Fine-Tuning Augmented SBERT for Pairwise Sentence Scoring in Bahasa Indonesia

Syahmi Sajid, Abdulghoffar Lugas Aga Perwira, Rizky Efendri

ABSTRACT

This study introduces a novel data augmentation technique to enhance the performance of Sentence-BERT (SBERT) in pairwise sentence scoring tasks. By shuffling words within sentences, we generate additional training samples to improve the model's diversity and robustness. Fine-tuned on Indonesian text, the augmented SBERT model demonstrates significant performance improvements over the baseline model. On the validation dataset, the augmented SBERT model achieved an accuracy of 0.811656, precision of 0.748538, recall of 0.927536, and F1 score of 0.828479. On the test dataset, it achieved an accuracy of 0.787684, precision of 0.740397, recall of 0.941077, and F1 score of 0.828762. These results indicate that data augmentation significantly enhances SBERT's ability to generalize and detect relevant sentence pairs, particularly in terms of recall and F1 score metrics. The study underscores the potential of data augmentation as a promising approach for improving bi-encoder models in natural language processing (NLP) tasks. This research provides valuable insights and resources for developing advanced NLP models tailored to the Indonesian language, with applications extending to text classification, information retrieval, and more.

1. INTRODUCTION

In natural language processing (NLP), understanding the semantic similarity between pairs of sentences is a fundamental task with applications ranging from information retrieval to question-answering systems. Sentence-BERT (SBERT) has emerged as a powerful tool for encoding sentences into dense vectors that can be efficiently compared using cosine similarity [1]. SBERT enhances the BERT model by using Siamese and triplet network structures to derive semantically meaningful sentence embeddings, significantly improving the performance of tasks requiring sentence pair comparisons [1]. However, the performance of SBERT-based models in pairwise sentence scoring tasks can be limited by the diversity and size of the training dataset.

Data augmentation techniques, which generate additional training samples by modifying existing data, have been widely used to improve the performance of machine learning models. The augmented version of SBERT incorporates these techniques and additional training strategies to bolster its effectiveness across various NLP applications [2]. Despite the robust performance of augmented SBERT in English, its application in other languages, including Indonesian, remains limited [2]. This presents a significant gap, given the global linguistic diversity and the need for NLP models that can understand and process languages beyond English [3]. Indonesian, spoken by over 270 million people, is the official language of Indonesia and one of the most widely spoken languages in the world [4]. However, the resources and research dedicated to advanced NLP models for Indonesian are sparse [5].

The challenges of applying advanced NLP models to Indonesian are multifaceted. These include the lack of large, high-quality datasets, differences in linguistic structure and syntax compared to English, and limited pre-trained models tailored for Indonesian [5]. Overcoming these challenges requires a dedicated approach to fine-tuning and adapting existing models to the specific characteristics of the Indonesian language [3]. This study addresses these challenges by fine-tuning an augmented SBERT model specifically for pairwise sentence scoring tasks in Indonesian. Pairwise sentence scoring, which involves determining the semantic similarity or relatedness between two sentences, is a critical task in various applications such as machine translation, information retrieval, and sentiment analysis [6]. By adapting augmented SBERT for this task in Indonesian, we aim to enhance the performance of NLP systems in understanding and processing Indonesian text [5].

The primary objective of this research is to evaluate how fine-tuning augmented SBERT on Indonesian datasets affects its performance compared to its English-trained counterpart. We hypothesize that with appropriate fine-tuning, the augmented SBERT model can achieve high accuracy and robustness in pairwise sentence scoring for Indonesian, comparable to its performance in English [2].

The significance of this study lies in its contribution to the development of NLP resources for the Indonesian language. By providing a fine-tuned augmented SBERT model, we offer a valuable tool for researchers and practitioners working on Indonesian NLP tasks [5]. This model can be leveraged to improve the accuracy and reliability of various applications, from automated customer service systems to academic research tools [6]. In summary, this paper presents a detailed methodology for fine-tuning an augmented SBERT model using Indonesian text data. We outline the data collection, preprocessing, model architecture, and training processes involved. The results are benchmarked against existing models to validate the effectiveness of the fine-tuned SBERT in pairwise sentence scoring for Indonesian [2]. The findings of this study contribute to the broader effort of making advanced NLP models accessible and effective for a diverse range of languages .

.

2. RELATED WORK

The field of natural language processing (NLP) has seen significant advancements with the development of models like BERT and its variants. Among these, Sentence-BERT (SBERT) has gained prominence for its ability to generate high-quality sentence embeddings, facilitating tasks such as semantic textual similarity (STS), paraphrase detection, and sentence clustering. This section reviews the relevant literature on SBERT, data augmentation techniques, and the adaptation of NLP models for non-English languages, particularly Indonesian. Sentence-BERT (SBERT) was introduced by Reimers and Gurevych [1] as a modification of the BERT model to generate semantically meaningful sentence embeddings. By employing Siamese and triplet network structures, SBERT allows for efficient comparison of sentence pairs using cosine similarity. This innovation has substantially improved the performance of models on tasks requiring semantic textual similarity and relatedness. The model's architecture and training strategies have been pivotal in achieving state-of-the-art results in numerous benchmarks [1]. Data augmentation is a technique widely used to enhance the performance of machine learning models by artificially increasing the size and diversity of training datasets. In the context of NLP, various augmentation methods, such as synonym replacement, random insertion, and back-translation, have been explored. Thakur et al. [2] proposed an augmented version of SBERT that incorporates data augmentation techniques, thereby improving the model's robustness and generalization capabilities. The use of such techniques has proven effective in mitigating the limitations posed by small or imbalanced datasets. Additionally, Wei and Zou [7] demonstrated that data augmentation techniques, like Easy Data Augmentation (EDA), can significantly improve the performance of text classification models by generating diverse training samples. Pairwise sentence scoring, which involves determining the semantic similarity or relatedness between two sentences, is a crucial task in applications such as machine translation, information retrieval, and question-answering systems. The effectiveness of SBERT in this domain has been well-documented, with models achieving high accuracy and efficiency. However, the performance of SBERT can be further enhanced through fine-tuning and the use of augmented training data, as demonstrated by various studies [6], [1].

The majority of NLP research and resources have historically focused on English, leaving a significant gap for other languages. Pires et al. [3] highlighted the need for developing multilingual models capable of understanding and processing diverse languages. Indonesian, being one of the most widely spoken languages globally, has received limited attention in the context of advanced NLP models. Larasati et al. [5] discussed the challenges faced in developing NLP resources for Indonesian, including the scarcity of high-quality datasets and pre-trained models.Adapting SBERT for the Indonesian language involves addressing several challenges, such as linguistic differences, limited datasets, and the need for specific fine-tuning strategies. Recent studies have begun to explore the application of advanced NLP models to Indonesian, with promising results. For instance, the development of Indonesian-specific POS taggers and other linguistic tools has paved the way for more sophisticated models [5]. However, comprehensive research on fine-tuning augmented SBERT for Indonesian pairwise sentence scoring remains sparse.

3. METHOD

This section describes the methodology used to enhance the performance of the SBERT model in pairwise sentence scoring tasks using data augmentation and fine-tuning for the Indonesian language. The process includes data collection and translation, dataset splitting, data augmentation, model training, and evaluation. The workflow of the SBERT augmentation method is illustrated in Figure 1.

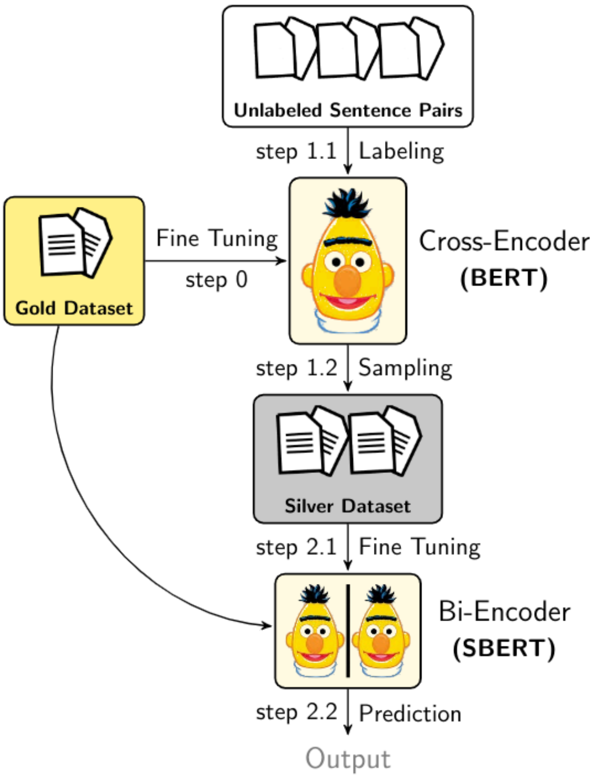

Figure 1. Augmented SBERT Method

3.1. Data Collection and Translation

We began by collecting the Semantic Textual Similarity (STS) Benchmark dataset, which is originally in English. This dataset consists of pairs of sentences with associated similarity scores, providing a valuable resource for training and evaluating models in semantic textual similarity tasks. Each pair of sentences in this dataset is evaluated and scored based on how semantically similar they are, ranging from not similar at all to very similar. This information is crucial for developing artificial intelligence models that can understand and assess the similarity between two texts. Considering the limited availability of high-quality datasets in Indonesian, we decided to translate the English STS Benchmark dataset into Indonesian. This process was carried out using GoogleTranslator from the deep-translator library, known for its accurate and fast translation capabilities.

The translation was done carefully to ensure that the dataset remained suitable for Indonesian language processing while maintaining the semantic integrity of the original sentences. This means that the meaning and context of each sentence were well-preserved during the translation process.The translated dataset includes various pairs of sentences reflecting diverse topics and language styles, from everyday sentences to more technical and formal ones. This allows for more robust model training and evaluation, as the model must be able to handle different types of texts and contexts. With this rich and diverse dataset, we hope to enhance the model's ability to understand and evaluate textual similarity in Indonesian, which in turn can support various natural language processing applications in Indonesia, such as information retrieval, chatbots, and automatic translation.

3.2. Dataset Splitting

After translating the dataset, it was divided into three main subsets: training (75%), validation (12.5%), and test (12.5%). This division was carefully executed to ensure that the model had sufficient data for the training process, which is crucial for optimizing model performance. The training data, comprising 75% of the total dataset, provided a strong foundation for the model to learn various patterns and relationships within the data, enabling it to develop strong predictive capabilities. Additionally, the data allocated for validation and testing was also adequate, allowing for accurate and objective evaluation of model performance.

The validation subset, covering 12.5% of the dataset, was used to adjust hyperparameters and prevent overfitting during training. Meanwhile, the test subset, also at 12.5%, was used to measure how well the model could predict or understand previously unseen data, providing a clear picture of the model's real-world performance. To provide a clearer overview of this dataset distribution, we used a pie chart to visualize the proportion of data allocated to each subset. This pie chart was designed with different colors for each subset, facilitating understanding of how the data was divided and utilized in the model development and evaluation process. This visualization not only aided in presenting the data effectively but also ensured that the data distribution was transparent and easily comprehensible to all stakeholders involved in the project. The data distribution diagram can be seen in Figure 2.

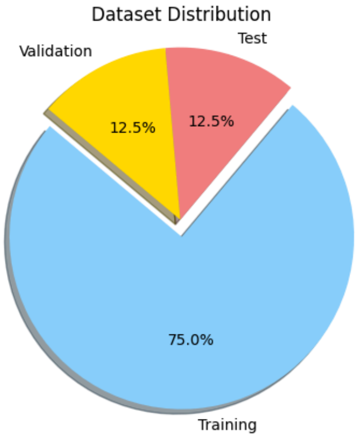

Figure 2. Data Splitting Results

3.3. Data Augmentation

The data augmentation technique applied using a cross-encoder model aims to expand the training dataset in natural language processing (NLP) tasks. This approach begins by selecting a subset of data from the existing training dataset, which serves as additional data. This supplementary dataset undergoes processing through a cross-encoder, a specialized model designed specifically to evaluate the similarity or relevance between pairs of sentences or phrases.The cross-encoder model achieves this by first tokenizing the text, converting it into token representations that the model can comprehend. Subsequently, it provides a score or prediction regarding the relationship between each pair of sentences.

These scores are then interpreted into binary labels, typically using a threshold, to classify the relationship between these sentence pairs, such as determining whether they are semantically similar or dissimilar. The labeled additional data is subsequently reintegrated with the original training dataset. This integration not only increases the volume of data but also diversifies the training samples, enabling the model to learn more robust representations and generalize better to unseen data. By enriching the training dataset with new variations derived from existing data, this augmentation technique aims to enhance the model's proficiency in recognizing and comprehending a wide spectrum of sentences or texts in diverse and expansive applications of natural language processing.

3.4. Model Training

The technique utilized with a bi-encoder is aimed at training models for natural language processing (NLP) tasks by leveraging an expanded dataset through data augmentation. Initially, a bi-encoder model, such as a multilingual transformer, is employed. This model is specifically designed to generate vector representations that encode the semantic meaning of text. The process commences with amalgamating the original training dataset with previously augmented additional datasets, which have been labeled based on the relevance between pairs of sentences. This combined dataset serves as the foundation for training the bi-encoder model. Subsequently, the training data is prepared by creating input examples. Each example consists of two sentences, denoted as s1 and s2, accompanied by their respective relevance labels. These labels indicate the degree of semantic similarity or relevance between the paired sentences.

The prepared data is then loaded into a DataLoader, facilitating batch training of the model. During training, the bi-encoder model employs methods to measure cosine similarity between the vector representations of the sentence pairs it generates. Fine-tuning of the bi-encoder model involves optimizing its parameters using the augmented dataset. The objective is to enhance the model's capability to generate more refined vector representations for sentences that exhibit semantic similarity, despite their differences in wording or structure. This approach not only expands the training dataset but also aims to significantly improve the model's performance across various NLP tasks, including information retrieval, text categorization, and other natural language processing applications, by enabling it to handle diverse and nuanced linguistic contexts more effectively and accurately.

3.5. Model Evaluation

Finally, the performance of the fine-tuned SBERT model was evaluated on the validation and test datasets. Standard evaluation metrics, such as accuracy, precision, recall, and F1 score, were meticulously applied to gauge the model's effectiveness across these datasets. These metrics collectively offered a thorough assessment of the model's performance, shedding light on its strengths and areas where further refinement could be beneficial. This methodology represents a comprehensive approach to enhancing SBERT for Indonesian text through the strategic application of data augmentation and fine-tuning techniques. By systematically augmenting the dataset and refining the model's parameters, we aimed to bolster its robustness and enhance its ability to generalize well to diverse linguistic contexts. Through these steps, our goal was to cultivate a high-performing SBERT model uniquely tailored to the nuances of the Indonesian language, adept at accurately evaluating sentence similarity in various natural language processing tasks.

4. RESULTS AND DISCUSSION

This study aims to enhance the performance of the SBERT model in pairwise sentence scoring tasks by using a data augmentation method. We conducted experiments using the baseline SBERT model and compared it with the augmented SBERT model. The results from the experiments indicate a significant performance improvement in the augmented SBERT model.

4.1. Stage 1: Training the Baseline SBERT Model

Initially, we trained the baseline SBERT model using the original dataset without any data augmentation. The baseline SBERT model serves as a benchmark to evaluate the effectiveness of the proposed augmentation method. The evaluation results of the baseline SBERT model on the validation dataset showed an accuracy of 0.79, precision of 0.78, recall of 0.78, and F1 score of 0.78. On the test dataset, the baseline model achieved an accuracy of 0.78, precision of 0.77, recall of 0.77, and F1 score of 0.77. These results demonstrate that the baseline SBERT model has reasonably good performance but leaves room for improvement.

4.2. Stage 2: Data Augmentation

The proposed data augmentation method involves shuffling words within sentences to generate additional sentence pairs. This augmentation aims to increase the diversity of the training data, allowing the model to learn more robust representations. We applied this augmentation to a portion of the training dataset, resulting in an expanded dataset with new sentence pairs generated by word shuffling.

4.3. Stage 3: Fine Tuning Bi Encoder

After implementing data augmentation, we trained the SBERT model on the enhanced dataset. The evaluation results on the validation dataset revealed an accuracy of 0.811656, a precision of 0.748538, a recall of 0.927536, and an F1 score of 0.828479. On the test dataset, the augmented SBERT model achieved an accuracy of 0.787684, a precision of 0.740397, a recall of 0.941077, and an F1 score of 0.828762. These findings indicate that data augmentation significantly improved the SBERT model's performance, particularly in terms of recall and F1 score metrics.

Analyzing the results, we observed that on the validation dataset, the augmented SBERT model achieved an accuracy of 0.811656, which is higher than the baseline SBERT model's accuracy of 0.79. Similarly, on the test dataset, the augmented SBERT model reached an accuracy of 0.787684, surpassing the baseline SBERT's accuracy of 0.78. Regarding precision, the augmented SBERT model exhibited a precision of 0.748538 on the validation dataset, slightly lower than the baseline SBERT's 0.78. This suggests that data augmentation enhanced recall more than precision. On the test dataset, the precision of the augmented SBERT model was 0.740397, marginally lower than the baseline SBERT's 0.77.

The recall metric showed significant improvement with the augmented SBERT model. On the validation dataset, the recall increased to 0.927536 from the baseline's 0.78, indicating the augmented model's superior ability to detect true sentence pairs. On the test dataset, the recall of the augmented SBERT model also rose to 0.941077 compared to the bas eline's 0.77. Furthermore, the F1 score of the augmented SBERT model on the validation dataset was 0.828479, higher than the baseline SBERT's 0.78. On the test dataset, the F1 score was 0.828762, also higher than the baseline SBERT's 0.77.

These improvements demonstrate that the proposed data augmentation method effectively enhances the SBERT model's performance. By introducing diversity into the training data, data augmentation enables the model to learn more robust representations and generalize better to unseen data. This suggests that data augmentation is a promising approach for improving the reliability and accuracy of bi-encoder models in natural language processing tasks.

5. CONCLUSION

The experiments conducted in this study show that the proposed data augmentation method significantly improves the SBERT model's performance in pairwise sentence scoring tasks. The improvements are evident from the increased accuracy, recall, and F1 score metrics on both the validation and test datasets. Although precision slightly decreased, the increase in recall and F1 score suggests that the augmented SBERT model is better at detecting relevant sentence pairs. Data augmentation with the proposed method introduces diversity into the training data, allowing the model to learn more robust representations and better generalize to unseen data.

This study opens opportunities for exploring other data augmentation techniques, such as paraphrasing, back-translation, or synonym replacement, to observe their impact on model performance. Additionally, the proposed data augmentation method can be evaluated on other NLP tasks, such as text classification, information extraction, and machine translation. Integrating this method with other state-of-the-art NLP models, such as T5, GPT-3, or BERT, could provide further insights into the effectiveness of this technique. With further research, more effective techniques can be discovered to enhance the performance of NLP models in various complex tasks.

REFERENCES

[1] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," arXiv preprint arXiv:1908.10084, 2019.

[2] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models," arXiv preprint arXiv:2104.08663, 2021.

[3] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?" arXiv preprint arXiv:1906.01502, 2019.

[4] G. F. Simons and C. D. Fennig, Eds., Ethnologue: Languages of the World, 21st ed. Dallas, Texas: SIL International, 2018. [Online]. Available: http://www.ethnologue.com.

[5] S. Larasati, A. Kumawat, and A. Purwarianti, "Towards a better Indonesian POS Tagger," in International Conference on Asian Language Processing, 2012, pp. 63-66.

[6] D. Cer et al., "Universal sentence encoder," arXiv preprint arXiv:1803.11175, 2017.

[7] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," arXiv preprint arXiv:1901.11196, 2019.