BERT-Based Hybrid Deep Learning with Text Augmentation for Sentiment Analysis of Indonesian Hotel Reviews

Maxwell Thomson, Hendri Murfi and Gianinna Ardaneswari
Department of Mathematics, Universitas Indonesia, Depok 16424, Indonesia

Keywords: Sentiment Analysis, Deep Learning, BERT, Hybrid Model, Text Augmentation.

Abstract: Indonesia's tourist industry plays a significant role in the country's economic growth. Despite being impacted by COVID-19, the occupancy rate of hotels in June 2022 reached 50.28%, surpassing the previous record of 49.17% in January 2020. As hotel occupancy rates rise, it becomes increasingly important to analyze customer reviews of hotels through sentiment analysis to categorize the emotions expressed in the reviews. While a BERT-based hybrid deep learning model has been shown to perform well in sentiment analysis, the nature of class imbalance is often a problem. To address this, the text augmentation method provides a solution to increase the amount of minority class training data through existing data. This paper evaluates five word-level text augmentation methods for the BERT-based hybrid model on classifying sentiments in Indonesian hotel reviews. Our simulations show that the text augmentation methods can improve the model performance for all datasets dan unit measures. Moreover, the random swap method achieves the highest precision and specificity on two of three datasets.

1 INTRODUCTION

The tourism sector is one of the sectors that have a significant impact on economic growth in Indonesia. According to data from The World Travel & Tourism Council, the tourism sector in Indonesia has the highest growth rate, with a rank of 9th in the world in 2018 (World Travel & Tourism Council, 2018). Due to the COVID-19 pandemic, the hotel room occupancy rate dropped to 12.67% in April 2020. However, it has recovered and reached 50.28% in June 2022, even higher than the rate of 49.17% in January 2020 before the pandemic (Badan Pusat Statistik. 2022).

As demand for hotel room occupancy increases, customer reviews of hotels become increasingly important. From the hotels' perspective, these reviews serve as a benchmark for evaluating and improving their services. In contrast, customers are a factor in determining the suitability of booking a hotel. One type of analysis that can be applied to these reviews is sentiment analysis (Sun et al., 2020), which is the process of extracting information such as opinions, views, sentiments, emotions, and evaluations of entities such as products, services, organizations, individuals, issues, events, and topics (Liu, 2015). In this case, sentiment analysis can also be applied to hotel reviews to detect the sentiments within them, such as positive and negative sentiments.

Several deep learning models commonly used in sentiment analysis include recurrent-based models such as Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM), as well as convolutional-based models such as Convolutional Neural Networks (CNN) (Zhang et al., 2018). Some simulations have shown that hybrid models built based on a combination of basic deep learning models have been demonstrated to yield improved performance. For example, the hybrid CNN-GRU model slightly outperformed both the CNN and GRU models in classifying sentiment in Indonesian e-commerce reviews (Gowandi et al., 2021).

Transformers have become a viral language model (Vaswani et al., 2017). One of the models developed from transformers is Bidirectional Encoder Representations from Transformers (BERT), a model used to contextually represent words in a sentence (Devlin et al., 2019). Several simulations show that the text representation of BERT has significantly improved the performance of hybrid deep learning compared to the text representation of fine-tuned embedding (Murfi et al., 2022). From another point of view, we can also say

Thomson, M., Murfi, H. and Ardaneswari, G.
BERT-Based Hybrid Deep Learning with Text Augmentation for Sentiment Analysis of Indonesian Hotel Reviews.
DOI: 10.5220/0012127400003541
In *Proceedings of the 12th International Conference on Data Science, Technology and Applications (DATA 2023)*, pages 468-473
ISBN: 978-989-758-664-4; ISSN: 2184-285X

that hybrid deep learning slightly improves the neural network's performance with the text representation of BERT.

Besides the text representation and the feature selection, another common problem of sentiment analysis is class imbalance. Class imbalance is when the number of observations in one class significantly differs from those in other classes. When a class imbalance occurs, the model in the learning process tends not to receive feedback from the minority class, thus not resulting in optimal performance on new minority class observations. On the other hand, the number of minority class observation data for sentiment analysis is usually much lower. To address this issue, text augmentation methods can collect additional minority class training data by creating new training data from existing data (Bayer et al., 2022; Wilie et al., 2020; Wei & Zou, 2019). This paper examines five word-level text augmentation methods, namely Random Swap (RS), Random Deletion (RD), Random Combination (RC), Text Embedding-based (FE), and Language Model-based (LM), for the BERT-based hybrid CNN-GRU model on classifying sentiment in Indonesian hotel reviews. Our simulations show that the text augmentation methods can improve the model performance for all datasets dan unit measures. Moreover, the random swap method achieves the highest precision and specificity on two of three datasets.

The structure of this paper is as follows: in Section 2, we explained the related works on text augmentation techniques. In Section 3, we briefly explain methods. We describe the experiments in Section 4 and the results in Section 5. Finally, a general conclusion about the results is presented in Section 6.

2 RELATED WORKS

Sentiment analysis has been a widely researched area in natural language processing, and several approaches have been proposed to tackle its associated challenges. One notable work in text augmentation techniques is the Easy Data Augmentation (EDA) method proposed by Wei and Zou in 2019 (Wei & Zou, 2019). EDA introduces four operations: synonym replacement, random insertion, random swap, and random deletion, to augment the training data for text classification tasks using CNN and Recurrent Neural Network (RNN) classifiers. These methods improved accuracy on five different text classification tasks with an average of 0.8% for complete datasets.

Additionally, Bayer, Kaufhold, and Reuter conducted a comprehensive survey on data augmentation for text classification (Bayer et al., 2022). This survey explores different levels of various augmentation methods and their impact on improving classification performance. This survey also discusses embedding-based and language model-based approaches for word-level augmentation. Unlike simple EDA methods, embedding-based processes utilize pre-trained word embeddings to generate new augmented training texts. In contrast, language model-based methods employ pre-trained language models to create new text instances.

In summary, our work draws upon the EDA technique introduced by Wei and Zou (2019) for simple text augmentation. It incorporates embedding-based and language model-based methods discussed in the survey by Bayer et al. (2022). By applying these augmentation methods to the sentiment analysis of Indonesian hotel reviews, we extend their application to a specific language and domain, contributing to the advancement of sentiment analysis in a unique context.

3 METHODOLOGY

3.1 Word-Level Text Augmentation

In the case of imbalanced datasets, text augmentation methods can increase the number of data in the minority class using the existing data to achieve a balanced distribution of classes in the training set, resulting in a more robust classification model. In the context of textual data, data augmentation methods can be grouped into four levels: character level, word or token level, phrase or sentence level, and document level (Bayer et al., 2022). In this research, five different types of word-level text augmentation methods were implemented; they are:

3.1.1 Random Swap (RS)

RS randomly swaps words found in the review sentence to generate new sentences. The aim is to produce different versions of the original sentence while preserving the meaning and context by randomly swapping words within the review sentence.

3.1.2 Random Deletion (RD)

RD randomly removes words in the review sentence to generate new sentences. The objective is to create new sentences by randomly eliminating words from the review sentence while retaining the meaning and context of the original sentence.

3.1.3 Random Combination (RC)

RC method combines the RS and RD methods in generating new sentences. In other words, in a set of augmented data, 50% of the augmented sentences are generated using the RS method, and 50% are generated using the RD method.

3.1.4 Text Embedding-Based (FE)

FE method utilizes pre-trained FastText embeddings to generate new sentences. Randomly selected words are first transformed into latent representation space, commonly in vectors (embeddings), where words with similar contexts are located nearby. Then, words with similar semantic contexts in the nearby latent space are selected to replace those in the initial sentence. The FastText model used in this research is the cc.id.300.vec model for the Indonesian language, each word is represented in a 300-dimensional vector representation.

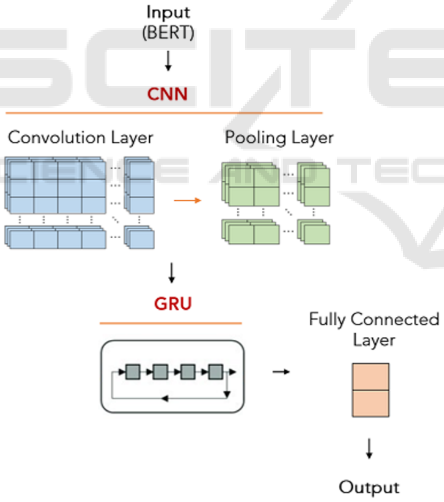

Figure 1: The architecture of BERT-based CNN-GRU model.

3.1.5 Language Model-Based (LM)

The LM method in this research employs a pre-trained BERT model to generate new sentences. Randomly selected words are masked, and surrounding context words are used to predict the masked words utilizing a language model, as the masked language modeling task was done. The pre-trained BERT model is the IndoBERT$_{\text{LARGE}}$ model, with 335.2 million parameters (Wilie et al., 2020).

It should be noted that RS and RD are two accessible data augmentation techniques introduced by Wei and Zou in 2019 (Wei & Zou, 2019). The text augmentation formula applied in this research adheres to Wei and Zou, expressed as $n = \alpha l$, where α is a parameter that indicates the percentage of words in a sentence that have been altered. $\alpha = 0.1$ is the optimal value based on their findings.

3.2 The BERT-Based CNN-GRU Model

The pre-trained IndoBERT$_{\text{BASE}}$ represents the hotel reviews (Wilie et al., 2020). Before hotel reviews are converted into IndoBERT$_{\text{BASE}}$ representations, some pre-processing steps exist. First, each hotel review was tokenized using the WordPiece model. Each review is segmented into sub-words in the vocabulary and added with unique tokens such as [CLS] and [SEP]. Then, padding is applied to each review text input with a maximum of 128 tokens to ensure that each IndoBERT$_{\text{BASE}}$ input has the same fixed length. Using the same WordPiece model, each token in the input sequence will be assigned a unique key, where each key represents a high-dimensional vector or embedding. The embeddings are then fed into the pre-trained IndoBERT$_{\text{BASE}}$ model as inputs and processed in the transformer encoder. IndoBERT$_{\text{BASE}}$ contains 12 encoder layers, 12 attention heads, and a hidden size 768. The output from the first encoder layer serves as input for the second encoder layer, and this process is repeated 12 times. The output of IndoBERT$_{\text{BASE}}$ is a contextualized embedding with a size of 128×768 for each input sequence.

The CNN component of the CNN-GRU consists of a single convolutional layer followed by a max-pooling layer with a pool size of 2. The output from the CNN is then fed into the GRU model, which consists of 200 hidden units. Finally, the output produced by the GRU model is fed into a fully connected layer to predict the final sentiment of the review, as shown in Figure 1.

Table 1: Dataset class distribution.

Dataset	Positive Class	Negative Class
Traveloka	1294 (64.7%)	707 (35.3%)
Tiket	1441 (71.9%)	564 (28.1%)
Pegi Pegi	1572 (78.5%)	430 (21.5%)

4 EXPERIMENTS

There are three datasets consisting of Indonesian-language hotel reviews retrieved from the following Online Travel Agencies: Traveloka (ww.traveloka.com), Tiket (www.tiket.com), and Pegi Pegi (www.pegipegi.com). The 6,008 hotel reviews were collected from 2017 to 2022 for some hotels in Jakarta, Bali, Semarang, and other regions. Labels of positive and negative sentiment on each hotel review were annotated manually. The distribution of class is shown in Table 1.

Table 2: Set of candidate hyperparameter values.

Layer	Hyperparameter	Value
CNN	Number of filters	200; 250; 300
	Filter size	3; 4; 5
	L2 CNN	0.001; 0.01
GRU	Number of units	100; 150; 200
	L2 kernel	0.001; 0.01
	L2 recurrent	0.001; 0.01
Fully Connected	L2 dense	0.001; 0.01

After the hotel reviews were labeled, the review data was subjected to pre-processing stages. The pre-processing steps are removing symbols and punctuation, removing emoticons, converting letters to lowercase, and removing words listed in the Indonesian language NLTK stop-word. Finally, the pre-processed data is split into training and testing sets, and the training data is used to implement text augmentation. The five text augmentation methods were applied to the minority class, the review with negative sentiment, to balance the sentiment class distribution across all three datasets.

Several hyperparameters need to be optimized during the model training. The hyperparameters and their candidate values are shown in Table 2. The hyperparameter tuning is not performed on all available combinations but on a subset of hyperparameter combinations using the Bayesian optimization method. The model training uses 50 epochs, Adam optimization technique with a learning rate of 1×10^{-3}, the batch size of 32, and the loss function of categorical cross-entropy is utilized. An early stopping method is applied with a patience value of 5. This means the learning process will stop if the validation loss does not change after five epochs.

5 RESULTS

The performance of deep learning models heavily depends on the weight initialization set before the learning process begins. As a result, the same deep learning model with the same architecture and hyperparameters can result in varying performance outcomes when executed at different times. Therefore, in this research, the hybrid CNN-GRU model was fitted five times, and the average performance was calculated as the final performance of the model. Table 3, Table 4, and Table 5 shows the results of different word-level text augmentation methods performed on three distinct datasets. Those tables also display the standard deviation to provide further insight into the data.

Table 3 demonstrates that the Traveloka dataset's precision and specificity metrics improved across all text augmentation methods compared to the baseline method, a model trained on data without augmentation. Among those methods, LM, which utilizes a pre-trained BERT model in the augmentation process, produced the best performance in terms of the AUC-PR metric with a lower standard deviation than the baseline. The RC method had the highest AUC-ROC value among other methods. On the other hand, the highest precision and specificity values were achieved with the implementation of the RS augmentation method, resulting in an increase of 1.67% and 4.71%, respectively, over the baseline. Meanwhile, FE produced the best recall/sensitivity results compared to other methods. All five text augmentation methods (RS, RD, RC, FE, and LM) resulted in higher precision and specificity values compared to the baseline, with increases of 1.67%, 0.91%, 0.49%, 0.62%, and 1.03% for precision, and 4.71%, 2.53%, 1.62%, 1.62%, and 2.71% for specificity, respectively.

Table 4 shows that the tiket.com dataset's precision and specificity metrics improved across all text augmentation methods compared to the baseline method. Among those methods, FE, which incorporates the FastText embedding model in its augmentation process, produced the best performance regarding AUC-ROC and AUC-PR metrics with a lower standard deviation than the baseline. The RS augmentation method resulted in the highest precision and specificity values, with an increase of 2.71% and 9.13%, respectively, over the baseline. Simultaneously, RD yielded the best recall/sensitivity results compared to other methods. All five text augmentation methods (RS, RD, RC, FE, and LM) resulted in higher precision and specificity values compared to the baseline, with increases of 2.71%, 0.01%, 2.03%, 1.64%, and 0.03% for precision, and

9.13%, 0.005%, 6.91%, 5.35%, dan 0.45% for specificity, respectively.

Table 5 of the Pegi Pegi dataset, all text augmentation methods improved precision and specificity metrics compared to the baseline method. Of those methods, RC combining RS and RD in the augmentation process yielded the best performance in AUC-ROC and AUC-PR metrics with a lower standard deviation than the baseline. The RD augmentation method achieved the highest precision and specificity values, which increased by 6.42% and 34.23%, respectively, over the baseline. Concurrently, LM produced the highest recall/sensitivity results compared to other methods. All five text augmentation methods (RS, RD, RC, FE, and LM) resulted in higher precision, specificity, AUC-ROC, and AUC-PR values compared to the baseline, with increases of 6.02%, 6.42%, 4.96%, 4.72%, and 1.45% for precision; 32.23%, 34.23%, 26.24%, 27.24%, and 7.97% for specificity; 1.56%, 1.15%, 2.63%, 1.5%, and 1.25% for AUC-ROC, and 0.58%, 0.41%, 0.94%, 0.44%, and 0.44% for AUC-PR, respectively.

Furthermore, sentiment analysis of hotel reviews from multiple locations in Indonesia can be subject to certain limitations. Previous research (Padilla et al., 2018; Gore et al., 2015) has shown the existence of visitor versus resident bias and geographic bias in sentiment expression. Visitors tend to express more positive sentiments in a city than residents, and opinions expressed through Twitter can vary geographically. It is essential to acknowledge that these factors may influence our results.

Table 3: Evaluation results on traveloka dataset.

Method	**Metric**				
	Precision	***Recall /Sensitivity***	***Specificity***	***AUC-ROC***	***AUC-PR***
Base	0.8899 ± 0.02580	0.9443 ± 0.03090	0.7829 ± 0.06150	0.9545 ± 0.00660	0.9743 ± 0.00400
RS	**0.9048 ± 0.02777**	0.8899 ± 0.02590	**0.8198 ± 0.06831**	0.9577 ± 0.00583	0.9769 ± 0.00302
RD	0.8981 ± 0.02013	0.9397 ± 0.01741	0.8027 ± 0.04870	0.9544 ± 0.00536	0.9739 ± 0.00360
RC	0.8944 ± 0.03037	0.9296 ± 0.02175	0.7957 ± 0.07010	**0.9583 ± 0.02232**	0.9706 ± 0.00696
FE	0.8955 ± 0.02352	**0.9459 ± 0.01658**	0.7957 ± 0.05363	0.9542 ± 0.00606	0.9741 ± 0.00407
LM	0.8992 ± 0.03463	0.9335 ± 0.02580	0.8042 ± 0.08090	0.9581 ± 0.00282	**0.9770 ± 0.00197**

Table 4: Evaluation results on tiket.com dataset.

Method	**Metric**				
	Precision	***Recall /Sensitivity***	***Specificity***	***AUC-ROC***	***AUC-PR***
Base	0.9187 ± 0.01805	0.9006 ± 0.03917	0.7946 ± 0.05648	0.9354 ± 0.00859	0.9717 ± 0.00477
RS	**0.9436 ± 0.02209**	0.8520 ± 0.04965	**0.8672 ± 0.06225**	0.9347 ± 0.00572	0.9728 ± 0.00272
RD	0.9189 ± 0.01992	**0.9041 ± 0.02890**	0.7946 ± 0.06270	0.9379 ± 0.00367	0.9749 ± 0.00163
RC	0.9374 ± 0.03191	0.8520 ± 0.04820	0.8495 ± 0.09550	0.9331 ± 0.00830	0.9716 ± 0.00380
FE	0.9338 ± 0.02717	0.8756 ± 0.05284	0.8371 ± 0.07829	**0.9417 ± 0.00313**	**0.9765 ± 0.00177**
LM	0.9190 ± 0.02224	0.8886 ± 0.02872	0.7982 ± 0.06605	0.8632 ± 0.01077	0.9703 ± 0.00591

Table 5: Evaluation results on pegi pegi dataset.

Method	**Metric**				
	Precision	***Recall /Sensitivity***	***Specificity***	***AUC-ROC***	***AUC-PR***
Base	0.9234 ± 0.02719	0.9714 ± 0.04541	0.6999 ± 0.12177	0.9651 ± 0.01218	0.9879 ± 0.00276
RS	0.9790 ± 0.00276	0.9504 ± 0.03093	0.9255 ± 0.01037	0.9802 ± 0.00397	0.9937 ± 0.00163
RD	**0.9827 ± 0.01010**	0.9282 ± 0.05209	**0.9395 ± 0.03713**	0.9762 ± 0.00461	0.9920 ± 0.00175
RC	0.9692 ± 0.02505	0.9663 ± 0.03282	0.8836 ± 0.09968	**0.9905 ± 0.00245**	**0.9973 ± 0.00067**
FE	0.9702 ± 0.00456	0.9745 ± 0.01003	0.8906 ± 0.01765	0.9796 ± 0.00737	0.9923 ± 0.00335
LM	0.9368 ± 0.00978	**0.9885 ± 0.00659**	0.7557 ± 0.04027	0.9772 ± 0.00785	0.9923 ± 0.00319

6 CONCLUSION

This study evaluates the impact of various word-level text augmentation methods on a hybrid CNN-GRU model with BERT representation for sentiment analysis of Indonesian hotel reviews. Our simulations show that the performance of each word-level text augmentation method varied across the datasets. All five word-level text augmentation methods (RS, RD, RC, FE, and LM) yielded higher precision and specificity than the baseline method. The methods increase the precision and specificity on average as much as 0.94% and 2.64% for the Traveloka dataset, 1.28% and 4.37% for the Tiket.com dataset, and 4.71% and 25.58% for the Pegi Pegi dataset. The RS method achieved the highest results for precision and specificity, yielding the most optimal performance on two datasets.

ACKNOWLEDGEMENT

The Directorate of Research and Development, Universitas Indonesia, funded this research under Hibah PUTI 2023 (Grant No. NKB-476/UN2.RST/HKP.05.00/2023).

REFERENCES

Badan Pusat Statistik. (2022). Tingkat penghunian kamar pada hotel bintang. Retrieved from https://www.bps.go.id/indicator/16/122/1/tingkat-penghunian-kamar-pada-hotel-bintang.html

Bayer, M., Kaufhold, M.-A., & Reuter, C. (2022). A survey on data augmentation for text classification. *arXiv preprint arXiv:2107.03158*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*, pages 4171–4186.

Gore, R. J., Diallo, S., & Padilla, J. (2015). You are what you tweet: Connecting the geographic variation in America's obesity rate to Twitter content. PLoS ONE, 10(9).

Gowandi, T., Murfi, H., & Nurrohmah, S. (2021). Performance analysis of hybrid architectures of deep learning for Indonesian sentiment analysis. In *Soft Computing in Data Sciences*, pages 18-27.

Liu, B. (2015). Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge: Cambridge University Press.

urfi, H., Syamsyuriani, , Gowandi, T., Ardaneswari, G., & Nurrohmah, S. (2022). Bert-based combination of convolutional and recurrent neural network for Indonesian sentiment analysis. *arXiv preprint arXiv:2211.05273*.

Padilla, J. J., Kavak, H., Lynch, C. J., Gore, R. J., & Diallo, S. Y. (2018). Temporal and spatiotemporal investigation of tourist attraction visit sentiment on Twitter. PLoS ONE, 13(6).

Sun, F., Chu, N., & Du, X. (2020). Sentiment analysis of hotel reviews based on deep learning. In *International Conference on Robots & Intelligent Systems* (ICRIS), pages 627–630.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Wei, J., & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP), pages 6381–6387.

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843-857.

World Travel & Tourism Council. (2018). Travel & tourism power and performance. London, UK: World Travel & Tourism Council.

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. WIREs Data Mining and Knowledge Discovery, 8(4).