efficiently. It's bidirectional, so it can be read from both directions. Special tokens are used: $\langle s \rangle$, $\langle /s \rangle$, [MASK].

The Next Sentence Prediction (NSP) task, which is used in the BERT's pretrained model, is missing from the RoBERTa model. The NSP model is used to determine whether phrase B follows phrase A. The RoBERTa model employs a dynamic masking pattern, which duplicates the training data ten times and masks the word differently in each epoch, resulting in a change in the masked token during training. In the training procedure, larger numbers of training sizes have also been found more helpful.

The authors tested the removal/addition of NSP loss and found that removing NSP loss improves the model's task performance, possibly a bit. This RoBERTa model can process long texts, but it is very expensive, so we can only use a pretrained model. The comprehensive design and execution of each component is covered further below.

3.2 Explanation of the Block Diagram

Dataset: In this work, the researchers obtained the dataset from kaggle [11] related to the corona virus. The dataset is in the JSON format so the researchers convert this format into a CSV file and then divide the information into paragraphs. The RoBERTa model uses input data in a particular format. For input formatting, this model (Yinhan Liu, 2019) uses a special token called $\langle s \rangle$, $\langle /s \rangle$, [MASK]. The first sentence begins with a $\langle s \rangle$ token, and each subsequent sentence ends with a $\langle /s \rangle$token. The words are randomly selected from each sequence and mask those randomly selected words by using a [MASK] token.

Embeddings: The Highly dimensional vectors can be translated into a relatively low-dimensional space using embedding. The RoBERTa pretrained model (A Robustly Optimized BERT Pretraining Approach) [7] convert the paragraphs and user query into embeddings. In this research, the model uses token embeddings, embeddings, and position embeddings for the conversion of embeddings. The model adds the tokens in token embeddings and the number of the sentences that are encoded into a vector in the segment embeddings, and the position of a word within that sentence that is encoded into a vector in position embeddings. These values are concatenated by the model.

RoBERTa Transformer: The RoBERTa [7] transformer is a collection of encoders. Each encoder incorporates a self-attention neural network and a feed forward neural network(FNN). The concept of attention has aided in the performance of neural machine translation programmes. In order to help the encoder to look at the other words in the input sentence, a self-attention layer is used. A FNN is a network of neurons with multiple layers in which all information travels exclusively in the forward direction. It has three layers: input, hidden, and output. The information enters the input nodes first, then passes via the hidden layers, and ultimately exits through the output nodes. There are no links in the network to feed the information coming out of the output node back into the network.

Self Attention: The encoding of a single word in the self-attention layer enables the encoder to inspect other phrases within the input text. There are three vectors: the vector query, the key vector and the vector value.

The Following is the Processing Step Within Self-attention

- The vectors are produced by multiplying weight matrices.
- Calculate the score by multiplying the query vector by the key vector.
- Distribute the score according to the square root of the dimension of the key vector.
- For the best results, use softmax and then score is standardised by Softmax.
- Multiply the softmax score by each value vector.
- Add the weighted value vectors at this position to obtain self-attention output.
- The output of self-attention is subsequently transferred into the next neural feed network.

4 Experiments and Results

The model examines both the embeddings of the research paper and the query using cosine similarity, then returns the top K closest similar paragraphs, along with their paper ID, paper title, abstract, and abstract summary, as shown in Fig. 2. This work focused on experimenting with sentence transformer models such as BERT, DistilBERT, RoBERTa, ALBERT, and DistilRoBERTa, as well as SVD topic modelling using the Latent Semantic Indexing Model (LSI). When compared to other models, it is determined that the RoBERTa model produces the best results. The similarity score for each model was calculated by averaging the top 100 query scores. The RoBERTa model is faster, generates a higher similarity score, and requires less runtime, which represents the length of time required to finish the programme.

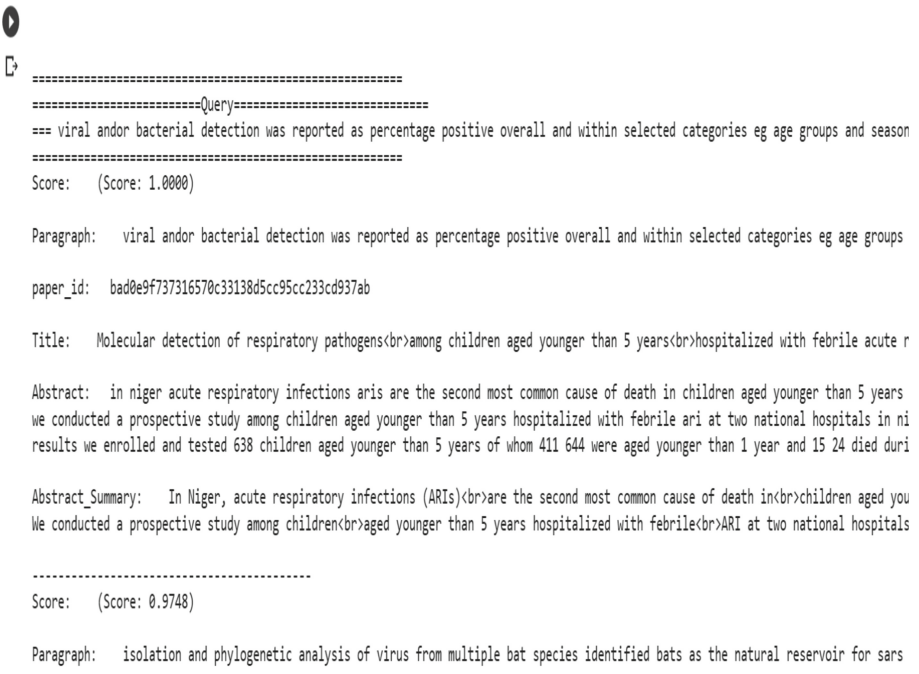

Fig. 2. The result

4.1 Latent Semantic Indexing Model Using SVD

The Latent Semantic Indexing (LSI) [10] technique is utilised in document data to uncover hidden concepts.The elements connected to these concepts will subsequently be shown for each document and word as vectors. Each entry in a vector represents the extent to which the document or phrase participates in the idea.

The aim is to unify documents and terms so that hidden document, document-term and term-semantic relationships may be disclosed. To do matrix decomposition, the researchers utilise Sklearn's TruncatedSVD. The n components option can be used to specify the number of subjects/topics. This work employed Bokeh, a Python data visualisation package, to illustrate the LSI model. As shown in the Fig. 3, The LSI model identifies words and texts that are similar in meaning. Then identify the topics that are useful for a variety of applications, including document clustering, structuring online content available for data mining, and making decisions. Topic modelling is a text mining methodology that identifies co-occurring terms in order to summarise massive collections of textual information. It facilitates the discovery of hidden concepts in documents, annotating them with these topics, and organising massive amounts of raw data.

A topic model [10] is a form of statistical model used to uncover the conceptual "topics" that consist of a collection of texts. Topic modelling is a popular

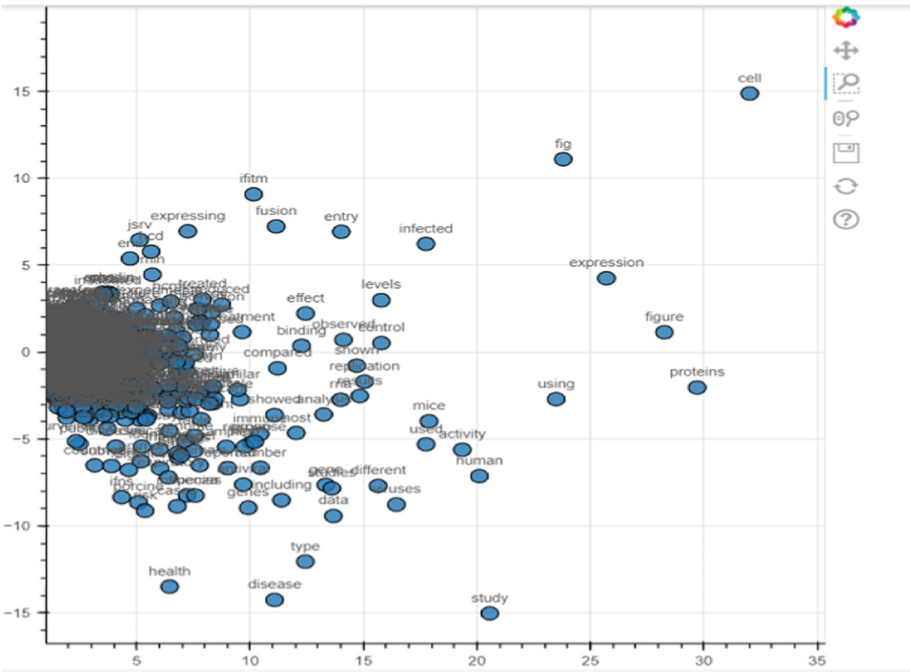

Fig. 3. The topics by using the LSI model

text-mining method for revealing hidden spatial features in a text. The Latent Semantic Analysis was used to identify 14 topics in this work. Topic modelling automatically detects hidden concepts in provided documents. It is used to find a collection of words from a given document. These word groups represent a topic. Figure 3 depicts the visualisation of different topics. The researchers used Python's popular NLTK library to import the stopwords into English. The stopwords are the frequently used words, so the researchers eliminate all stopwords to plot the most commonly used topics.

4.2 Sentence Transformer Models

BERT: The BERT [2] is a bidirectional transformer-based language model. The bidirectional nature of the sentence transformer model allows it to read text sequences from both directions, making the text sequence more meaningful. The BERT employs the NSP task, in which the model attempts to predict whether a sentence A is a continuation of a sentence B. To embed the research paper and input query, the model uses bert-base-nli-mean-tokens. Prior to input word sequences into BERT, the MLM replaces 15% of each sequence's words with a [MASK] token. The model then helps to identify the unmasked values of the masked words using the information contained in the sequence's non-masked words.

RoBERTa: The RoBERTa [7] is an improved BERT Pretraining Approach. It is bidirectional and self-attention techniques improve the efficiency and accuracy of the sentence transformer model in text classification. This language model can process long texts efficiently. The RoBERTa eliminates the NSP task, which in BERT's pre-trained model incorporates dynamic masking by duplicating the training data ten times such that each sequence is masked into ten distinct patterns. The authors tested the removal and addition of NSP loss and found that removing NSP loss improves the model's task performance a bit. To embed the research paper and input query, the model uses the nli-roberta-base-v2 package.

ALBERT: The ALBERT [8] model is a lite variant of BERT that beats assert models for a variety of benchmark datasets due to its minimal memory usage, although it takes longer to train. For evaluation, the researchers used the ALBERT base model, using a sentence transformer. When compared to equivalent BERT models, the ALBERT model has a decreased parameter size. In this work, to embed the research paper and input query, the model uses the twmkn9/albert-base-v2-squad2 package. When compared to other sentence transformer models, the ALBERT model takes a long time to compute.

DistilBERT: The DistilBERT [5], is a distilled version of BERT that uses the distillation method. The approach is then fine-tuned to execute similarly to its larger competitors on a variety of tasks. Pre-training a smaller, faster, and lighter model is less expensive. The whole output distribution of a big neural network can be approximated using a smaller network once it has been trained. To embed the research paper and input query, the model uses the nq-distilbert-base-v1 package.

DistilRoBERTa: The DistilRoBERTa [7] is installed using the Transformer package from the sentence transformer. To embed the research article and query, the model is implemented using the nli-distilroberta-base-v2 package. This model is a distillation of the RoBERTa-base model. It is trained in the same manner as DistilBERT. When compared to RoBERTa, this model uses four times less training data.

The Table 1 shows the similarity scores and programme runtime for each sentence transformer model. When compared to other models, the RoBERTa model is found to offer the best outcomes. The similarity score was obtained by averaging the top 100 query scores for each model. The RoBERTa model is faster and produces a higher similarity score. The runtime is displayed in seconds and explains how long it took to complete the programme.

When compared to the RoBERTa model, ALBERT has a slightly higher similarity score, but the processing time is much longer. So, based on the COVID 19 dataset, All these models conclude that RoBERTa is the best semantic search engine. Here, use the SBERT sentence-transformers package, which makes it extremely straightforward to use BERT, DistilBERT, RoBERTa, ALBERT and

Table 1. Sentence transformer models evaluation

Models	Score (approximate score based on 100 query)	Time (runtime of program in secs)
RoBERTa	0.99994	2081.515028476715
ALBERT	0.99997	7009.639680147171
DistilBERT	0.99873	3101.0192477703094
DistilRoBERTa	0.94985	1067.1314284801483
BERT	0.99901	5267.991491317749

DistilRoBERTa for sentence embedding. The researchers installed sentence - transformers using pip and bokeh, a Python data visualisation package, to illustrate the LSI model.

5 Conclusion

The COVID19 text classification is at the forefront of a number of NLP applications. In this work, a semantic search engine is proposed that utilises different sentence transformer models such as BERT, DistilBERT, RoBERTa, ALBERT and DistilRoBERTa. For the COVID19 semantic search engine, the researchers compared the performance of several techniques extensively. The results reveal that the RoBERTa outperforms the competition in this job. The researchers also use topic modelling to visualise the most important subjects in the acquired dataset. Instead of reading complete research papers and articles, this model provides answers to health-care employee's questions. Future studies could include assessing a semantic search engine employing OpenAI's Generative Pre-trained Transformer (GPT) models. These GPT models are some of the most powerful language models available.

References

1. Ait, C., Hubner, M., Hennig, L.: Fine-tuning Pre-Trained Transformer Language Models to Distantly Supervised Relation Extraction. arXiv preprint arXiv:1906.08646 (2019)
2. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
3. Kassim, J.M., Rahmany, M.: Introduction to semantic search engine. In: International Conference on Electrical Engineering and Informatics, vol. 2, pp. 380–386. IEEE (2009)
4. Patel, M.: TinySearch- Semantics based Search Engine using Bert Embeddings. arXiv preprint arXiv:1908.02451 (2019)
5. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)

6. Guo, X., Ma, J., Li, X.: LSTM-based neural network model for semantic search. In: Yang, H., Qiu, R., Chen, W. (eds.) INFORMS-CSS 2019. SPBE, pp. 17–25. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-30967-1_3
7. Liu, Y., et al.: Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
8. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: a lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942 (2019)
9. Patel, M.: TinySearch-Semantics based Search Engine using Bert Embeddings. arXiv preprint arXiv:1908.02451. Twitter as a tool for the management and analysis of emergency situations: a systematic literature review. Int. J. Inf. Manag. **43**, 196–208 (2019)
10. Akashram. Topic-modeling (2019). www.kaggle.com/akashram/topic-modeling-intro-implementation
11. Allen Institute for AI. COVID-19 Dataset (2020). www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks?taskId=568
12. Alammar, J.: The Illustrated Transformer (2020). http://jalammar.github.io/illustrated-transformer/
13. Nair, P.C., Gupta, D., Devi, B.I.: A survey of text mining approaches, techniques, and tools on discharge summaries. In: Gao, X.-Z., Tiwari, S., Trivedi, M.C., Mishra, K.K. (eds.) Advances in Computational Intelligence and Communication Technology. AISC, vol. 1086, pp. 331–348. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-1275-9_27
14. Menon, R.R.K., Joseph, D., Kaimal, M.R.: Semantics-based topic inter-relationship extraction. J. Intell. Fuzzy Syst. **32**(4), 2941–2951 (2017)
15. Akhil dev, R., Menon, R.R.K., Bhattathiri, S.G.: An insight into the relevance of word ordering for text data analysis. In: 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), pp. 207–213 (2020). https://doi.org/10.1109/ICCMC48092.2020.ICCMC-00040
16. Asritha, P., Prudhvi Raja Reddy, P., Pushpitha Sudha, C., Neelima, N.: Intelligent text mining to sentiment analysis of online reviews. In: ICASISET (2021). https://doi.org/10.4108/eai.16-5-2020.2303907
17. Karumudi, G.V.N.S.K., Sathyajit, R., Harikumar, S.: Information retrieval and processing system for news articles in English. In: 2019 9th International Conference on Advances in Computing and Communication (ICACC), pp. 79–85 (2019). https://doi.org/10.1109/ICACC48162.2019.8986223
18. Baladevi, C., Harikumar, S.: Semantic representation of documents based on matrix decomposition. In: International Conference on Data Science and Engineering (ICDSE) 2018, pp. 1–6 (2018). https://doi.org/10.1109/ICDSE.2018.8527824