

Word embedding for contextual similarity using cosine similarity

Yessy Asri¹, Dwina Kuswardani², Amanda Atika Sari², Atikah Rifdah Ansyari¹

¹Faculty of Energy and Telematics, Study Program Information Systems, Institute Technology of PLN, Jakarta, Indonesia

²Faculty of Energy and Telematics, Study Program Informatics Engineering, Institute Technology of PLN, Jakarta, Indonesia

Article Info

Article history:

Received Mar 13, 2024

Revised Nov 5, 2024

Accepted Nov 11, 2024

Keywords:

Augmented data

Contextual similarity

Cosine similarity

IndoBERT

Word embedding

ABSTRACT

Perspectives on technology often have similarities in certain contexts, such as information systems and informatics engineering. The source of opinion data comes from the Quora application, with a retrieval limit of the last 5 years. This research aims to implement Indo-bidirectional encoder representations from transformers (BERT), a variant of the BERT model optimized for Indonesian language, in the context of information system (IS) and information technology (IT) topic classification with 414 original data, which, after being augmented using the synonym replacement method, The generated data becomes 828. Data augmentation aims to evaluate the performance of models by using synonyms and rearranging text while maintaining meaning and structure. The approach used is to label the opinion text based on the cosine similarity calculation of the embedding token from the IndoBERT model. Then, the IndoBERT model is applied to classify the reviews. The experimental results show that the approach of using IndoBERT to classify SI and IT topics based on contextual similarity achieves 90% accuracy based on the confusion matrix. These positive results show the great potential of using transformer-based language models, such as IndoBERT, to support the analysis of comments and related topics in Indonesian.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Yessy Asri

Faculty of Energy and Telematics, Study Program Information Systems, Institute Technology PLN

Jakarta 11750, Indonesia

Email: yessyasri@itpln.ac.id

1. INTRODUCTION

A topic in the form of an opinion becomes a reference in determining certain interests, goals, or other things. However, it often seems complicated and confusing when topics have different categories in the same field. Especially in the technology category, which has similarities between several parts, such as the topic of informatics engineering and the topic of information systems, both topics have many enthusiasts, but from enthusiasts such as prospective students who have an interest, some of them experience confusion in understanding the differences and approaches or similarities between the topics of informatics engineering and information systems. So, they tend to have various questions covering aspects of both topics by wanting an opinion or opinion on a platform as a source of information.

In text data, it does not only have one or two sentences but can form a long paragraph. A field becomes the topic of a series of sentences. However, despite having the same field, there will be differences in categories [1]. This can be analyzed according to the context. A number of documents are given term frequency-inverse document frequency (TF-IDF) and cosine similarity model calculations to see their similarity [2]. In this scope, context is the main concern, so a model is needed that can see the entire context

of sentence construction. Bidirectional encoder representations from transformers (BERT) is one of the techniques that can be used in this process. BERT models have proven effective in performing language processing tasks such as semantic text [3] and syntax [4]. In terms of language, IndoBERT is a BERT development that can be used for Indonesian data. In addition to IndoBERT indicators, data augmentation techniques are also performed with the aim of expanding the data to have certain variations. Merging is performed, followed by sequence randomization, before dividing the text into two columns for text similarity calculation using cosine similarity by utilizing word embedding from IndoBERT. A simple text similarity label is 0 meaning not similar and 1 meaning similar [5].

In the era of digitalization, there are a number of platforms as information-sourcing tools. Quora is a question-and-answer app founded by Adam D'Angelo and a colleague. This application was initially only available in the English version; until April 2018, Quora was present in the Indonesian version (<https://www.idntimes.com/tech/trend/mahda-lena/keunggulan-aplikasi-quora-c1c2>). Reporting from statistics by the director of product management, Quora is a website and application-based platform with an achievement of 300 million visitors in creating a set of questions from various topics posted by users and answers from other users who have greater insight into a topic [6], [7]. However, prospective students often have difficulty filtering and analyzing Quora user opinions that are relevant to their needs. In addition, sometimes workers in the marketing field have a little hassle when identifying and filtering the information obtained, such as information about the most relevant question topics with informative opinions related to the difference between informatics engineering and information systems, so that later the data can be used as a reference in promotional materials and other marketing matters. Therefore, an effective method is needed to analyze the opinions of Quora users to help in reading topics in the field of technology, especially between the two topics, namely informatics engineering and information systems. One technique that can be used is a Transformers-based natural language processing (NLP) technique, such as BERT [8]. Such models have proven to be very effective in performing language processing tasks such as capturing syntax [4] and sentence-semantic text [3]. On a language basis, BERT has been developed using the IndoBERT model, which is used for Indonesian language data.

Classification by implementing the IndoBERT model in embedding Indobenchmark or IndoBERT has an accuracy level of 87% with online article content data in research in the building of informatics, technology, and science (BITS) journal, "Clickbait Classification Model on Online News with Semantic Similarity Calculation Between News Title and Content" [1]. In this study, the authors performed text similarity analysis techniques using the IndoBERT model to classify SI and IT topics in contextual similarity-based opinions. The difference in this study lies in the labeling used; in the journal, no labeling is done, while in this study, labeling is done using cosine calculations by utilizing embedding tokens in contextual calculations. Another difference lies in the measurement of similarity being limited to the title and content of one news story, not the whole news story, and using semantics, while this research does a shuffle to calculate the similarity of sentences randomly in context so that the accuracy of the model can be seen.

2. RELATED RESEARCH

IndoBERT was developed and trained specifically for the Indonesian language so that it can provide more accurate results in analyzing text in Indonesian. In a study entitled "clickbait classification model on online news with semantic similarity calculation between news title and content" [1], in the classification technique performed, it was found that IndoBERT had an accuracy rate in semantic similarity of 87%. In a study entitled "identification of text similarity using class indexing based and cosine similarity for complaint document classification" [5], using text similarity identification techniques using class indexing-based and cosine similarity methods to classify complaint documents, the accuracy of the research was 84.12%. The next research, entitled "Indonesian news classification using IndoBERT" [9], conducts news recommendations based on recommendations by comparing the IndoBERT model with XGB. The highest accuracy when implementing IndoBERT is 94.5%.

From several studies that become references for this research, the author uses the IndoBERT model in measuring text similarity with the cosine method on contextual similarity using the Python programming language. It is known that the implementation of the IndoBERT model in analyzing Indonesian-based data has high accuracy when fine-tuning. In addition, cosine similarity, as a vector calculation for clustering before performing semantic similarity, is used to measure text similarity in language structure and context. There are similarities from several previous studies, namely using the IndoBERT model for classification. The difference lies in the use of contextual similarity, which produces labeling from the results of cosine calculations in text similarity analysis. In this research leads to text similarity analysis for new data containing opinions with two topics, namely the topic of informatics engineering and the topic of information systems, by utilizing the augmentation process to expand the dataset which then shuffles the data to produce a random sequence as a reference that the data is not the same as the results of augmentation when separating

into two columns with the column names text 1 and text 2, cosine similarity for the process of calculating text similarity with the aim of labeling which utilizes embedding tokens from IndoBERT, implementation of the IndoBERT model in topic classification, and binaryclass confusion matrix as model testing.

Based on [8], semantic labeling is the process of mapping attributes in data sources to ontology classes as an important step when integrating heterogeneous data. In the research “Semantic Labeling: A Domain-Independent Approach,” similarity metrics are approached as a comparison feature for labeled domain data. It is explained that in semantic labeling, attribute values have an important role in identification with the same semantic type. The similarity approach carried out in the study has different metrics, including Jaccard similarity as a modification for numerical values and TF-IDF for textual data.

Cosine similarity is a common method to perform data similarity, as in the journal “improving patient clustering by incorporating structured label relationships in similarity measures” [9], which uses cosine similarity to classify patient similarity. The use of IndoBERT is done when the available data uses Indonesian [10]. This is because IndoBERT is specially trained for Indonesian, as shown in Figure 1 [11].

Model	#Params	#Layers	#Heads	Emb. Size	Hidden Size	FFN Size	Language Type	Pre-train Emb. Type
Scratch	15.1M	6	10	300	300	3072	Mono	-
fastText-cc-id	15.1M	6	10	300	300	3072	Mono	Word Emb.
fastText-indo4b	15.1M	6	10	300	300	3072	Mono	Word Emb.
IndoBERT-lite _{BASE}	11.7M	12	12	128	768	3072	Mono	Contextual
IndoBERT _{BASE}	124.5M	12	12	768	768	3072	Mono	Contextual
IndoBERT-lite _{LARGE}	17.7M	24	16	128	1024	4096	Mono	Contextual
IndoBERT _{LARGE}	335.2M	24	16	1024	1024	4096	Mono	Contextual
mBERT	167.4M	12	12	768	768	3072	Multi	Contextual
XL _M -R _{BASE}	278.7M	12	12	768	768	3072	Multi	Contextual
XL _M -R _{LARGE}	561.0M	24	16	1024	1024	4096	Multi	Contextual
XL _M -ML _M _{LARGE}	573.2M	16	16	1280	1280	5120	Multi	Contextual

Figure 1. IndoNLU benchmark

Figure 1 shows the type of indobenchmark. The parameters in the research adjust the model type and data size. In this research, labeling is done with the text similarity method, namely cosine similarity, which utilizes word embedding from the indoBERT model and then performs classification based on contextual similarity [12], [13] using the model. Previously, the data will be expanded using the data augmentation method, namely synonym replacement, to perform variations so as to emphasize the model providing word embedding and cosine similarity, calculating similarity in data that has been varied.

2.1. Text similarity

Text similarity is the measurement of text similarity, which is the basis of NLP tasks. Text similarity is defined as the similarity between two texts. Not only that, text similarity also considers a broader context perspective in analyzing the semantic properties of two words [14]. The method of measuring text similarity involves two aspects, including:

2.1.1. Text distance

There are three ways of measuring text distance based on length, distribution, and semantic objects, one of which is cosine distance. The cosine measurement measures the cosine angle between the two texts. Judging from the cosine of 0° being 1 and the cosine of 90° being 0, the similarity value lies in the numbers -1 to 1, where the cosine measure is related to orientation. As osin the following formula [14]:

$$\cos\theta = \frac{a \cdot b}{\|a\| \cdot \|b\|} = \frac{\sum_{i=1}^n a_i * b_i}{\sqrt{\sum_{i=1}^n (a_i)^2} * \sqrt{\sum_{i=1}^n (b_i)^2}}$$

2.1.2. Text representation

Text representation performs calculations directly as numerical features that are similar in lexical and semantic ways. Lexical similarity is done through different measurements, while semantic similarity is introduced through string-based, corpus-based, semantic text matching, and graph structure-based

methods [14]. Text similarity research often uses the cosine similarity formula because it provides intuitive interpretations and values, which range between -1 and 1. The formula is scale-consistent, where values close to 1 are interpreted as a high degree of similarity, while values close to -1 indicate dissimilarity. In the context of text similarity research on document data, cosine similarity also shows robustness to dimensional differences, demonstrating the flexible nature that makes it commonly used.

Cosine similarity can be considered a text similarity calculation technique in the framework of text representation, with a focus on the category of semantic text matching to assess the similarity between text and documents. In addressing the complexity of sentence meaning and vector representations that take into account inter-word and contextual relationship patterns, the BERT model is a relevant choice. The combination of cosine similarity calculation with the use of embedding tokens from the IndoBERT model is chosen as a method for contextual similarity-based labeling, considering the complexity of data arising from the relationship between sentences to form paragraphs.

2.2. Bidirectional encoder representations from transformers (BERT)

BERT is the latest NLP algorithm developed by Google. It was first introduced by Google AI researchers in 2018. BERT utilizes the transformer model in learning contextual relationships between words in a text, where the transformer has two mechanisms, namely encoder and decoder. However, BERT only requires an encoder. BERT uses a bidirectional approach and performs sequential reading of text inputs, allowing the model to learn the context of words based on the surrounding words. In the encoder input, the sequence of tokens will be embedded into a vector, which will then be passed on to the neural network and output vector and generated according to the input [15]. Figure 2 shows of BERT architecture.

Figure 2 shows BERT utilizes the Transformer architecture to learn contextual relationships between words in a text. Transformers have two mechanisms: an encoder and a decoder. However, BERT only has an encoder mechanism that takes a bidirectional approach and reads text input sequentially, allowing the model to learn context based on surrounding words. BERT has BERT-base with as many as 12 encoder layers, 768 hidden nodes, 12 attention heads, and about 110,000,000 parameters, and BERT-large with as many as 24 encoder layers, 1024 hidden nodes, 16 attention heads, and about 340,000,000 parameters [3].

Token embedding in the context of BERT refers to the numerical vector representation of a token generated by a BERT model. BERT is one of the transformer architectures that has proven to be very effective in natural language understanding tasks, such as question understanding, language translation, and other tasks. Figure 3 shows of embedding token BERT [16]. Some detailed points about token embedding in BERT include tokenization, embedding layers, position embeddings, segment embeddings, fine-tuning, and bidirectional context. Embedding tokens in BERT provides a rich and contextual representation for each word or subword in the text, allowing the model to better cope with natural language understanding tasks.

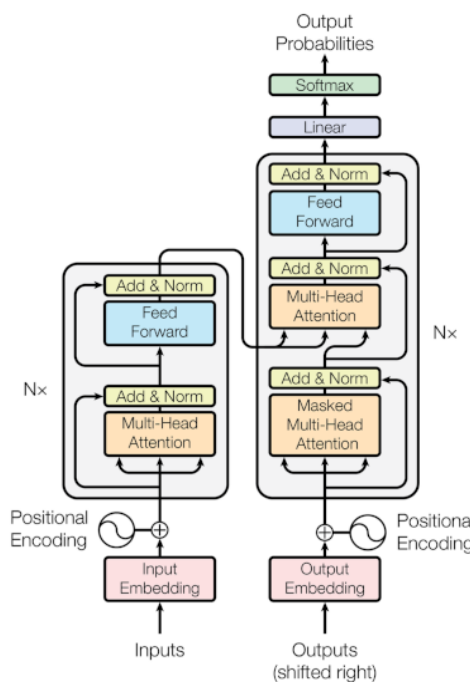


Figure 2. BERT architecture

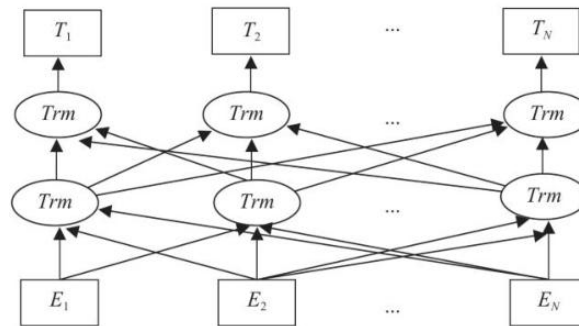


Figure 3. Embedding token BERT

3. METHOD

Figure 4 illustrates the workflow methodology for embedding word for contextual similarity using cosine similarity. In the first stage, information is searched on opinion data on the topic of informatics engineering and information systems on the Quora platform. Data retrieval in this study uses a sampling technique called simple random sampling, where data is taken from the two required topics, namely informatics engineering and information systems, and from both topics will have a number of data from the two topics in the same data so that it only has one opinion column that will be used in the next stage of data preparation. Furthermore, the data preparation process includes web scraping, data pre-processing, data labeling, and data splitting for modeling used in the research. In one of step, there have step for themself. Figure 5 shows the data preparation workflow and Figure 6 shows about data preprocessing workflow[17]. Figure 6 shows the pre-processing data workflow[18]. After obtaining opinion data from the two topics that will be the object of research, then, data pre-processing includes case folding, deleted unique characters, slang words, and tokenization by adding a step to remove numbers that are no longer needed.

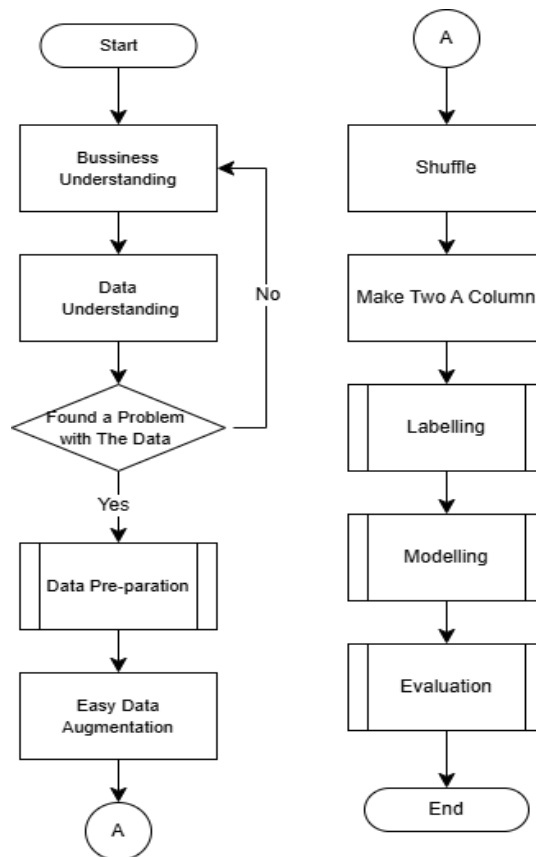


Figure 4. The research workflow

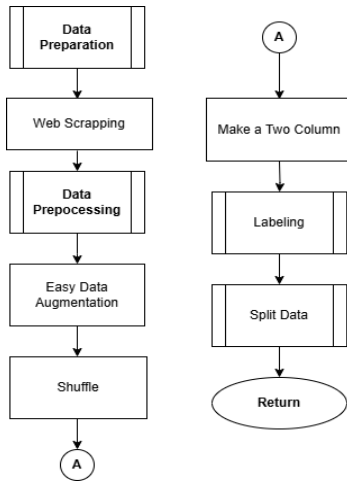


Figure 5. Data preparation workflow

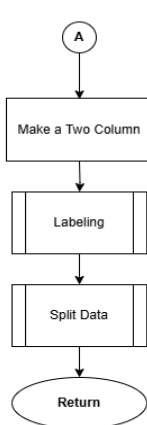
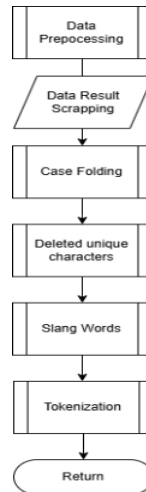


Figure 6. The pre-processing data workflow



4. RESULTS AND DISCUSSION

Data augmentation is a collection of algorithms that create synthetic data by making small changes to existing data, aiming to expand the amount of training data in deep neural network learning [14]–[16]. This technique is useful for observing model failures and improving their performance. Data augmentation is an important step in model training, helping to overcome the limitations of limited data. It is also considered a cost-efficient way to increase data size, reduce training errors, and produce more accurate predictions. Many machine learning projects rely on data augmentation as their critical success factor [18].

Therefore, a data augmentation method is performed to perform sentence variation using synonym replacement. In doing so, a random word will be searched and changed into another word that has the same meaning. However, the process does not involve words that are categorized as stopwords. In the changed word, it is possible for the embedding word to have a different number value from the actual word. The results will be combined with the data before the changes are made, and to add to the improvement, randomization of the order will be carried out so that the division of the two columns will not have the same sentence, although it is possible to have similar positions in randomization.

In the amount of 414 data, it is still relatively small for the use of a series of deep learning processes; therefore, data augmentation is carried out using the synonym replacement technique, which takes one word at random and replaces it with the same sentence where the sentence is in the equation contained in WordNet. The process can be seen in Figure 7, with the result in Figure 8.

```

@staticmethod
def validate(**kwargs):
    if 'p' in kwargs:
        if kwargs['p'] > 1 or kwargs['p'] < 0:
            raise TypeError("p must be a fraction between 0 and 1")
    if 'sentence' in kwargs:
        if not isinstance(kwargs['sentence'].strip(), str) or len(kwargs['sentence'].strip()) == 0:
            raise TypeError("sentence must be a valid sentence")
    if 'n' in kwargs:
        if not isinstance(kwargs['n'], int):
            raise TypeError("n must be a valid integer")

def __init__(self, stop_words=None, random_state=None):
    self.stopwords = stop_words.words('indonesian') if stop_words is None else stop_words
    self.sentence = None
    self.p = None
    self.n = None
    self.random_state = random_state
    if isinstance(self.random_state, int):
        random.seed(self.random_state)
    else:
        raise TypeError("random_state must have type int")

def add_word(self, new_words):
    """Insert word"""
    synonyms = list()
    counter = 0
    while len(synonyms) < 1:
        random_word_list = list([word for word in new_words if word not in self.stopwords])
        random_word = random_word_list[random.randint(0, len(random_word_list) - 1)]
        synonyms = self._get_synonyms(random_word)
        counter += 1
        if counter >= 10:
            return new_words # See Issue 14 for details
    random_synonym = synonyms[0] # TODO
    random_idx = random.randint(0, len(new_words) - 1)
    new_words.insert(random_idx, random_synonym)
    return new_words

def synonym_replacement_top_n(self, sentence: str,

```

Figure 7. Source code of synonym replacment

	kalimat	hasil_replacement
0	beberapa hal yang general pakai version contro...	beberapa hal yang general pakai version contro...
1	lulusan yang mendominasi di gojek tokopedia te...	lulusan yang mencengkeram di gojek tokopedia t...
2	yang paling mendasar adalah pemahaman arsitek...	yang paling mendasar adalah pemahaman arsitek...
3	salah satu iklan sebuah sekolah formal setingk...	salah satu iklan sebuah sekolah formal setingk...
4	teman saya lulusan diploma tiga kampus swasta ...	teman saya lulusan diploma tiga kampus swasta ...

Figure 8. Result of segmentation data

After augmenting the data so that it has 828 sentences where the data is indirectly sequential when performing the merge command, randomization is performed on the data to avoid similarities between sentences from the original data and sentences from the augmented data that will be calculated. Figure 9 is the result of the data shuffle. Before performing calculations using IndoBERT on measuring text similarity with cosine, the data is divided into two columns, namely text 1 and text 2, which can be seen in Figure 10.



Figure 9. Result of shuffle data

	teks 1	teks 2
0	ptn teknik informatika terbaik tapi persaingan...	ooba buka jawaban saya tentang tujuan tersulit...
1	kalau anda tidak mengorek sama sekali tentang ...	sebelum memilih jurusan pastikan anda lebih su...
2	nanti bisa sekarang juga bisa belajar itu inve...	berikut adalah contoh program dalam bahasa c u...
3	kalau kamu anak ipa sistem informasi kalau kam...	business analyst seiring dengan meningkatnya k...
4	prospek kerja saya yakin bagus sistem maklum...	berdasarkan preferensi pribadi dan teman teman s...

Figure 10. Division results into two columns

The next step is data labeling based on cosine similarity calculations. Cosine similarity provides labeling by calculating the similarity of opinion data on the information engineering label and the information systems label after calculating the similarity between the two text columns. Then the labeling uses the average value as a threshold to produce balanced data, because if you use a threshold based on the highest or lowest value, it will result in a total of one label having hundreds more data than other labels, which only have tens of thousands of data points. The labeling states that 0 is not similar, while 1 is similar. Figure 11 shows of the labelling process.

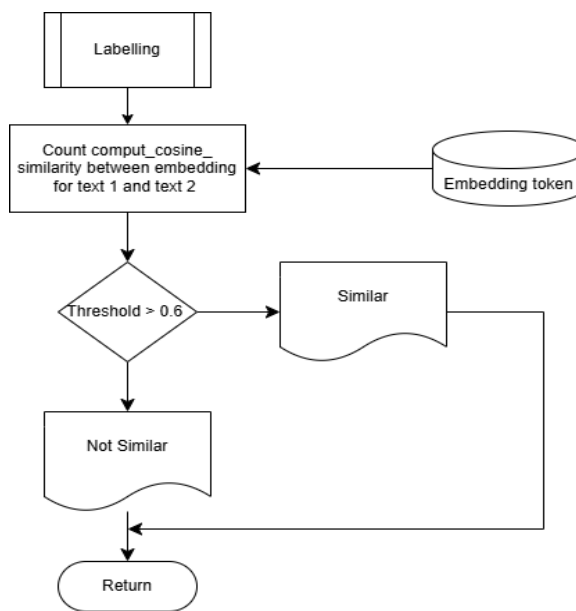


Figure 11. Labelling process

Data labeling is based on cosine similarity calculations, which can be seen in Figures 12 and 13 for the results of searching for the highest, lowest, and average values. Then the labeling uses the average value as a threshold to produce balanced data, because if you use a threshold based on the highest or lowest value, it will result in a total of one label having hundreds more data than other labels, which only have tens of thousands of data points. The labeling states that 0 is not similar, while 1 is similar. The labeling results can be seen in Figure 14.

	teks 1	teks 2	Teks 1 Embeddings	Teks 2 Embeddings	Cosine Similarity
0	ptn teknik informatika terbaik tapi persaingan...	coba buka jawaban saya tentang tujuan tersulit...	[3.25543843e-02 8.32005880e-01 -2.2012803e-...	[0.8770394 0.82026577 -0.2311910 0.135448...	0.8092
1	kalau anda tidak mengorek sama sekali tentang ...	sebelum memilih jurusan pastikan anda lebih su...	[-2.33479142e-01 1.85190784e-01 -1.04837723e-...	[2.15938830e-01 4.33998840e-01 -1.37478322e-...	0.6815
2	nanti bisa sekarang juga bisa belajar itu inve...	berikut adalah contoh program dalam bahasa o u...	[1.03754497e+00 1.50273514e+00 2.59330503e-...	[-6.90191209e-01 4.39789957e-01 2.3936697e-...	0.2197
3	kalau kamu anak ipa sistem informasi kalau kam...	business analyst seiring dengan meningkatnya k...	[8.50444734e-01 9.31510448e-01 2.88731303e-...	[-2.40344258e-01 8.47591777e-01 -8.91180110e-...	0.5011
4	prospek kerja saya yakin bagus sistem maklum...	berdasarkan prefensi pribadi dan teman teman s...	[3.89403843e-01 8.25589120e-01 3.59340101e-...	[5.8208921e-01 1.09259212e+00 -2.05144719e-...	0.7223

Figure 12. Cosine similarity result

```
Highest Cosine Similarity Score: 0.9056
Lowest Cosine Similarity Score: 0.2197
Average Cosine Similarity Score: 0.6006386473429952
```

Figure 13. Search for highest, lowest, and average values of cosine values

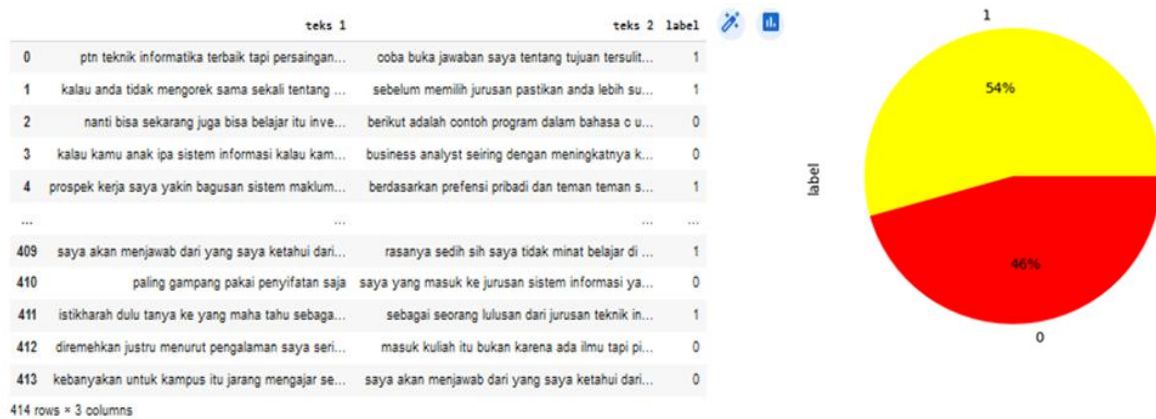


Figure 14. Data labeling results

After cosine similarity measurement utilizing IndoBERT word embedding as labeling, the IndoBERT model is used to perform contextual similarity-based classification. In this study, the data sharing index was 80:10:10, with training data being 80%, while validation data and testing data were 10% each. The flow of the IndoBERT model requires training data, and validation data will be at the BertTokenizer model stage by having various parameters that support the process, including max_length 128, batch_size 32, epoch 5, and learning rate 1e-5. Model size and the number of parameters give an indication of the complexity and memory requirements of a model. Models with a large number of parameters tend to require more computational resources and memory. This information is useful for evaluating whether the model is suitable for use on resource-constrained devices. By understanding the size and complexity of the model, decisions regarding the implementation and use of the model can be better considered, especially in the context of resource constraints on the device.

The next stage is fine-tuning, adjusting the feature representation that already exists in BERT [19]–[21] according to the characteristics of the dataset. After going through the model training stage using training data and validation data, the next step is to test the model with testing data. When making predictions on testing data, the model will produce a percentage probability of confidence related to the predictions made. The higher the probability given by the model, the more confident the model is in its prediction results. This process provides insight into the model’s level of confidence in the prediction results

given on data that was not used during the training and validation processes. There is a probability column for the percentage of confidence that the model places the text in that category. The confusion matrix is a technique that is often used in the classification of model results based on stimulated objects [22], [23]. The confusion matrix performed is a type of binary class confusion matrix. Figure 15 is an image of the placement of the binary class confusion matrix [6], [16], [22], [24], [25], and Figure 16 is the result of the confusion matrix from the research. The evaluation of the model in this research uses a confusion matrix, which will calculate accuracy values in the text similarity analysis for the classification that has been carried out. From the confusion matrix, the accuracy value of IndoBERT is 90%.

TP	FN
FP	TN

Figure 15. Confusion matrix

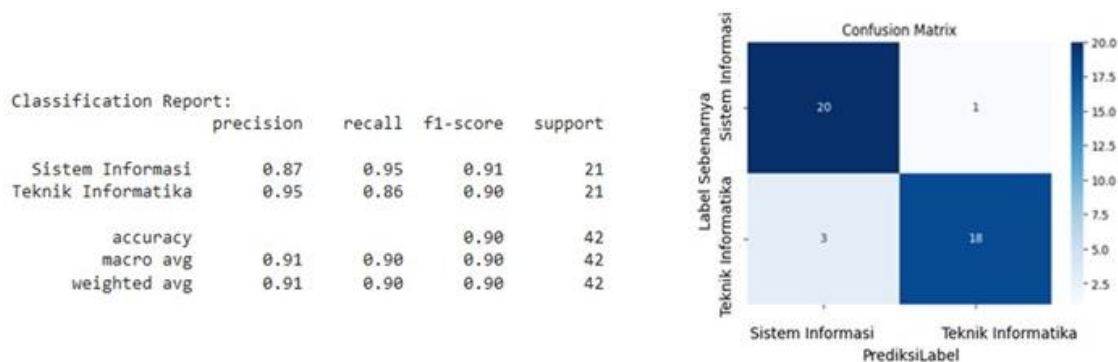


Figure 16. Confusion matrix result

5. CONCLUSION

Based on the results of the research and discussion, it can be concluded that the data similarity analysis process consists of data collection, data pre-processing, data augmentation, calculating the similarity of text in the data, labeling the data according to the average value of the results of the cosine similarity calculation, which will become the threshold. The calculation of text similarity becomes labeling with two labels, namely label 0 is not similar while label 1 is a similar label. In the IndoBERT model, label 0 is represented as an IS topic classification and label 1 as an IT label classification. In this research, it was concluded that the IndoBERT model had calculation accuracy results using a confusion matrix worth 90% in calculations and understanding of contextual text because it had been trained specifically for data in Indonesian. In order to perfect the resulting text similarity analysis, there are several suggestions, including adding Google Translate to the programming before doing slang words for several words in sentences that use foreign languages so that it can improve context calculations, and it is hoped that different Indobenchmarks will be used, such as IndoBERTLite or IndoBERTLarge in type indobenchmark-p1 or type indobenchmark-p2.

ACKNOWLEDGEMENTS

Our deepest gratitude goes to the Institute for Community Service Research (LPPM) of the Institute Technology of PLN for funding the community service activities for fiscal year 2023.




REFERENCES

- [1] H. A. Ahmadi and A. Chowanda, "Clickbait classification model on online news with semantic similarity calculation between news title and content," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 4, Mar. 2023, doi: 10.47065/bits.v4i4.3030.




- [2] L. Sahu and B. R. Mohan, "An improved K-means algorithm using modified cosine distance measure for document clustering using Mahout with Hadoop," *2014 9th International Conference on Industrial and Information Systems (ICIIS)*, Gwalior, India, 2014, pp. 1-5, doi: 10.1109/ICIINFS.2014.7036661.
- [3] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: pre-training of deep bidirectional transformers for language understanding," *Naacl-Hlt 2019*, pp. 4171–4186, 2018.
- [4] G. Jawahar, B. Sagot, and D. Seddah, "What does BERT learn about the structure of language?," in *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2020, pp. 3651–3657, doi: 10.18653/v1/p19-1356.
- [5] J. Ye, "Improved cosine similarity measures of simplified neutrosophic sets for medical diagnoses," *Artificial Intelligence in Medicine*, vol. 63, no. 3, pp. 171–179, 2015, doi: 10.1016/j.artmed.2014.12.007.
- [6] Z. Imtiaz, M. Umer, M. Ahmad, S. Ullah, G. S. Choi, and A. Mehmood, "Duplicate questions pair detection using siamese MaLSTM," *IEEE Access*, vol. 8, pp. 21932–21942, 2020, doi: 10.1109/ACCESS.2020.2969041.
- [7] E. Uzun, "A novel web scraping approach using the additional information obtained from web pages," in *IEEE Access*, vol. 8, pp. 61726–61740, 2020, doi: 10.1109/ACCESS.2020.2984503.
- [8] A. Vaswani *et al.*, "Attention is all you need," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017, pp. 5999–6009.
- [9] B. Juarto, "International journal of intelligent systems and applications in engineering Indonesian news classification using IndoBERT," *Original Research Paper International Journal of Intelligent Systems and Applications in Engineering IJISAE*, vol. 2023, no. 2, 2023.
- [10] L. Wu *et al.*, "Word mover's embedding: from word2vec to document embedding," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, 2018, pp. 4524–4534, doi: 10.18653/v1/d18-1482.
- [11] M. Pham, S. Alse, C. A. Knoblock, and P. Szekely, *Semantic labeling: A domain-independent approach*, vol. 9981. Cham: Springer International Publishing, 2016.
- [12] J. Lambert, A.-L. Leutenegger, A. Baudot, and A.-S. Jannot, "Improving patient clustering by incorporating structured label relationships in similarity measures," *medRxiv*, 2023.
- [13] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: a benchmark dataset and pre-trained language model for Indonesian NLP," in *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, 2020, pp. 757–770, doi: 10.18653/v1/2020.coling-main.66.
- [14] 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.
- [15] D. Rothman, "Transformers for natural language processing: Build, train, and fine-tuning deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, and GPT-3," *Packt Publishing*, 2022.
- [16] B. Wilie *et al.*, "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," 2020.
- [17] U. Shafique and H. Qaiser, "A comparative study of data mining process models (KDD, CRISP-DM and SEMMA)," *International Journal of Innovation and Scientific Research*, vol. 12, no. 1, pp. 217–222, 2014.
- [18] J. M. Wu, Y. Belinkov, H. Sajjad, N. Durrani, F. Dalvi, and J. Glass, "Similarity analysis of contextual word representation models," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4638–4655, doi: 10.18653/v1/2020.acl-main.422.
- [19] C. Biemann and M. Riedl, "Text: Now in 2D! a framework for lexical expansion with contextual similarity," *Journal of Language Modelling*, vol. 1, no. 1, 2013, doi: 10.15398/jlm.v1i1.60.
- [20] C. Shorten, T. M. Khoshgoftar, and B. Furht, "Text data augmentation for deep learning," *Journal of Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00492-0.
- [21] D. Haba, "Data augmentation with python enhance deep learning accuracy with data augmentation methods for image, text, audio, and tabular data," 2023.
- [22] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," in *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2019, pp. 6382–6388, doi: 10.18653/v1/d19-1670.
- [23] M. Bayer, M. A. Kaufhold, B. Buchhold, M. Keller, J. Dallmeyer, and C. Reuter, "Data augmentation in natural language processing: A novel text generation approach for long and short text classifiers," *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 1, pp. 135–150, 2023, doi: 10.1007/s13042-022-01553-3.
- [24] M. Hasnain *et al.*, "Evaluating trust prediction and confusion matrix measures for web services ranking," *IEEE Access*, vol. 8, pp. 90847–90861, 2020.
- [25] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition*, vol. 91, pp. 216–231, 2019, doi: 10.1016/j.patcog.2019.02.023.

BIOGRAPHIES OF AUTHORS






Yessy Asri    she was born on October 13, 1976, in Padang city. She is the first of three children, born to father Asri Ma'arif and mother Edinar. Starting education in the S1 Department of Informatics Engineering, then continuing her Masters education in the Department of Information Systems, and currently, she is continuing his doctoral education in the Department of Information Technology. Currently, she serves as Head of the Information Systems Study Program at the PLN Institute of Technology (ITPLN) in Jakarta and is also a lecturer at several well-known universities in Jakarta. She is also currently active as the Chief Editor of the ITPLN KILAT Journal, is active as an instructor at several training institutions, and is also currently an active BNSP assessor. With perseverance and high motivation to continue learning and trying, until now the author has produced many writings in several SINTA-accredited national journals and indexed international journals on intellectual property rights (HaKI), carried out many community service activities (PkM), and received several awards. She can be contacted at email: yessyasri@itpln.ac.id.






Dwina Kuswardani    she was born in Surabaya on July 25, 1962. She graduated S1 in the Computational Mathematics Study Program, Faculty of Mathematics and Natural Sciences, University of Indonesia, in 1988. She graduated S2 in the Master of Computer Science Study Program, Faculty of Computer Science at the University of Indonesia in 2009. She graduated S3 in the Doctoral Study Program, Faculty of Computer Science, University of Indonesia, in 2016. Currently, she is a permanent lecturer in the Informatics Engineering Study Program, Faculty of Energy Telematics, Institut Teknologi PLN (ITPLN). She teaches courses in data mining, machine learning, and image processing. She can be contacted at email: dwina@itpln.ac.id.



Amanda Atika Sari    she was born in Pasuruan city. On December 10, 2001, she was the first child of two children of the late Mr. Cahya Anang Santoso and Mrs. Lasmiyati. She started her education at the S1 Department of Informatics Engineering. Currently, she is a freelancer in the field of arts and entertainment and is actively following the community entertainment and is actively participating in the community. She has a high level of motivation to keep trying and learning about new things. She also has experience in web design during her internship. includes the biography here. She can be contacted at email: amanda1931182@itpln.ac.id.



Atikah Rifdah Ansyari    she was born in Barru city. On May 06, 2005, she was the last child of four children of the late Mr. Muhammad Kasang and Mrs. Syarifah. She started her education at Information Systems with awardee scholarship Aperti BUMN. Currently, she is a freelancer in marketing and admission, actively as a prompt engineer and telemarketer. She always expands experience and knowledge by innovating, being creative, having high motivation with altruist quotient, high ambition, initiative, and keen to seek new challenges. She can be contacted at email: atikah2232001@itpln.ac.id.