Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 2, April 2025, Page. 741-754
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.2.2606

Information Retrieval Related to Information Regarding Covid-19 Using Transformers Architecture

Wiktasari[1], Prayitno*[2], Vinda Setya Kartika[3], Eri Eli Lavindi[4], Naufal Reki Ardhana[5], Rucirasatti Nariswana[6]

[1,2,3,4,5,6]Computer Engineering Technology, Electrical Engineering, Semarang State Polytechnic, Indonesia

Email: [2]prayitno@polines.ac.id

Received : Aug 5, 2024; Revised : Sep 3, 2024; Accepted : Sep 5, 2024; Published : Apr 26, 2025

Abstract

The spread of the COVID-19 virus has occurred exponentially, necessitating advanced search technologies that provide accurate information. The primary challenge in searching for COVID-19 related information involves the diversity and rapid changes in data, as well as the need to understand specific medical contexts. Unstructured information sources, such as research articles, news reports, and social media discussions, add complexity to retrieving relevant and up-to-date information. As the volume of data and information related to the COVID-19 pandemic increases, there is a pressing need for effective and accurate information retrieval systems. Transformer architecture, known for its capabilities in natural language processing and managing complex contexts, offers great potential to enhance search quality in the healthcare domain. BERT is a deep learning model that performs searches based on specific queries, with search results sorted accordingly. The ranking process uses BERT architecture to compare the performance of transformer encoders, specifically between bi-encoders and cross-encoders. A bi-encoder is an architecture where two separate encoders process two different inputs, such as queries and documents. In contrast, a cross-encoder processes two texts simultaneously using a single encoder, allowing the model to capture contextual interactions between them. Research indicates that cross-encoder performance is significantly better than bi-encoder for cases with relatively small data sets. Evaluation results show that the NDCG score for bi-encoder is 0.89, while for cross-encoder it is 0.9. The mAP score for bi-encoder is 0.7, and for cross-encoder, it is 0.89. Both bi-encoder and cross-encoder achieved an MRR score of 1.0.

Keywords : *BERT, Bi-Encoder, Cross-Encoder, Information Retrieval, Transformers.*

1. INTRODUCTION

The spread of the Covid-19 virus occurs exponentially beyond regional borders through close contact between humans which spreads through transportation, both land and air. Various efforts have been made to prevent the spread of the virus which has quickly changed its status to a global pandemic. All areas of life are experiencing changes, starting from the economic, health, education, social and other fields, forcing residents to stay at home to avoid spreading the virus [1].

This condition encourages breaking the chain of virus spread in various ways. One solution is to provide correct information [1]. The amount of information currently available is so large that it is necessary to retrieve information quickly and correctly according to needs [2]. Information retrieval is a method of retrieving text from documents according to a given query which aims to obtain relevant documents. This method consists of retrieving information from parts of the document such as such as image captioning, paragraph, caption, and question answering [3].

Several studies regarding information seeking have been carried out by researchers. Research conducted by Otegi et al developed an information system that can help biosanitation experts to access, consult and analyze publications related to COVID-19. The information system developed is

P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 2, April 2025, Page. 741-754
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.2.2606

Questioning Answering to receive questions related to COVID-19, the answers to these questions present documents ranked according to the level of relevant experts. The information retrieval model is SciBERT which is used to retrieve and extract answers with the results of ranking documents with the best answers [4].

Research conducted by Alexander Turchin et al regarding BERT has been proven to be able to solve many problems related to NLP. This research calculates accuracy by identifying text related to complex medical terms. From this research, the recall was more than 80% and the precision was above 75%. By using various categories, an F1-Score is obtained with a range from 0.0 to 0.860 [5].

It has been said that the BERT (Bidirectional Encoder Representations from Transformers) model architecture is state-of-the-art for tasks including entity recognition, question answering, and text extraction [54]. In contrast to previous models, BERT does not provide you a single word embedding for every word after training. It offers a model that, given the entire sentence, creates a word integration for each word that appears in the statement [6].

Research conducted by Kevin Peyton et al used SBERT (Sentence BERT). They use this method with an API framework, namely transformers which uses an encoder-decoder architecture which has the ability to interpret the input obtained. Phrase classification testing and results in the form of an F1 comparison score. The average F1 score obtained from Google of 0.96 and Microsoft QnA of 0.96 [7].

The BM25 approach was utilized in research by Khalisma et al. to search for news in the Indonesian language. In order to ascertain the R-rank of documents for queries containing R-relevant documents, this study used R-Precision evaluation. 300 documents are utilized as training data, and 12 queries are used as testing data to yield the best r-precision value in Q1 and Q2, with a value of 1. Because all pertinent documents were placed at the top, this value was achieved [8].

Uthirapathy et al conducted research on analyzing topics and opinions regarding climate change discussed by the public via social media such as Twitter, Facebook and Weibo. The methods used are Latent Dirichlet Allocation (LDA) and BERT which are Deep Learning techniques for conducting sentiment analysis of the data sets used. Sentiment was labeled as news, sports, neutral and anti. The research results showed that the performance of the sentiment classification model was with a precision value of 91.35%, recall was 89.65% and accuracy of 93.50% [10].

Research conducted by Oliaee et al analyzed data regarding traffic accidents. By using BERT data of 750.00 it will be classified based on the type of injury experienced when an accident occurs. The results of this research are an accuracy value of 84.2% and an area under the receiver operating curve (AUC) of 0.93 [11].

Research conducted by Liu et al regarding initial diagnosis regarding the accuracy and efficiency of maintenance of power system equipment to prevent overhead transmission lines. The model used is BERT which is considered effective in increasing the effectiveness of information extraction needed to prevent overhead transmission lines. The results of research using RoBERTa had better accuracy with a value of 92.22% [12].

Research conducted by Sheher Bano et al regarding to tackle the summarization of scholarly articles are currently exploring novel approaches that utilize deep learning models such as BERT (Bidirectional Encoder Representations from Transformers). Unfortunately, due to input length limitations, BERT is not as effective at summarizing long papers. We suggest an innovative method to identify a better solution. This method combines the strength of a transformer network that has been pre-trained on large amounts of self-supervised datasets (BERT) [13]. Kalamajit et al.'s research on BERT classification issues yielded an F1-score weight of 85% for each class [14].

Document search applications were examined using FMeasure, recall, and precision in other studies. The BM25 method has a higher precision value than the PLSA method because, when searching for documents, it finds several that do not match the query and for which the document weight is not

Vol. 6, No. 2, April 2025, Page. 741-754
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.2.2606

detected; in contrast, when searching for documents using the PLSA method, it finds both documents that match the query and documents that do not. The test above demonstrates that the recall value stays at 100%, meaning that the system is able to locate all documents in the document collection that match the query [15].

2. METHOD

The research flowchart can be seen in Figure 1. In Figure 1 you can see that the first step taken was collecting data from the alodokter.com website. Data collected using the web crawling method. The data processing or reprocessing process is carried out to prepare data so that further processing can be carried out.

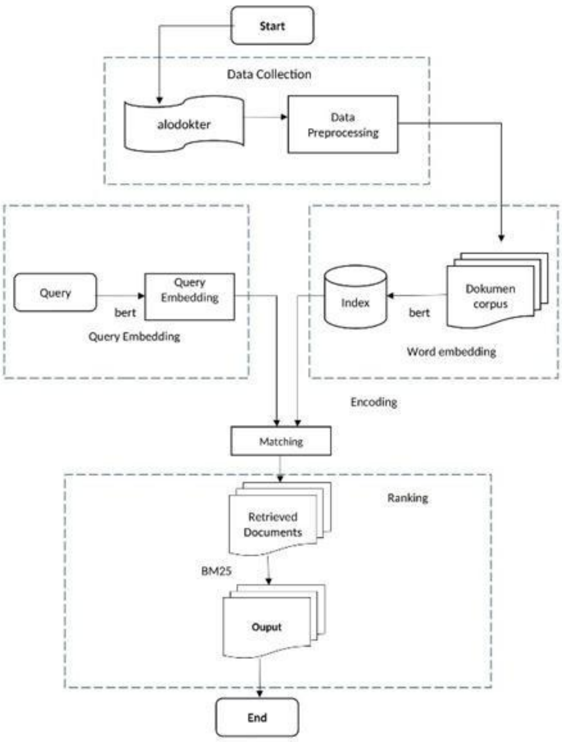


Figure 1. Flowchart System

The next step is to carry out word embedding for the document data and for the given query. Word embedding is done to change the text data format into numeric data. The next step is to match the given query with the available data (encoding). Encoding is carried out to find data in available documents that matches the given query. The next step is to rank the data resulting from the matching process. The next step displays search results and rankings according to the given query.

2.1. Data Preprocessing

Data preprocessing is a step in preparing data for further processing. The process carried out is changing the form of data according to needs [8]. The steps that will be taken in this research are case folding, stopwords, tokenizing, and stemming.

Case folding is a process for changing text data in the form of uppercase letters to lowercase. Stopword is the process of eliminating words that have no meaning. Tokenization is the process of breaking sentences into smaller units. The final step is stemming, which is the process of changing words that have affixes into their base words [8][9].

In this study, data was taken from the health website alodokter.com. Data was taken using the web crawling method. Because this research aims to test models for matching and ranking, the data that is crawled

Vol. 6, No. 2, April 2025, Page. 741-754
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.2.2606

will be adjusted to the query that will be tested, mixed with other data that has prepositions that are similar to the query. For example, the query given is COVID-19, then the crawl results contain the word COVID-19 and the data is mixed with other data that contains almost similar words, for example cobalt. So that when the model is run the results obtained can be analyzed for suitability and the order in which they appear.

The first step of preprocessing is collecting data from the alodokter.com website and random data from other documents.

2.1.1 Parsing

The first step in web crawling the alodokter.com website is parsing the data, namely changing the HTML data format to plain text. The data that is parsed from the web is url_path, image_url, category, title, short_description, and article. There are issues with obtaining data from HTLM; specifically, the table and div tags are among the HTML tags that encounter issues. To solve this issue, special parsing for the tag must be done.

2.1.2 Sanitizer

The next step is sanitizer, namely cleaning unnecessary data such as punctuation and changing the letter pattern to lowercase Remove Punctuation and changing the letter pattern to lowercase.

2.1.3 Remove Stopword

This step, also known as filtering, involves selecting key words from the token results using a wordlist or stoplist algorithm to maintain the relevant words and remove the less important ones.

2.1.4 Tokenisasi

Tokenization is the process of dividing text into manageable chunks for study at a later time. Tokens include words, numbers, symbols, punctuation, and other significant elements.

2.1.5 Stemming

The next step is stemming, which is a tactic that is also required to determine the number of different indexes from a single set of data so that a word with a suffix or a prefix will return to its original form. With the help of this simple Python library called sastrawi, you can reduce Indonesian inflected words (Bahasa Indonesian) to their stem form.

Then, from the preprocessing steps that have been carried out, the data is obtained in the form of a dataframe with 100 data containing words that match the query to be given. The data obtained can be seen in table 1.

Table 1. Dataframe

url_path	….	Article
https://www.alodok...	.....	covid19 adalah penyakit...
https://www.alodok...	….	virus corona atauâ severe..
https://www.alodok...	….	vaksin moderna adalah…

The data obtained is then changed to jsonl (json line) format and combined with data that comes from other documents but does not contain the specified query. So the final data is a combination of data that contains a query and data that does not contain a query. The two documents were then randomized and 6222 data were obtained.

Vol. 6, No. 2, April 2025, Page. 741-754
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.2.2606

2.2. BERT

BERT is the latest technique from NLP for processing text data. BERT has a more competitive architecture compared to other NLP models. This model uses layer transformers with a pre-training process using diverse self-supervised objectives so that it has more comprehensive capabilities and performance enhancement [13][14].

In this study, BERT model leverages workers' labor skills as sentence-transformers. This work schedule is used to manage text, images, and numbers. Based on Pytorch and Transformers, this framework offers several models that have already been studied and can be easily used in this research, making the process of creating new models easier. Because there are many models that have already been studied, researchers can easily select a model that best suits their research needs. The work-related stress in this area slows down the fine-tuning process and makes it easier to make more efficient decisions, which improves the performance of the analysis in the BERT application.

2.3. Word Embedding

Text modeling can be analogous to mathematical modeling. However, the text data format must first be converted into vector form. Encoding is the step taken to convert text into vector form. The BERT method uses a transformer framework consisting of an encoder. The encoder functions to read input data in the form of text. The input is read in the form of a sequence which must first be converted into vector form when it will be processed by the encoder [10][11].

In this research, the data is in jsonl (json line) format which is stored in a file. The word embedding process is carried out by decoding using UTF-8 to convert non-numeric data into numeric data. The decoded data is taken based on the lines in the jsonl file.

2.4. Transformers

Text modeling can be analogous to mathematical models. However, the text data format must first be converted into vector form. Encoding is the step taken to convert text into vector form. The BERT method uses a transformer framework consisting of an encoder that performs block endocing and labeling. The encoder functions to read input data in the form of text. The input is read in the form of a sequence which must first be converted into vector form when it will be processed by the encoder [10] [16].

BERT which is based on ranking or Neural Ranking Models is classified based on how queries and documents are encoded according to the layer they have. There are two types of encoders, namely bi-encoders which are more time efficient encoders and cross-encoders which are more precise encoders in finding data that matches a given query [20]. In this research will compare bi-encoders with cross-encoders by determining evaluation values based on MRR, mAP and NDCG.

2.5. BM25

BM25 is a method used to rank word search results in a corpus document. BM25 is the Best Match class which has the best formula because it is effective and accurate in returning results in the order based on the given query [15][16].

In this research, the encoding results will be ranked using bm25_scores, which is a function used to sort based on the given query. The score for each encoder will be calculated and the 5 highest values will be taken.

Data modeling using BERT using a bi-encoder and cross-encoder using the queries "covid-19", "virus" and "corona". The encoding results can be seen in Figure 2. The results of the modeling obtained five data in order of highest score to lowest score. The highest score value for the query "covid19" using the bi-encoder shows a value of 0.511 and for the cross-encoder it shows a value of

Vol. 6, No. 2, April 2025, Page. 741-754
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.2.2606

0.920. The highest score value for the query "virus" using the bi-encoder shows a value of 0.582 and for the cross-encoder it shows a value of 0.707. The highest score value for the query "corona" using the bi-encoder shows a value of 0.426 and for the cross-encoder it shows a value of 0.630. By using three different queries the score results obtained by cross-encoding give a higher score than bi-encoder because cross-encoder provides better performance when using data that is not too large.

```
[ ] search(query = "Covid19")

Input question: Covid19
Top-5 lexical search (BM25) hits
        0.881   virus corona penyebab covid19 masih terus be
        0.953   sejak awal kemunculannya virus corona penyeb
        0.993   virus corona penyebab covid19 masih terus be
        0.998   sebagian orang mungkin beranggapan bahwa den
        1.000   covid19 adalah penyakit akibat infeksi virus
Top-5 Bi-Encoder Retrieval hits
        0.511   covid19 adalah penyakit akibat infeksi virus
        0.459   virus corona penyebab covid19 masih terus be
        0.414   virus corona penyebab covid19 masih terus be
        0.351   sejak awal kemunculannya virus corona penyeb
        0.346   Qviding FIF is an association football club
Top-5 Cross-Encoder Re-ranker hits
        0.970   covid19 adalah penyakit akibat infeksi virus
        0.940   virus corona penyebab covid19 masih terus be
        0.925   sejak awal kemunculannya virus corona penyeb
        0.893   virus corona penyebab covid19 masih terus be
        0.722   vaksin moderna adalah vaksin untuk melindung

[ ] search(query = "Virus")

Input question: Virus
Top-5 lexical search (BM25) hits
        1.000   Poliomyelitis, or polio, is a virus that
        1.000   Acellular or non-cellular life is life th
        1.000   Glandular fever is a viral infection caus
        0.993   virus corona penyebab covid19 masih terus
        1.000   covid19 adalah penyakit akibat infeksi vi
Top-5 Bi-Encoder Retrieval hits
        0.582   virus corona penyebab covid19 masih terus
        0.510   virus corona penyebab covid19 masih terus
        0.483   sejak awal kemunculannya virus corona pen
        0.473   covid19 adalah penyakit akibat infeksi vi
        0.425   Anthrax, or splenic fever, is a disease.
Top-5 Cross-Encoder Re-ranker hits
        0.707   Poliomyelitis, or polio, is a virus that
        0.535   Acellular or non-cellular life is life th
        0.336   virus corona penyebab covid19 masih terus
        0.321   Influenza, better known as the flu and so
        0.253   vaksin sinopharm adalah vaksin untuk menc

search(query = "Covid")

Input question: Covid
Top-5 lexical search (BM25) hits
        0.993   virus corona penyebab covid19 masih teri
        1.000   205||0||23||0||19||0||17||0||264||0
        1.000   Manute Bol (born 16 October 1962 - died
        1.000   Naoki Matsuyo (born 9 April 1974) is a 1
        1.000   John Devon Roland Pertwee (7 July 1919 -
Top-5 Bi-Encoder Retrieval hits
        0.426   Covasna (, ) is a county (judet) of Rom
        0.425   Cobalt(III) fluoride, also known as cob
        0.423   covid19 adalah penyakit akibat infeksi \
        0.416   Cobalt(II) fluoride, also known as coba
        0.412   Qviding FIF is an association football
Top-5 Cross-Encoder Re-ranker hits
        0.630   covid19 adalah penyakit akibat infeksi \
        0.580   virus corona penyebab covid19 masih teri
        0.564   sejak awal kemunculannya virus corona pe
        0.445   virus corona penyebab covid19 masih teri
        0.243   The Safavids () were a dynasty of ruler
```

Figure 2. Search Results with The Query Given

Encoding with a cross-encoder produces a score with a value above 1, this will be normalized to a value in the range 0-1 using the sigmoid function. The normalization results can be seen in Figure 3.

Vol. 6, No. 2, April 2025, Page. 741-754
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.2.2606

```
search(query = "Covid19")

Input question: Covid19
Top-5 lexical search (BM25) hits
        0.881   virus corona penyebab c
        0.953   sejak awal kemunculanny
        0.993   virus corona penyebab c
        0.998   sebagian orang mungkin
        1.000   covid19 adalah penyakit
Top-5 Bi-Encoder Retrieval hits
        0.511   covid19 adalah penyakit
        0.459   virus corona penyebab c
        0.414   virus corona penyebab c
        0.351   sejak awal kemunculanny
        0.346   Qviding FIF is an assoc
Top-5 Cross-Encoder Re-ranker hits
        0.970   covid19 adalah penyakit
        0.940   virus corona penyebab c
        0.925   sejak awal kemunculanny
        0.893   virus corona penyebab c
        0.722   vaksin moderna adalah v
```

Figure 3. Score Normalization with Sigmoid

2.6. Evaluation

The final step of this research is to evaluate the process that has been carried out. This research compares three evaluation methods of query ranking results carried out by the offered models. The evaluations used are MRR, mAP and NDCG.

MRR (Mean Reciprocal Rank) is an evaluation method that determines unique correlation which is commonly used to assess recommendation systems. The MRR score is obtained from calculating all the data included in the ranking [18]

$$\text{MRR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i} \tag{1}$$

Mean Average Precision (mAP) is an evaluation method by calculating the average of average precision (AP). The mAP value is determined by calculating all the average AP values to find the intersection area on the recall and precision curves [19].

$$\text{mAP@t} = \frac{1}{k}\sum_{k=0}^{k-1}\text{AP@t} \tag{2}$$

NDCG (Normalized Discounted Cumulative Gain) determines the position of the highest hits resulting from the ranking. The highest value indicates the most recommended value [19].

$$\text{nDCG} = \frac{1}{Q}\sum_{q=1}^{Q}\frac{DCG_p^{(g)}}{IDCG_p^{(g)}} \tag{3}$$

3. RESULT

This research aims to test the information retrieval model with the BERT model using a transformers architecture which is considered more efficient than previous NLP models. The BERT architecture used is a transformers framework using an encoder for the process of matching documents with queries. This research uses two encoders, namely bi-encoder and cross-encoder. The results of document encoding are then ranked using BM25. The scores from the two encoders are then evaluated using the MRR, mAP and NDCG methods to see which encoder is more efficient for searching data using queries.

The results of data encoding using three queries using 6222 data produced a score which can be seen in table 2. The first ranking score for the query "covid19" using a bi-encoder showed a value of 0.511 and of the 5 highest data there was 1 data that was not relevant to the query. The first ranking score for the query "covid19" using a cross-encoder shows a value of 0.920 and of the 5 highest data there is no data that is not relevant to the query. The first ranking score for the query "virus" using a bi-

Vol. 6, No. 2, April 2025, Page. 741-754
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.2.2606

encoder shows a value of 0.582 and of the 5 highest data there is 1 data that is not relevant to the query. The first ranking score for the query "virus" using a cross-encoder shows a value of 0.707 and of the 5 highest data there is no data that is not relevant to the query. The first ranking score for the query "covid" using a bi-encoder shows a value of 0.426 and of the 5 highest data there are 4 data that are not relevant to the query. The first ranking score for the query "covid" using a cross-encoder shows a value of 0.630 and of the 5 highest data there are 2 that are not relevant to the query.

Table 1 Score Encoding Results

Query	Encoder	Score	Encoder	Score
Covid-19	Bi-encoder	0.511	Cross-encoder	0.970
		0.459		0.940
		0.414		0.925
		0.351		0.893
		0.346		0.722
Virus	Bi-encoder	0.582	Cross-encoder	0.707
		0.510		0.535
		0.483		0.336
		0.473		0.321
		0.425		0.253
Covid	Bi-encoder	0.426	Cross-encoder	0.630
		0.425		0.580
		0.423		0.564
		0.416		0.445
		0.412		0.243

Based on the score results and the relevance of the data obtained based on the given query, it can be concluded that the cross-encoder has better performance in information retrieval with better relevance results compared to the bi-encoder. The score shows the similarity of a pair of sentences with the input query with a value of 0-1, where the closer the value is to 1, the greater the similarity value. For the 3 queries given, it can be seen that the cross-encoder has a higher score compared to the bi-encoder. This shows that the cross-encoder has higher performance in obtaining data with a high level of similarity compared to the bi-encoder according to the given query.

From table 2, it can be shown that the cross-encoder is more robust since it can handle complex input interactions when two inputs are presented in a cooperative manner. This makes it possible to learn representations that are highly contextual and relevant by extending the input context in a cooperative manner. For example, in optimization or debugging tasks, cross-encoder can effectively capture information and the relationship between the query and the document more effectively than a model that captures both of them in an imprecise way. However, because the cross-encoder combines two inputs in a single way, it frequently requires more computing power and processing time. This could be useful in large-scale applications or real-time time windows.

It is biased to favor work-related learning tasks that require contextual understanding and input-to-output interaction, such as document review tasks or relevancy assessment tasks. An example of a specific example would be a data-driven writing system similar to the one used in this study, where data and queries have a crucial relationship.

Conversely, the bi-encoder splits the encoding process into two distinct steps, which may reduce its ability to capture input-to-output interactions. The representation obtained for each individual input may not fully capture the details of the relationship between the two. Though, in terms of computer performance, bi-encoders are often faster in terms of inference since their representations for queries and documents may be accessed independently and then compared. This makes it an excellent choice for systems that require quick responses, as search sensors.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 6, No. 2, April 2025, Page. 741-754
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.2.2606

Table 2. Evaluation result

Encoder	Query	mAP	MRR	NDCG
Bi-encoder	Covid19	0.7	1.0	0.89
	Virus	0.5	1.0	0.82
	Covid	0.16	0.33	0.41
Cross-encoder	Covid19	0.83	1.0	0.94
	Virus	0.83	1.0	0.91
	Covid	0.7	1.0	0.93

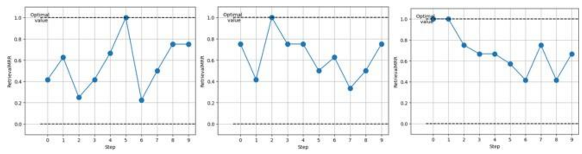
Table 3 shows the results of the evaluation scores obtained. The first evaluation using mAP showed that the cross-encoder had a higher score, namely 0.83, while the bi-encoder had a score of 0.7. This shows that the average value of the average precision (AP) of the cross-encoder is higher than that of the bi-encoder. The second evaluation using MRR for both encoders produces the same value, namely 1.0, so that the unique correlation value for the recommendation system for both encoders can run well. The third evaluation using NDCG shows that the cross-encoder gives a score of 0.94 for the highest rank, while the bi-encoder has a score of 0.89 for the highest score.

4. DISCUSSIONS

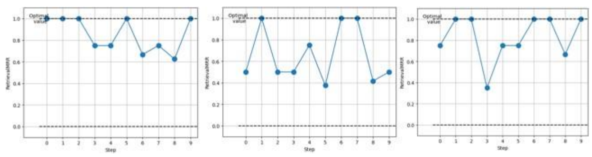
Previous research [8] using BERT for text identification cases with various categories obtained F1-Scores ranging from 0.0 to 0.860. This shows that BERT's performance is good enough for text classification. However, some BERT architectures still have shortcomings because they have not carried out a comprehensive pre-training process. So the new architecture using transformers is considered to have good performance.

In this research, text matching is carried out by comparing two types of encoder transformers in the BERT framework architecture, namely bi-encoder and cross-encoder. The score obtained for each encoder will be calculated and then ranked. The data used in this research was 6222 data originating from the health website alodokter.com which had been adjusted to the query that would be given mixed with random data that did not contain the query. The score value for the top ranking using the cross-encoder is 0.970 while for the bi-encoder it is 0.511.

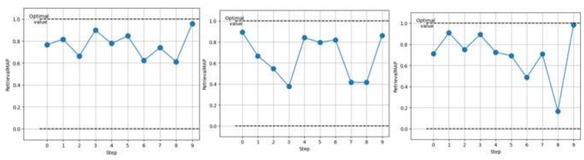
MRR
Bi-encoder



Cross-encoder

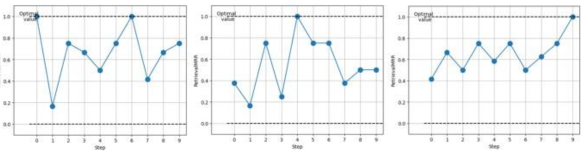


mAP
Bi-encoder

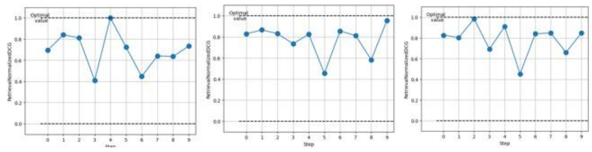


Cross-encoder

Vol. 6, No. 2, April 2025, Page. 741-754
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.2.2606



NDCG
Bi-encoder



Cross-encoder

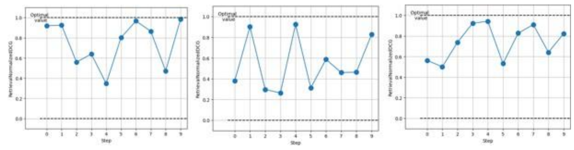


Figure 3. Evaluation Comparison Chart with MRR, mAP and NDCG.

With a more variable performance and a more robust mean approximate precision across all queries, the bi-encoder indicates instabilities in query processing. In every query, the cross-encoder yields results that are more consistent and stable. This indicates that, in comparison to Bi-encoder, Cross-encoder performs better in terms of identifying relevancy and handling a variety of query types. Figure 4 shows that the ranking scores obtained from the two encoders will be compared and the resulting match will evaluate its accuracy with the given query. Three evaluation methods are used to determine the appropriate model to produce documents that match the query. The results of the comparison of evaluation methods using mAP, MRR and NDCG can be seen in Figure 4. Evaluation using MRR shows the score for both encoders reaches a value of 1.0. This shows that the unique correlation for the recommendation system for both encoders can work well. For evaluation with mAP, the highest score for the cross-encoder reaches a value above 0.89, while for the bi-encoder the highest value is 0.7, so the average average precision (AP) of the cross-encoder is higher than the bi-encoder. For evaluation using NDCG, it shows a relatively high score, namely 0.94, while the bi-encoder shows a score of 0.89, this shows that the cross-encoder has a higher score for the top ranking. Based on research conducted, the cross-encoder has a higher score, so it can be concluded that the cross-encoder has better performance in matching data with the given query.

Measures the average precision of a model across all queries. The cross-encoder scores higher than the bi-encoder in mAP, with a maximum of 0.89 compared to 0.7. This suggests that the cross-encoder has better precision in ranking relevant documents.

Measures the average rank at which the first relevant document is retrieved. Both encoders achieve a perfect score of 1.0 in MRR, indicating that for the most part, the first relevant document is retrieved at the highest possible rank.

Measures the relevance of the results while accounting for the position of relevant documents. The cross-encoder scores higher (0.94) compared to the bi-encoder (0.89), indicating better performance in ranking the most relevant documents higher.

5. CONCLUSION

In this study, researchers tried to analyze the data matching model, in this case the data used is text data. The science used is information retrieval using Natural Language Processing. The goal of information retrieval is to find information that is relevant to a given query. BERT is a method that can be used for information retrieval. Researchers use a framework called transformers, the newest BERT

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.2606

architecture and has the advantage of better model pre-training compared to other BERT architectures. Researchers compared two encoder transformers, namely using a bi-encoder and a cross-encoder.

The data used in this research was 6222 data originating from alodokter.com and random text data selected to test the accuracy of the model presented. The final result of this research is a comparison of two encoders, namely bi-encoder and cross-encoder and the resulting score value for the highest ranking using the NCGD evaluation with a cross-encoder score of 0.94 while the bi-encoder shows a score of 0.89. So it can be concluded that performance of cross-encoder is better than bi-encoder in matching data for information retrieval.

In this study, researchers analyzed text data using information retrieval techniques with Natural Language Processing (NLP). They employed BERT, specifically the latest transformer architecture, to improve model pre-training. The researchers compared two encoder models: bi-encoder and cross-encoder. The dataset comprised 6,222 entries from alodokter.com and additional random text data. The study evaluated the models' performance using the NCGD metric. The results indicated that the cross-encoder outperformed the bi-encoder, achieving a score of 0.94 compared to 0.89 for the bi-encoder. Therefore, the cross-encoder demonstrated superior performance in data matching for information retrieval.

Based on the findings of the study, here are some suggestions for further research and practical applications:

1. Exploration of Additional Metrics: While the NCGD metric provided useful insights, incorporating additional evaluation metrics (such as Precision, Recall, F1-Score, and Mean Reciprocal Rank) could offer a more comprehensive assessment of model performance. This would help in understanding different aspects of the models' effectiveness in various contexts.
2. Dataset Expansion: Although the dataset of 6,222 entries is substantial, exploring larger and more diverse datasets could help in validating the generalizability of the cross-encoder model. Incorporating data from different domains or sources might reveal insights into the model’s robustness and adaptability.
3. Comparative Analysis with Other Models: Including other advanced transformer models or variations, such as GPT-4 or T5, could provide a broader comparison and potentially highlight even more effective approaches for text data analysis and retrieval.
4. Fine-Tuning and Optimization: Investigate the impact of fine-tuning hyperparameters and model architectures on the performance of the cross-encoder. Optimizing these aspects might further enhance the model’s efficiency and accuracy.
5. Real-World Application Testing: Implement the cross-encoder in real-world scenarios to evaluate its performance in practical applications. Testing the model on live data or within specific industry contexts can offer insights into its effectiveness beyond controlled study conditions.
6. User Feedback and Adaptation: Gathering feedback from end-users who interact with the model in practical settings can provide valuable information on its usability and effectiveness. Adapt the model based on this feedback to better meet user needs.
7. Performance Analysis in Various Languages: If the dataset includes or can be expanded to include text in different languages, evaluating the performance of the cross-encoder across various languages could be beneficial, especially for applications in multilingual environments.
8. Exploring Model Interpretability: Investigate methods to enhance the interpretability of the cross-encoder model. Understanding why the model makes certain decisions can be crucial for trust and transparency in information retrieval systems.

These suggestions aim to build upon the study's findings and explore ways to enhance the model’s performance and applicability in diverse scenarios.

Vol. 6, No. 2, April 2025, Page. 741-754
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.2.2606

REFERENCES

[1] E. Andre, K. Anuprit, P. Romain, H. Kazuma, Y. Wenpeng, R. Dragomir, and S. Richard, "COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization," *npj Digital Medicine*, vol. 68, pp. 21–37, 2021.

[2] C. Tanmoy, L. G. Valerio, M. Vincenzo, and S. Giancarlo, "Information retrieval algorithms and neural ranking models to detect previously fact-checked information," *Neurocomputing*, vol. 557, pp. 66–80, 2023.

[3] G. Yan and C. Georgina, "Improving visual-semantic embeddings by learning semantically-enhanced hard negatives for cross-modal information retrieval," *Pattern Recognition*, vol. 137, pp. 159–272, 2023.

[4] O. Arantxa, S. V. Iñaki, S. Xabier, P. Anselmo, L. Borja, and A. Eneko, "Information retrieval and question answering: A case study on COVID-19 scientific literature," *Knowledge-Based Systems*, vol. 240, pp. 60–72, 2022.

[5] T. Alexander, M. Stanislav, and Z. Marinka, "Comparison of BERT implementations for natural language processing of narrative medical documents," *Informatics in Medicine Unlocked*, vol. 36, pp. 11–19, 2023.

[6] J. L. Z. M. João and C. A. d. C. Cristiano, "The HoPE Model Architecture: a Novel Approach to Pregnancy Information Retrieval Based on Conversational Agents," *Journal of Healthcare Informatics Research*, vol. 6, pp. 253–294, 2022.

[7] K. P. Kevin and U. Saritha, "A comparison of chatbot platforms with the state-of-the-art sentence BERT for answering online student FAQs," *Results in Engineering*, vol. 17, pp. 10–25, 2023.

[8] F. Khalisma, Indriati, and P. A. Putra, "Pencarian Berita Berbahasa Indonesia Menggunakan Metode BM25," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 3, pp. 2589–2595, 2019.

[9] T. P. Diaz and B. S. Erwin, "Sentiment Analysis on Social Media with Glove Using Combination CNN and RoBERTa," *Rekayasa Sistem dan Teknologi Informasi*, vol. 3, pp. 457–563, 2023.

[10] E. U. Samson and S. Domnic, "Topic Modelling and Opinion Analysis On Climate Change Twitter Data Using LDA And BERT Model," *Procedia Computer Science*, vol. 218, pp. 908–917, 2023.

[11] H. O. Amir, D. Subasish, L. Jinli, and R. Ashifur, "Using Bidirectional Encoder Representations from Transformers (BERT) to classify traffic crash severity types," *Natural Language Processing Journal*, vol. 3, pp. 100–112, 2023.

[12] L. Jiaxin, W. Shuai, Z. Qianqian, Z. Chenchen, Z. Guogang, Z. Zijian, and L. Xuchen, "Overhead transmission line condition assessment based on intention classification and slot filling using optimized BERT model," *Energy Reports*, vol. 9, pp. 838–846, 2023.

[13] S. Bano, S. Khalid, N. M. Tairan, H. Shah, and H. A. Khattak, "Summarization of scholarly articles using BERT and BiGRU: Deep learning-based extractive approach," *Journal of King Saud University*, vol. 557, p. 101739, 2023.

[14] K. Kamaljit and K. Parminder, "BERT-CNN: Improving BERT for Requirements Classification using CNN," *Procedia Computer Science*, vol. 2018, pp. 2604–2611, 2023.

[15] P. TJasman, G. H. Milda, and R. Rizky, "Implementasi dan Perbandingan Metode Okapi BM25 dan PLSA pada Aplikasi Information Retrieval," in *Proc. Seminar Nasional Informatika dan Aplikasinya (SNIA)*, 2021.

[16] A. Haya, E. Tamer, and J. J. Bernard, "Improving conversational search with query reformulation using selective contextual history," *Data and Information Management*, vol. 7, pp. 100–125, 2023.

[17] T. J. Hai, Z. B. Y. Bo, H. Chao, L. W. H. Wen, X. W. B., F. B. Yu, and R. Li, "Application of graph neural network and feature information enhancement in relation inference of sparse knowledge graph," *Journal of Electronic Science and Technology*, vol. 21, pp. 100–124, 2023.

[18] A. H. Ramadhan, S. Casi, and A. N. Ratna, "Deteksi Pelanggaran Parkir pada Bahu Jalan Tol dengan Intelligent Transportation System Menggunakan Algoritma Faster R-CNN," *e-Proceeding of Engineering*, vol. 9, pp. 1047–1060, 2022.

[19] W. Beinan, "A parallel implementation of computing mean average precision," *arXiv preprint*, arXiv:2206.09504v1 [cs.CV], 2022.

Jurnal Teknik Informatika (JUTIF) Vol. 6, No. 2, April 2025, Page. 741-754
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.2606

[20] R. Mehrdad, F. Vahid, A. Sajad, J. J. S. M., and O. Mourad, "A novel healthy and time-aware food recommender system using attributed community detection," *arXiv preprint,* arXiv:2103.06523v2 [cs.IR], 2021.

Vol. 6, No. 2, April 2025, Page. 741-754
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.2.2606