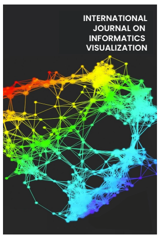
INTERNATIONAL JOURNAL
ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv



Performance Improvement of Cosine Similarity Algorithm with Bidirectional Encoder Representations from Transformers on Abstract Document Similarity Detection

Musthofa Galih Pradana [a,*], Nindy Irzavika [a], Nurhuda Maulana [a], Jesselyn Mu [a], Valtrizt Khalifah Wari [a]

[a] Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jakarta, South Jakarta, DKI Jakarta, Indonesia
Corresponding author: [*]musthofagalihpradana@upnvj.ac.id

***Abstract*—In thesis courses or final projects, students are required to be able to conduct research by the science they are engaged in, find innovations, solve problems, and foster a culture and critical mindset. However, the issue that is often encountered is plagiarism. Plagiarism is taking a work that can be in the form of someone else's opinion and making it seem as if it is your own. The step in applying technology that can be done is to carry out early detection of the similarity of documents written by students. In this case, the document that will be detected is an abstract that must be collected by students when submitting a thesis title. The algorithm used is a cosine similarity algorithm, which is computationally efficient because of its ease of interpretation and compatibility with large-scale data. This research was carried out using two schematic approaches: bidirectional encoder representations from transformers (BERT) and not bidirectional encoder representations from transformers (BERT). The corpus data used in this study was 1450 data of student thesis abstract documents, with the test using 10 data to see the performance of the cosine similarity algorithm in detecting the similarity of abstract documents. The results showed that documents with optimization using the Bidirectional Encoder Representations from Transformers (BERT) approach had better results, with an average performance improvement of 23.48%.**

***Keywords*— Abstract; cosine similarity; bidirectional encoder representations; transformers.**

Manuscript received 28 Jun. 2024; revised 17 Sep. 2024; accepted 8 Oct. 2024. Date of publication 31 Mar. 2025.

I. Introduction

The Preamble to the 1945 Constitution and Articles 31 and 32 state that one of the roles of the state government is to educate people's lives [1]. Education is related to the future of a country and should be prioritized. Moreover, the progress or retreat of a country is primarily determined by the quality of its human resources. One of the critical levels for developing human resources is at the university level [2]. The process of education, research, and service, known as the Tridharma of Higher Education, is an obligation. Research significantly impacts higher education since it can encourage innovation and knowledge development and solve complex problems in various disciplines. The research activities in the academic community involve both lecturers, as facilitators in the Tridharma of Higher Education, and their students. From the students' perspective, completing a thesis or final project can encourage them to participate actively in research activities. In a thesis or final project course, students must be able to conduct research following the science they are engaged in, find innovations, and solve problems. Currently, the acceleration and advancement of artificial intelligence-based technology make it easier to access data and information while promoting plagiarism [3]. Plagiarism can be defined as taking a work, including someone's opinion, and making it seem theirs. Plagiarism, if left unchecked, will cause massive corrosion of the noble values of education [4].

In the context of higher education as a place of production and transfer of knowledge, the government issues regulations on the Prevention and Control of Plagiarism in Higher Education [5]. However, it does not stop there; a preventive way is needed for this act of plagiarism, one of which is by utilizing technology [6]. Technology does have two blades; as long as it is used wisely, it will benefit human work.

Technology can be used to detect the similarity of students' manuscript abstracts early when submitting a thesis title. The manuscript is compared to another to decide on the admission of students' thesis titles. Many algorithms can be applied to detect the similarity of text documents [7], one of which is the cosine similarity algorithm. It is computationally efficient because of the ease of interpretation and compatibility with large-scale data [8]. This is certainly relevant to students' abstracts, which, from time to time, will undoubtedly grow more extensive and need a machine-learning approach [9]. This study comprehensively examines the application of the cosine similarity algorithm, combined with word weighting using TF-IDF, to obtain more optimal and accurate detection results using several test scenario approaches to strengthen the research position. Using the notion of proximity, the cosine similarity algorithm finds instances of plagiarism or document similarity between thesis and final project papers. The Bidirectional Encoder Representations from Transformers (BERT) method and the additional scheme of applying cosine similarity will be integrated, and the outcomes will be further examined and compared. The selection of bidirectional encoder representations from transformers (BERT) and cosine similarity algorithms is predicated on their prowess in determining the degree of document similarity.

II. Material and Method

A. State of the Art

This research discusses the detection of the similarity of students' abstracts submitted with the thesis title. The algorithm used in this detection is cosine similarity and weighting using TF-IDF. To strengthen the position of the research, the following are some previous researches that apply identical algorithms or similar cases. Prior research can show approaches for detecting similarity. Moreover, each approach has its advantages and disadvantages. For example, the corpus method is more optimal for cross-language, and the semantic approach has good results but needs more resources. In contrast, the graph structure method needs to rely on learning good graph representation to perform well.

Previous research suggested that the semantic approach produces the most optimal result [10]. Research on document similarity detection in NLP uses a semantic approach: understanding and extracting meaning from text using computational techniques. The results of this study suggest that various methods exist to obtain embeddings from text, which are then used to detect similarities in bag-of-words-based documents such as Word2vec and GloVe, TF-IDF, Word Mover spacing, and Smooth Inverse Frequency. Moreover, the novel model outperformed all other models and, therefore, can be used to capture semantic information from input text effectively [11]. Research on the semantic approach for short text detection is still rare due to its limited application, and to perform this semantic detection, corpus-based, knowledge-based, and DL-based detections can be conducted [12].

Measuring word similarity with a semantic approach becomes a solution to finding optimal results. Previous research succeeded in improving detection results with Discourse Representation Structure, which shows more optimal results [13]. The results of the experimental evaluation confirmed that the proposed model improved the performance of the textual semantic similarity measure compared to the sentence embedding model, achieving an accuracy of 88.35% [14]. The research on checking final project documents using the TF-IDF algorithm could calculate the level of similarity of the final project's title submitted by students of the Padang State Polytechnic. The test results show that the process of calculating the degree of title similarity can be done quickly [15]. Similar results were also found to check ransomware detectioon the Confusion Matrix, with 94% accuracy [16]. Moreover, a previous study showed that students used cosine similarity and the Nazief-Adriani algorithm for stemming. It suggested the choice of words, considered as keywords in the answer key, greatly affected the results of the system assessment and obtained a cosine law of 89.5% [17].

The document similarity detection used the Goods and Price Planning Information System (SiPaGa) application for the codification search process. This study produces cosine similarity and TF-IDF weighting calculations and is expected to be applied to the SiPaGa application for more accurate results. Cosine similarity and TF-IDF algorithms are expected to improve the accuracy of product codification searches. Therefore, OPD can choose the product code as desired [18]. Moreover, previous research shows that product searches in e-commerce apply TF-IDF and Word2Vec, as well as cosine similarity, to calculate similarities between objects. From the study above, it can be concluded that Word2Vec takes more time to process data than using TF-IDF, making TF-IDF more efficient in terms of time [19]. The similarity of news articles among several sites can be measured using the cosine similarity algorithm by translating Hindi news articles into English and then comparing them with English news articles. Furthermore, cosine similarity, Jaccard similarity, and Euclidean distance were measured to calculate the news similarity score. The results can be used for effective and efficient identification [20]. Meanwhile, another study checking the similarities on news portals used cosine similarity and TF-IDF on the Microsoft News portal and obtained an accuracy of 80.77% [21].

Similarity detection has long been used in information search and machine learning domains for multi-purpose text mining, while this study is carried out in combination with clustering techniques. This study compares many methods for measuring document similarity, resulting in the conclusion that PDSM, Cosine, and Jaccard were superior to Euclidean, Manhattan, and Kulback–Leibler [22]. Text Classification has received significant attention recently, including a centroid-based approach and Bayesian naïve multinomial classifier, support vector machine, and neural network. The classification algorithm involves cosine similarity and weighting of the IDF TF, with the results being much more optimal [23]. The research from Hanifi states that using Doc2vec and Cosine Similarity algorithms helps the integration process for the initial analysis phase of the inventive design, which takes and collects essential knowledge from scientific data. Applying these two algorithms can optimize the data collection time in the initial analysis phase of the inventive design process and can significantly improve the accuracy of the information collected [24].

On the other hand, cosine similarity can be combined with the bidirectional encoder representations from the Transformers (BERT) model of the Transformer architecture, which allows bidirectional processing to better understand the context of words in sentences [25]. The BERT model in Natural Language Processing can obtain curve, accuracy, sensitivity, and specificity results of 0.96, 0.89, 088, and 0.89, respectively [26]. BERT is also often used to solve the problem of computational complexity and a vast memory footprint [27]. Moreover, the reliability of the BERT model can optimize good prediction results to determine the score of computer security vulnerability level [28]. This BERT model emphasizes transformer models designed to process data with high capacity and effectiveness [29]. BERT ability is also good at maximizing data with estimated data on unlabeled datasets [30] and can improve performance in word similarity detection [31]. In addition, the approach with BERT can also improve results and optimize results in word embeddings . Applying transformer-based architecture and proper sampling techniques significantly improves the performance of BERT [32]. Experimental results with BERT indicate that proposed model improves the performance of the baselines on 24 NLP tasks [33]. Deep learning models based on a combination of BERT with Bidirectional Long ShortTerm Memory (BiLSTM) and Bidirectional Gated Recurrent Unit (BiGRU) algorithms have a good result [34].

Previous research suggested that the most optimal method or approach is to use semantic analysis. However, it requires resources that tend to be larger than others, and the results are more accurate. One of the algorithms in the semantic approach is cosine similarity with the TF-IDF weighting method. This gets optimal results, so it will be suitable for detecting the similarity of students' abstracts. Another reason for computing is that the algorithm is efficient because of the ease of interpretation and has a match for large-scale data. In this study, two test scenarios are carried out in the corpus data by comparing the performance of the results obtained using the cosine similarity algorithm in determining the level of similarity of abstract documents with the scenarios using Bidirectional Encoder Representations from Transformers (BERT) and without using BERT. This demonstrates the benefits of Bidirectional Encoder Representations from Transformers (BERT), which can raise the accuracy of document similarity detection.

B. *Methodology*

This study adopts and modifies the Cross-Industry Standard Process for Data Mining (Crisp-DM) method. This methodology is used as a Non-Proprietary Standard Methodology for data mining [35].The methodology flow is shown in Figure 1.

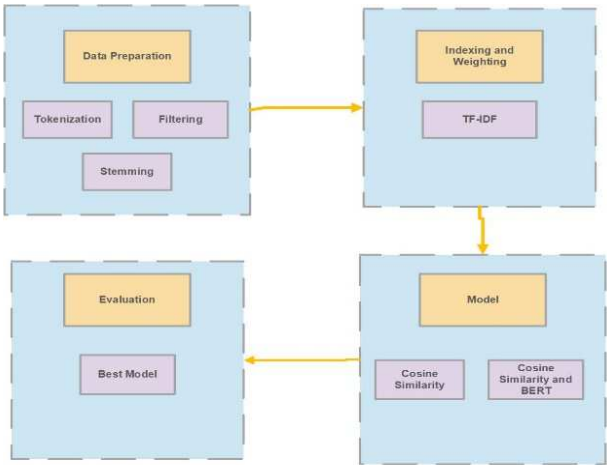


Fig. 1 Methodology

The detailed methodology flow is discussed in the following section.

III. Result and Discussion

Data from student abstract documents at a university within the Faculty of Computer Science is utilized. Abstract data from the last two years will be used for testing and training. This data was taken from 4 departments in the faculty of computer science. The flow of the methodology is explained in detail as follows:

A. *Data Preparation*

The description of the corpus data used is shown in Table I and Table II.

TABLE I
RAW DATA

Corpus Document
Perkembangan ilmu pengetahuan dan teknologi di indonesia mengalami kemajuan yang sangat pesat kemajuan yang paling dirasakan kehidupan masyarakat saat ini salah satunya adalah kemajuan di bidang teknologi informasi dan komunikasi ilmu pengetahuan dan teknologi secara umum memiliki keterkaitan yang erat khususnya dalam bidang pendidikan islam salah satunya pesantren pesantren merupakan lembaga pendidikan berbasis agama islam yang dikembangkan secara pribumi oleh masyarakat website pondok pesantren asshiddiqiyah 2 tangerang mempunyai banyak kekurangan kekurangan tersebut menjadi kendala website website masih bersifat semi online perlu adanya pembaharuan fitur seperti pendaftaran peserta didik baru secara full online tujuan dari penelitian ini adalah untuk mengetahui pengukuran tingkat kapabilitas website pondok pesantren asshiddiqiyah 2 tangerang dengan framework cobit khususnya pada bidang pesantren dengan aspek domain dss dan mea metode penelitian ini adalah metode kuantitatif dengan menyebarkan kuisioner kepada kurang lebih 94 responden pengguna serta nilai capability level dan gap analysis dari framework cobit hasil penelitian ini menunjukkan bahwa uji validitas dengan tingkat signifikansi 0202 dan uji reliabilitas dengan nilai cronbachs alpha sebesar 0905 masing-masing dinyatakan valid dan reliabel kemudian rata-rata hasil capability level pada subdomain dss01 dss03 dss05 mea01 dan mea02 adalah sebesar 14 level ini berada pada level 1 performed process dari level target yang ingin dicapai pada sisi it website pondok pesantren asshiddiqiyah 2 tangerang yaitu level 4 predictable process

TABLE II
RAW DATA 2

Corpus Document
saat ini daerah dapat mengatur sendiri rumah tangganya, oleh karena itu daerah diberikan kewenangan untuk menggali potensi sumber penerimaan yang ada dimana salah satunya berasal dari sektor pajak daerah, salah satu yang ingin dioptimalkan adalah pendapatan asli daerah (pad) dari jenis pajak reklame. saat ini di kota depok masih banyak reklame yang tidak memiliki izin maupun tidak diperpanjang izinnya tentu hal ini dapat mengurangi pad kota depok. oleh karena itu, saat ini diperlukan sistem yang dapat memonitoring reklame di kota depok. dalam sistem yang dirancang ini penulis melakukan pembahasan masalah dengan menggunakan metode pieces dan pengembangan sistem menggunakan air terjun yang diharapkan berbasis client server dengan arsitektur 3-tier. harapan penulis proses monitoring reklame dengan menggunakan web yang menerapkan jaringan vps dapat mempermudah para pemohon maupun staf dalam menyelesaikan pekerjaannya.

The data obtained is shown in Table I and Table II following stages were carried out:

1) Tokenization: This process describes the description initially as a sentence into a word. This method works well for breaking words up into tokens, which facilitates word identification, as shown in Table III and Table IV.

TABLE III
TOKENIZATION

Corpus Document
"perkembangan" "ilmu" "pengetahuan" "dan" "teknologi" "di" "Indonesia" "mengalami" "kemajuan" "yang" "sangat" "pesat" "kemajuan" "yang" "paling" "dirasakan" "kehidupan" "masyarakat" "saat" "ini" "salah" "satunya" "adalah" "kemajuan" "di" "bidang" "teknologi" "informasi" "dan" "komunikasi" "ilmu" "pengetahuan" "dan" "teknologi" "secara" "umum" "memiliki" "keterkaitan" "yang" "erat" "khususnya" "dalam" "bidang" "pendidikan" "islam" "salah" "satunya" "pesantren" "pesantren" "merupakan" "Lembaga" "pendidikan" "berbasis" "agama" "islam" "yang" "dikembangkan" "secara" "pribumi" "oleh" "masyarakat" "website" "pondok" "pesantren" "asshiddiqiyah" "2" "tangerang" "mempunyai" "banyak" "kekurangan" "kekurangan" "tersebut" "menjadi" "kendala" "website" "website" "masih" "bersifat" "semi" "online" "perlu" "adanya" "pembaharuan" "fitur" "seperti" "pendaftaran" "peserta" "didik" "baru" "secara" "full" "online" "tujuan" "dari" "penelitian" "ini" "adalah" "untuk" "mengetahui" "pengukuran" "tingkat" "kapabilitas" "website" "pondok" "pesantren" "asshiddiqiyah" "2" "Tangerang" "dengan" "framework" "cobit" "khususnya" "pada" "bidang" "pesantren" "dengan" "aspek" "domain" "dss" "dan" "mea" "metode" "penelitian" "ini" "adalah" "metode" "kuantitatif" "dengan" "menyebarkan" "kuisioner" "kepada" "kurang" "lebih" "94" "responden" "pengguna" "serta" "nilai" "capability" "level" "dan" "gap" "analysis" "dari" "framework" "cobit" "hasil" "penelitian" "ini" "menunjukkan" "bahwa" "uji" "validitas" "dengan" "tingkat" "signifikansi" "0202" "dan" "uji" "reliabilitas" "dengan" "nilai" "cronbachs" "alpha" "sebesar" "0905" "masing" "masing" "dinyatakan" "valid" "dan" "reliabel" "kemudian" "rata" "rata" "hasil" "capability" "level" "pada" "subdomain" "dss01" "dss03" "dss05" "mea01" "dan" "mea02" "adalah" "sebesar" "14" "level" "ini" "berada" "pada" "level" "1" "performed" "process" "dari" "level" "target" "yang" "ingin" "dicapai" "pada" "sisi" "it" "website" "pondok" "pesantren" "asshiddiqiyah" "2" "Tangerang" "yaitu" "level" "4" "predictable" "process"

TABLE IV
TOKENIZATION 2

Corpus Document
"saat" "ini" "daerah" "dapat" "mengatur" "sendiri" "rumah" "tangganya" "oleh" "karena" "itu" "daerah" "diberikan", "kewenangan" "untuk" "menggali" "potensi" "sumber" "penerimaan" "yang" "ada" "dimana" "salah" "satunya" "berasal" "dari" "sektor" "pajak" "daerah" "salah" "satu" "yang" "ingin" "dioptimalkan" "adalah" "pendapatan" "asli" "daerah" "(pad)" "dari" "jenis" "pajak" "reklame" "saat" "ini" "di" "kota" "depok" "masih" "banya" "reklame" "yang" "tidak" "memiliki" "izin" "maupun" "tidak" "diperpanjang" "izinnya" "tentu" "hal" "ini" "dapat" "mengurangi" "pad" "kota" "depok" "oleh" "karena" "itu" "saat" "ini" "diperlukan" "sistem" "yang" "dapat" "memonitoring" "reklame" "di" "kota" "depok" "dalam" "sistem" "yang" "dirancang" "ini" "penulis" "melakukan" "pembahasan" "masalah" "dengan" "menggunakan" "metode" "pieces" "dan" "pengembangan" "sistem" "menggunakan" "air" "terjun" "yang" "diharapkan" "berbasis" "client" "server" "dengan" "arsitektur" "3-tier" "harapan" "penulis" "proses" "monitoring" "reklame" "dengan" "menggunakan" "web" "yang" "menerapkan" "jaringan" "vps" "dapat" "mempermudah" "para" "pemohon" "maupun" "staf" "dalam" "menyelesaikan" "pekerjaannya"

2) Filtering: This stage filters irrelevant words or stop words. This will affect the overall results of the analysis because its function is to minimize the use of words that have less impact and will affect the overall results of the analysis, as in Table V and Table VI.

TABLE V
FILTERING

Corpus Document
ini, rata, bagi, di, dan, adanya, masih, untuk, dari, kemudian, masing, satunya, paling, dalam, jadi, yaitu, oleh, punya

TABLE VI
FILTERING 2

Corpus Document
ini, oleh, itu, yang, pad, maupun, dan, para, maupun, di, dalam, saat, adalah, dalam

3) Stemming: The stemming stage transforms words into their basic form. The main goal is to reduce the variation in a word's representation. The stem results are shown in Table VII and Table VIII.

TABLE VII
STEMMING

Corpus Document
kembang ilmu tahu teknologi indonesia alami maju pesat maju rasa hidup masyarakat saat satu maju bidang teknologi informasi komunikasi ilmu tahu teknologi cara umum milik kait erat khusus bidang didik islam satu pesantren rupa lembaga didik basis agama islam kembang cara pribumi masyarakat website pondok pesantren asshiddiqiyah 2 tangerang banyak kurang sebut kendala website website masih sifat semi online perlu baharu fitur seperti daftar serta didik baru cara full online tuju teliti adalah untuk tahu ukur tingkat kapabilitas website pondok pesantren asshiddiqiyah 2 tangerang framework cobit khusus bidang pesantren aspek domain dss mea metode teliti metode kuantitatif sebar kuisioner kurang lebih 94 responden guna serta nilai capability level gap analysis framework cobit hasil teliti tunjuk uji validitas tingkat signifikansi 0202 uji reliabilitas nilai cronbachs alpha besar 0905 masing nyata valid reliabel rata hasil capability level subdomain dss01 dss03 dss05 mea01 mea02 besar 14 level level 1 performed process level target capai sisi it website pondok pesantren asshiddiqiyah 2 tangerang yaitu level 4 predictable process

TABLE VIII
STEMMING 2

Corpus Document
saat ini daerah dapat atur sendiri rumah tangga oleh karena itu daerah beri wenang untuk gali potensi sumber terima yang ada mana salah satu asal dari sektor pajak daerah salah satu yang ingin optimal adalah dapat asli daerah pad dari jenis pajak reklame saat ini di kota depok masih banyak reklame yang tidak milik izin maupun tidak panjang izin tentu hal ini dapat kurang pad kota depok oleh karena itu saat ini perlu sistem yang dapat memonitoring reklame di kota depok dalam sistem yang rancang ini tulis laku bahas masalah dengan guna metode pieces dan kembang sistem guna air terjun yang harap bas client server dengan arsitektur 3-tier harap tulis proses monitoring reklame dengan guna web yang terap jaring vps dapat mudah para mohon maupun staf dalam selesai kerja

B. Indexing and Weighting

TF-IDF technique is used in text mining and natural language processing to evaluate how important a word is in a document relative to a set of documents (corpus). Some of the results of the TF-IDF calculations are shown in Table IX.

TABLE IX
TF-IDF

Corpus	Inverse Document Frequency
(0.317)	0.0559
(0.459)	0.0300
(0.381)	0.0559
(0.76)	0.0559
(0.158)	0.0370
(0.412)	0.0559
(0.322)	0.1119
(0.301)	0.0559
(0.4)	0.0559
(0.317)	0.0559

C. Model

The model used compares cosine similarity and optimization process using Bidirectional Encoder Representations from Transformers (BERT).

D. Evaluation

The model comparison results from the two scenarios will be sought for optimal results and used as an alternative to the best model obtained. The corpus data will be created and tested with a two-schematic approach, namely detection using the cosine similarity algorithm and schema using cosine similarity along with Bidirectional Encoder Representations from Transformers (BERT) using data testing. These results will be compared with the similar scores between documents with two different scenarios tested.

1) Scenario 1: Cosine Similarity: The results of data testing are shown in Table X.

TABLE X
RESULT IN COSINE SIMILARITY

Corpus	Result
Corpus 1	0.1396
Corpus 2	0.0998
Corpus 3	0.1641
Corpus 4	0.2532
Corpus 5	0.2081
Corpus 6	0.1167
Corpus 7	0.1663
Corpus 8	0.1108
Corpus 9	0.2119
Corpus 10	0.3202

2) Scenario 2: Cosine Similarity with Bidirectional Encoder Representations from Transformers (BERT): Combining Cosine Similarity with Bidirectional Encoder Representations from Transformers (BERT) will be optimized with the rich vocabulary embedded into the Bidirectional Encoder Representations from Transformers (BERT) model. The working steps in Bidirectional Encoder Representations from Transformers (BERT) are as follows:

Step 1: Large amounts of training data
BERT is specially designed to work on larger word counts. So, with a more diverse vocabulary than the built-in BERT model combined with the data corpus in the study, the model will have more vocabulary and variety.

Step 2: Masked Language Model
Masked Language Model (MLM) enables bidirectional learning from text. Because word analysis is carried out in two directions, the word identification process becomes more profound, and the possibilities and approaches to word meaning can be captured more broadly.

After these 2 steps, what is done is to apply fine-tuning to the BERT model with the following steps:

Step 1: Get the dataset.
The dataset used in this study is corpus data from abstracts obtained in the repository. This data will be reference data combined with reliable problem-solving or translation skills from Bidirectional Encoder Representations from Transformers (BERT).

Step 2: Start exploring.
This stage begins with data labeling, which is carried out by identifying and learning cosine similarity and bidirectional encoder representations from transformers (BERT).

Step 3: Data monitoring.
This stage of the Bidirectional Encoder Representations from Transformers (BERT) will request processing from the Central Processing Unit (CPU) so that it can execute the assigned command.

Step 4: Processing.
This process repeats the pre-processing stage. The Remodel performs processes such as tokenizing to ensure that the return results are more optimal.

Step 5: Design the final input pipeline.
The training and testing data are fed into the Bidirectional Encoder Representations from Transformers (BERT) architecture pipeline to be processed to identify word/sentence similarity.

Step 6: BERT classification model
Bidirectional Encoder Representations from Transformers (BERT) will identify words and sentences to be represented in the form of numbers similar to the corpus data you already have.

Step 7: Updating and saving.
At this stage, the Bidirectional Encoder Representations from Transformers (BERT) will store the model's results for identifying and detecting word similarities.

An overview of the trial scenario using the additional Bidirectional Encoder Representations from Transformers (BERT) model is shown in Table XI.

TABLE XI
RESULT IN COSINE SIMILARITY AND BERT

Corpus	Result
Corpus 1	0.4439
Corpus 2	0.3128
Corpus 3	0.4403
Corpus 4	0.4010
Corpus 5	0.6097
Corpus 6	0.3548
Corpus 7	0.3567
Corpus 8	0.3894
Corpus 9	0.3720
Corpus 10	0.4589

E. Comparison

Based on the results of the data testing carried out with 2 schemes, namely cosine similarity and cosine similarity with

Bidirectional Encoder Representations from Transformers (BERT), the comparison can be visualized in Figure 2.

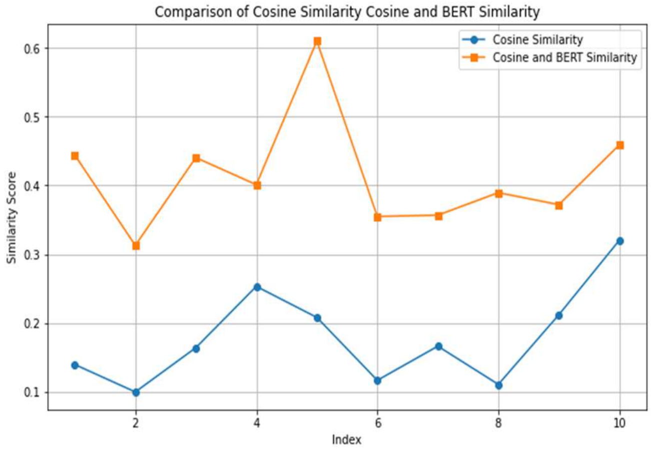


Fig. 2 Comparison Result

The detailed results of the tests carried out are shown in Table XII.

TABLE XII
DETAILED COMPARISON RESULT

Corpus	Cosine	Cosine Bert
Corpus 1	0.1396	0.4439
Corpus 2	0.0998	0.3128
Corpus 3	0.1641	0.4403
Corpus 4	0.2532	0.4010
Corpus 5	0.2081	0.6097
Corpus 6	0.1167	0.3548
Corpus 7	0.1663	0.3567
Corpus 8	0.1108	0.3894
Corpus 9	0.2119	0.3720
Corpus 10	0.3202	0.4589

A comparison of the average value of document similarity in 10 test data with cosine similarity and cosine similarity with BERT can be seen in Figure 3.

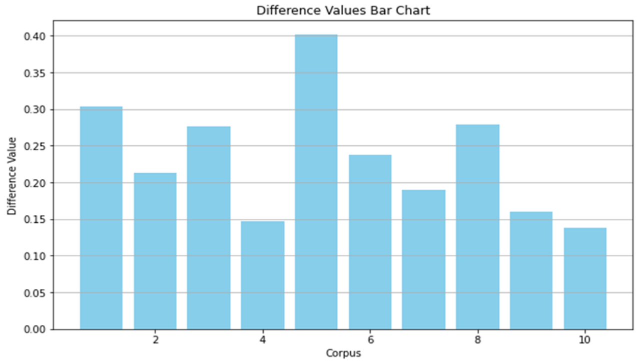


Fig. 3 Gap Cosine and BERT

The test findings indicate that cosine similarity with the inclusion of the BERT method has a more excellent similarity value when compared to the document similarity value or similarity generated by cosine similarity in 10 test data. A 23.48% increase in performance indicates an overall improvement. These findings demonstrate the superiority of using Bidirectional Encoder Representations from Transformers (BERT) over not using BERT. This is because BERT produces representative vectors from a text token, improving detection results and emphasizing sentence context for a broader range of complex words and sentences.

IV. Conclusion

The results showed that combining the Bidirectional Encoder Representations from Transformers (BERT) method in Cosine Similarity could increase the similarity detection value of students' thesis abstract by 23.48% compared to the cosine similarity algorithm. This can be one of the optimization techniques for improving the performance of the cosine similarity algorithm in detecting document similarity. The Bidirectional Encoder Representations from Transformers (BERT) method has a more complex range of understanding words or sentences. For future research, it is suggested that vocabulary and sentences be enriched to provide an overview of BERT's ability to detect similarities. The fact that students must first submit the title of their final project and abstract for the research to compare it to prior documents that have been saved in the repository offers promise for additional innovation in the prevention of plagiarism.

Acknowledgment

We thank the Institute for Research and Community Service (LPPM) at Universitas Pembangunan Nasional "Veteran" Jakarta for funding assistance for this research on the schema RISCOP and to every colleague who helped in this research.

References

[1] K. Keuangan, "Sistem Pendidikan Nasional," JDIH Kemenkeu. [Online]. Available: https://jdih.kemenkeu.go.id/fulltext/1989/2tahun~1989uupenj.htm

[2] Pemerintah Indonesia, "Undang-Undang Nomor 4 Tahun 2014 Tentang Penyelenggaraan Pendidikan Tinggi dan Pengelolaan Perguruan Tinggi," *Standar Nasional Pendidikan*, p. 37, 2014.

[3] V. Chandere, S. Satish, and R. Lakshminarayanan, "Online plagiarism detection tools in the digital age: A review," *Ann Rom Soc Cell Biol*, vol. 25, no. 1, pp. 7110–7119, 2021.

[4] K. P. dan Kebudayaan, *Peraturan Menteri Pendidikan Nasional Republik Indonesia Nomor 17 Tahun 2010*, vol. 4, no. 4. 2010, pp. 1921–1926.

[5] K. Ristek, *Peraturan Menteri Pendidikan, Kebudayaan Riset dan Teknologi Tentang Integritas Akademik dalam Menghasilkan Karya Ilmiah*, vol. 3, no. 2. 2021, p. 6.

[6] A. Kleebayoon and V. Wiwanitkit, "Artificial Intelligence, Chatbots, Plagiarism and Basic Honesty: Comment," *Cell Mol Bioeng*, vol. 16, no. 2, pp. 173–174, Apr. 2023, doi: 10.1007/s12195-023-00759-x.

[7] D. Chandrasekaran and V. Mago, "Evolution of Semantic Similarity—A Survey," *ACM Comput Surv*, vol. 54, no. 2, pp. 1–37, Mar. 2022, doi: 10.1145/3440755.

[8] H. Steck, C. Ekanadham, and N. Kallus, "Is Cosine-Similarity of Embeddings Really About Similarity?," in *Companion Proceedings of the ACM Web Conference 2024*, New York, NY, USA: ACM, May 2024, pp. 887–890. doi: 10.1145/3589335.3651526.

[9] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695, Sep. 2021, doi: 10.1007/s12525-021-00475-2.

[10] J. Wang and Y. Dong, "Measurement of text similarity: A survey," *Information (Switzerland)*, vol. 11, no. 9, pp. 1–17, 2020, doi:10.3390/info11090421.

[11] S. Agarwala, A. Anagawadi, and R. M. Reddy Guddeti, "Detecting Semantic Similarity of Documents Using Natural Language Processing," *Procedia CIRP*, vol. 189, pp. 128–135, 2021, doi:10.1016/j.procs.2021.05.076.

[12] M. Han, X. Zhang, X. Yuan, J. Jiang, W. Yun, and C. Gao, "A survey on the techniques, applications, and performance of short text semantic

similarity," *Concurr Comput*, vol. 33, no. 5, pp. 1–17, 2021, doi:10.1002/cpe.5971.

[13] M. Farouk, "Measuring text similarity based on structure and word embedding," *Cogn Syst Res*, vol. 63, pp. 1–10, 2020, doi:10.1016/j.cogsys.2020.04.002.

[14] S. Das, N. Deb, A. Cortesi, and N. Chaki, "Sentence embedding models for similarity detection of software requirements," *SN Comput Sci*, 2021, doi: 10.1007/s42979-020-00427-1.

[15] D. Meidelfi, - Yulherniwati, I. Rahmayuni, T. Hidayat, and D. Chandra, "TF-IDF Implementation for Similarity Checker on The Final Project Title," *International Journal of Advanced Science Computing and Engineering*, vol. 3, no. 1, pp. 40–52, 2021, doi:10.30630/ijasce.3.1.3.

[16] M. Argene, C. Ravenscroft, and I. Kingswell, "Ransomware Detection via Cosine Similarity-Based Machine Learning on Bytecode Representations," *Authorea*, Aug. 2024, doi:10.22541/au.172348750.00074165/v1.

[17] R. R. Et.al, "The Similarity of Essay Examination Results using Preprocessing Text Mining with Cosine Similarity and Nazief-Adriani Algorithms," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 3, pp. 1415–1422, 2021, doi:10.17762/turcomat.v12i3.938.

[18] S. Sintia, S. Defit, and G. W. Nurcahyo, "Product Codefication Accuracy With Cosine Similarity and Weighted Term Frequency and Inverse Document Frequency (Tf-Idf)," *Journal of Applied Engineering and Technological Science*, vol. 2, no. 2, pp. 14–21, 2021, doi: 10.37385/jaets.v2i2.210.

[19] H. Kusniyati and A. A. Nugraha, "Analysis of Matric Product Matching Between Cosine Similarity with Term Frequency-Inverse Document Frequency (TF-IDF) and Word2Vec in PT. Pricebook Digital Indonesia," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 6, no. 1, pp. 105–112, 2020, doi: 10.32628/cseit195672.

[20] R. Singh and S. Singh, "Text Similarity Measures in News Articles by Vector Space Model Using NLP," *Journal of The Institution of Engineers (India): Series B*, vol. 102, no. 2, pp. 329–338, 2021, doi:10.1007/s40031-020-00501-5.

[21] G. Yunanda, D. Nurjanah, and S. Meliana, "Recommendation System from Microsoft News Data using TF-IDF and Cosine Similarity Methods," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 1, pp. 277–284, 2022, doi: 10.47065/bits.v4i1.1670.

[22] A. A. Amer and H. I. Abdalla, "A set theory based similarity measure for text clustering and classification," *J Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00344-3.

[23] K. Park, J. S. Hong, and W. Kim, "A Methodology Combining Cosine Similarity with Classifier for Text Classification," *Applied Artificial Intelligence*, vol. 34, no. 5, pp. 396–411, 2020, doi:10.1080/08839514.2020.1723868.

[24] M. Hanifi, H. Chibane, R. Houssin, and D. Cavallucci, "Problem formulation in inventive design using Doc2vec and Cosine Similarity as Artificial Intelligence methods and Scientific Papers," *Eng Appl Artif Intell*, vol. 109, 2022, doi: 10.1016/j.engappai.2022.104661.

[25] S. Ravichandiran, *Getting Started with Google BERT*. Packt Publishing, 2021.

[26] I. Soldevilla and N. Flores, "Natural Language Processing through BERT for Identifying Gender-Based Violence Messages on Social Media," in *2021 IEEE International Conference on Information Communication and Software Engineering (ICICSE)*, IEEE, Mar. 2021, pp. 204–208. doi: 10.1109/icicse52190.2021.9404127.

[27] Z. Liu, G. Li, and J. Cheng, "Hardware Acceleration of Fully Quantized BERT for Efficient Natural Language Processing," in *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, IEEE, Feb. 2021, pp. 513–516. doi:10.23919/date51398.2021.9474043.

[28] M. R. Shahid and H. Debar, "CVSS-BERT: Explainable Natural Language Processing to Determine the Severity of a Computer Security Vulnerability from its Description," in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, Dec. 2021, pp. 1600–1607. doi:10.1109/icmla52953.2021.00256.

[29] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.

[30] D. Grießhaber, J. Maucher, and N. T. Vu, "Fine-tuning BERT for Low-Resource Natural Language Understanding via Active Learning," in *Proceedings of the 28th International Conference on Computational Linguistics*, Stroudsburg, PA, USA: International Committee on Computational Linguistics, 2020, pp. 1158–1171. doi:10.18653/v1/2020.coling-main.100.

[31] N. Peinelt, D. Nguyen, and M. Liakata, "tBERT: Topic Models and BERT Joining Forces for Semantic Similarity Detection," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 7047–7055. doi:10.18653/v1/2020.acl-main.630.

[32] J. Tracz, P. Wójcik, K. Jasinska-Kobus, R. Belluzzo, R. Mroczkowski and I. Gawlik, "BERT-based similarity learning for product matching", *Proceedings of Workshop on Natural Language Processing in E-Commerce*, pp. 66-75, 2020.

[33] K. A. Mazhar, M. Brodtbeck, and G. Gühring, "Similarity learning of product descriptions and images using multimodal neural networks," *Natural Language Processing Journal*, vol. 4, p. 100029, Sep. 2023, doi: 10.1016/j.nlp.2023.100029.

[34] A. S. Talaat, "Sentiment analysis classification system using hybrid BERT models," *J Big Data*, vol. 10, no. 1, p. 110, Jun. 2023, doi:10.1186/s40537-023-00781-w.

[35] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Comput Sci*, vol. 181, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.