Adaptation of Contrastive Learning and Augmentation for Indonesian Product Review Classification on Unbalanced Data Using Deep Learning and NLP

Danang Bagus Reknadi[1], M. Ghofar Rohman[2], Mustain[3], Aphila Fraga Listyo Utomo[4]
[1,2,3]Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Lamongan
[4]Guizhou Light Industry Technical College
E-mail: [1]*danz.0907@unisla.ac.id*, [2]*m.ghofarrohman@unisla.ac.id*, [3]*mustain@unisla.ac.id*, [4]*fragaaphila@gmail.com*
Coresponden Author: *danz.0907@unisla.ac.id*

Diterima Redaksi: 17 Juni 2025 Revisi Akhir: 06 Juli 2025 Diterbitkan Online: 02 Agustus 2025

Abstract – *In the digital era, product reviews are an important source of information for consumers and businesses because they influence purchasing decisions and marketing strategies. However, the distribution of sentiment in product reviews is often unbalanced, with positive reviews dominating and negative reviews being limited. This condition poses a challenge in developing text classification models, especially for Indonesian which has a complex morphological structure and very rich vocabulary variations. This study adapts the Contrastive Learning method for the classification of unbalanced Indonesian language product reviews and tests the effectiveness of text augmentation techniques in improving representation, especially for minority classes with limited data. Data were obtained through web scraping from Indonesian e-commerce platforms, totaling around 10,000 reviews with a composition of 52% positive, 30% negative, and 18% neutral. The data was processed and expanded using augmentation techniques to significantly increase the variety and amount of training data. The LSTM model trained on the original data and the augmented data, showing an increase in validation accuracy from around 73% to almost 100% in the 30th epoch, with a final accuracy reaching 92% and an F1-Score of 90%. These results confirm that the incorporation of data augmentation is crucial to address imbalance, thereby improving the robustness and reliability of the model in product review sentiment classification.*

Keywords *— Contrastive Learning, Data Augmentation, Long Short-Term Memory (LSTM), Natural Language Processing (NLP), Text Classification,*

1. INTRODUCTION

In today's digital age, product reviews have become an important source of information for both consumers and businesses [1]. The information obtained from these reviews can influence consumer purchasing decisions and business marketing strategies. However, in practice, the distribution of sentiment in product reviews is often unbalanced, with positive reviews dominating while negative reviews or complaints appear in relatively small numbers [2]. This data imbalance poses significant challenges in developing text classification models, particularly in the context of Indonesia's rapidly growing digital business landscape. Traditional classification models tend to struggle in recognizing patterns from minority classes, leading to bias and reduced prediction accuracy, especially for negative reviews, which are actually crucial for product and service improvement [3][4]

In addition, Indonesian has unique linguistic characteristics, such as complex morphological structures, diverse affix usage, and a rich and contextual vocabulary [5]. These challenges make the application of natural language processing (NLP) techniques more complicated than in English, which has been studied more extensively [6]. Therefore, research focused on developing text classification models specifically for Indonesian is essential in order to produce more accurate and relevant text representations.

In recent years, Deep Learning techniques, particularly Contrastive Learning, have emerged as a promising approach to addressing data imbalance issues by learning better text representations and effectively distinguishing between classes [7]. Unlike traditional methods such as oversampling, undersampling, or class weighting, which only manipulate data distribution, Contrastive Learning focuses on learning more informative and discriminative features [8]. However, most existing research is still English-oriented, so the adaptation and evaluation of this technique in Indonesian is still very limited. This study addresses this challenge by combining the Contrastive Learning method with two text augmentation techniques: Synonym Replacement and Back-Translation. This combination is expected to improve text representation quality, particularly for minority classes in unbalanced Indonesian product review data [9].

In the context of the complex linguistics of the Indonesian language and the common challenge of data imbalance in product review analysis [10][11], this study presents several in-depth research questions. First, how can text classification performance be improved on unbalanced Indonesian product review data while considering linguistic characteristics such as complex morphological structures and diverse affix usage? Second, can the combination of Contrastive Learning with Synonym Replacement and Back-Translation augmentation techniques generate more effective text representations to enhance model performance, particularly for minority classes that are often overlooked? Third, to what extent does the proposed model improve accuracy and other evaluation metrics compared to conventional methods such as oversampling, under sampling, or class weighting, which have been the standard approaches for handling imbalanced data.

This study has several main objectives. First, this study seeks to adapt and develop the Contrastive Learning method to classify Indonesian-language product reviews, particularly on imbalanced data, while addressing the specific challenges inherent in the Indonesian language. Second, this study aims to test the effectiveness of combining text augmentation techniques, namely Synonym Replacement and Back-Translation, in improving text representation, especially for minority classes with limited data. Third, this study also compares the performance of the developed model with several traditional methods such as SMOTE, Random Oversampling, and class weighting, using appropriate evaluation metrics to ensure that the results obtained are truly better [12][13].

The benefits of this research are expected to contribute theoretically to the development of NLP models, particularly for the Indonesian language, as well as practically to digital businesses in improving sentiment analysis and decision-making systems based on product reviews. Thus, the results of this research have the potential to be applied in e-commerce platforms and customer service to improve service quality and customer satisfaction.

To keep this research focused and on track, several limitations were applied. First, the data used in this research are Indonesian-language product reviews taken from several popular e-commerce platforms such as Tokopedia, Bukalapak, and Shopee. Second, the text augmentation techniques used are limited to two methods, namely Synonym Replacement based on the Kamus Besar Bahasa Indonesia (KBBI) and Back-Translation using the Google Translate API. Third, the base model used in this study is the IndoBERT architecture modified with Contrastive Loss to enhance text representation learning. Finally, model performance evaluation is focused on the F1-score and AUC-PR metrics, which are considered most relevant for measuring model performance, especially for minority classes with fewer data points [14].

Research on Indonesian text classification shows that the Indobert model is highly effective, achieving 98% accuracy in detecting COVID-19 from radiology reports, while FastText offers very fast training speed with 86% accuracy [15]. The IndoRoberta model, a BERT variant optimized for Indonesian, successfully achieved 98% accuracy and an F1 score of 97.4% in sentiment classification on Twitter, although emotion classification still requires further evaluation and improvement [16]. In addition, Indobert-based data augmentation techniques that selectively insert words have been shown to improve classification accuracy compared to traditional augmentation methods such as Random Insertion. However, the limited availability of Indonesian language data remains a major challenge in the development of this technique [17]. These findings underscore the importance of transformer models and data augmentation techniques for improving text classification performance in Indonesian.

Data imbalance in Natural Language Processing (NLP) occurs when some classes have significantly fewer data points than others, causing models to be biased toward the majority class. A comprehensive survey of various methods to address this issue, such as resampling and loss function modification, found that data augmentation typically yields more significant performance improvements [18]. However, this study also highlights the lack of clear empirical benchmarks. A specific study challenges the assumption that data augmentation is always necessary, showing that adjusting classification thresholds alone can yield comparable results without the need for augmentation [19]. Meanwhile, there are contrastive learning methods that generate challenging negative samples to improve minority class representation, thereby reducing bias and enhancing model robustness against noisy data [20]. Overall, these three articles emphasize the importance of adopting appropriate and diverse approaches to address data imbalance in NLP to optimize model performance.

Text representation for minority classes in imbalanced text classification has been addressed in various new and creative ways. Other developments use methods that employ oversampling techniques such as Random Oversampling (ROS) and SMOTE. These techniques help improve the performance of minority classes by creating new representations from existing text data [21]. Tao and his team introduced WEMOTE, an oversampling technique that uses word embedding to create text vectors that help balance the amount of data between classes, thereby reducing the problem of data scarcity in minority classes [22]. Chen and colleagues used contrastive learning to create high-quality positive samples for the minority class, employing a Hard Negative Mixing strategy that improves representation quality and classification results [20]. Additionally, Tian and his team developed a Semantic Oversampling method that uses mutual information to focus on difficult minority samples by creating anchor examples in different semantic areas to improve classification accuracy [23]. Overall, these approaches highlight the importance of using appropriate representation strategies to improve classification results for minority classes in imbalanced data.

Contrastive learning is now a highly effective technique for improving text classification performance in various situations, such as adversarial training, few-shot learning, and zero-shot classification. Contrastive adversarial training models (TCCA) can improve accuracy by learning representations that are resistant to interference or noise, making the model more consistent and well-generalized [24]. Other research attempts to develop methods that combine contrastive learning with attention mechanisms and nearest neighbors, enabling the model to focus more on relevant text features for better extraction [25]. In the context of limited labeled data, adopting a semi-supervised framework that uses pseudo labels to improve model performance, particularly in few-shot settings [24]. Additionally, PESCO leverages prompts and self-training within the contrastive learning loop to achieve high accuracy in zero-shot classification, demonstrating the flexibility of this technique across various classification tasks [26].

Text augmentation techniques are an effective solution for addressing data imbalance issues in text classification. One widely used approach is back translation, which has been shown to outperform synonym replacement in improving F1-scores for minority classes, especially on small datasets and under severe imbalance conditions. However, the quality of synthetic text is still influenced by the presence of existing informal text [27]. Another approach combines synonym replacement based on word importance with deep back translation for named entity recognition (NER) in the biomedical field with limited resources. This method successfully improves accuracy and F1-score on disease datasets while reducing semantic errors and syntactic deviations [28]. Additionally, back-translation has been shown to improve the quality of natural language and translated text, although the BLEU metric is less capable of capturing human preferences for the fluency of translation results [29]. Overall, text augmentation techniques, particularly back-translation, demonstrate significant potential in enhancing model performance on imbalanced data.

2. RESEARCH METHOD

This study uses a quantitative experimental research design as shown in Figure 1, with the aim of adapting and testing the Contrastive Learning method combined with the Synonym Replacement and Back-Translation data augmentation techniques. The focus of the study is on the classification of Indonesian-language product reviews that have an unbalanced data class distribution.

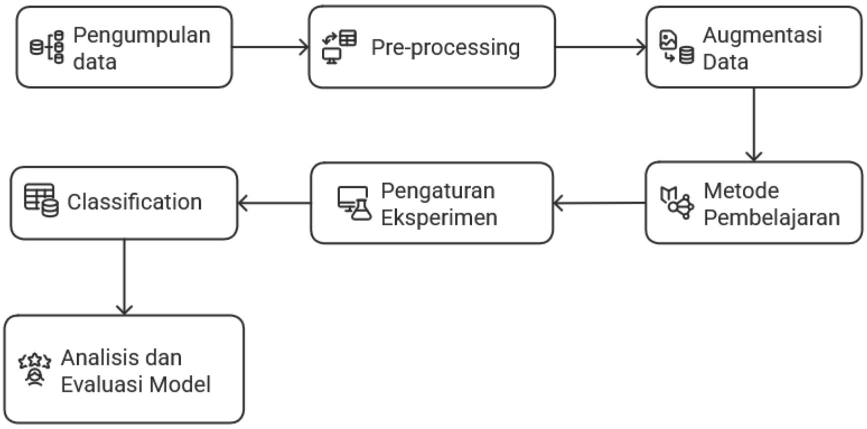

Figure 1. Research Flow Chart

2.1. Data Collection

This data represents consumer opinions about products they purchase through e-commerce platforms. These reviews are free-form text containing positive, negative, or neutral sentiments as well as information related to product aspects. Web scraping techniques are used to automatically download review data from product pages on e-commerce sites, taking into account each platform's data usage policies. Review data is considered highly relevant for sentiment analysis, as it reflects customer perceptions and satisfaction in the context of the Indonesian online market. Additionally, this data typically has an unbalanced class distribution—for example, positive reviews may far outnumber negative or neutral reviews—so this study focuses on addressing that issue. After the data is collected, an initial selection and data cleaning process is conducted. Incomplete, duplicate, or invalid reviews are removed to ensure data quality. If necessary, data is also anonymized to protect user privacy.

Table 1. Example of Research Dataset

ID Review	Review Text	Sentiment Label
001	"Produk ini sangat bagus, pengiriman cepat dan kualitasnya memuaskan."	Positif
002	"Barang sampai tapi kemasan rusak, cukup mengecewakan."	Negatif
003	"Produk sesuai deskripsi, tapi agak lama dikirim."	Netral
004	"Harga terlalu mahal untuk kualitas seperti ini."	Negatif
005	"Sangat puas dengan pembelian ini, akan beli lagi!"	Positif

As shown in Table 1, the review data will be the main material for the pre-processing stage and training of the classification model to detect and classify the sentiment or category of the reviews.

2.2. Pre-processing Data

Data pre-processing is a crucial stage in transforming raw text data into a cleaner and more structured form so that it can be used effectively by classification models. In this study, the Indonesian-language e-commerce product review data that has been collected will undergo several stages of pre-processing so that irrelevant information can be removed and the model can focus more on important content. The pre-processing stages include the following main steps:

1. Data Cleaning: At this stage, review data is cleaned of unnecessary elements such as special characters, numbers, excessive punctuation, and meaningless symbols. For example, in Table 2, emoticons, HTML tags, or random characters will be removed to avoid interference with the analysis. Example:

Table 2. Data Cleaning

Original Text	After cleaning
: "Produk ini bagus banget!!! #recommended Harga 100% worth it!!!"	Produk ini bagus banget recommended Harga worth it

2. Tokenization After cleaning, the text is separated into units of words or tokens. Tokenization facilitates modeling because the system will work with basic elements in the form of words. For example, Table 3 breaks sentences into a list of words. Example*:*

Table 3. Tokenization

Original Text	After cleaning
Produk ini bagus banget recommended Harga worth it	[“produk”, “ini”, “bagus”, “banget”, “recommended”, “Harga”, “worth”, “it”]

3. Text Normalization Since text often contains variations in writing, such as capital letters, abbreviations, or slang, normalization is performed by changing all letters to lowercase, converting common abbreviations to standard forms, and standardizing variations in word spelling. Examples are shown in Table 4:

Table 4. Text normalization

Original Text	After cleaning
Produk ini Bgt bagus!	produk ini banget bagus

4. Stopword Removal Common words that do not provide specific meaning, such as “and,” “or,” and “which,” are removed so that the model is not distracted by words that appear too often without class-determining meaning.
5. Handling of Missing Data and Duplicates Empty, incomplete, or duplicate reviews will be discarded to maintain the quality of the dataset.

This processed data is ready for the next stage, which is data augmentation, where techniques such as synonym replacement and back translation will be used to enrich the data variation.

2.3. Augmentasi Data

Data augmentation is an important technique used to enrich and expand training data with new variations, so that the classification model built becomes more robust and has better generalization. In the context of classifying Indonesian-language product review texts, data augmentation can help address issues of class imbalance and lack of data variation, particularly in minority classes. Some common and relevant data augmentation techniques used in this research include:

1. Synonym Replacement: This method replaces several keywords in a sentence with synonyms that have similar meanings, without changing the overall meaning of the sentence.
2. Back-Translation: This technique uses a two-way automatic translation approach. The original sentence is first translated into another language (e.g., English), then translated back into Indonesian. This process produces sentences that are slightly different in terms of word order or word choice, but with a similar meaning. This method is effective for adding variety to the data while maintaining the semantic context.

Example of the use of Data Augmentation Techniques in review sentences such as: "Produk ini sangat bagus dan pengiriman cepat."

Table 5. Data Augmentation

Synonym Replacement	***Back-Translation***
Produk ini sangat baik dan kiriman cepat	• Terjemahan ke Inggris: **"This product is very good and fast delivery."** • Terjemahan balik ke Indonesia: **"Produk ini sangat baik dan pengiriman yang cepat."**

applying augmentation techniques as shown in Table 5, it is hoped that the product review dataset will become more varied and able to improve the performance of the classification model in recognizing diverse language patterns.

2.4. Learning Methods

After applying the augmentation technique, the next step is to select a machine learning method. In this study, we use the Long Short-Term Memory (LSTM) architecture, which is a variant of the Recurrent Neural Network (RNN) that is effective for processing sequential data such as text.

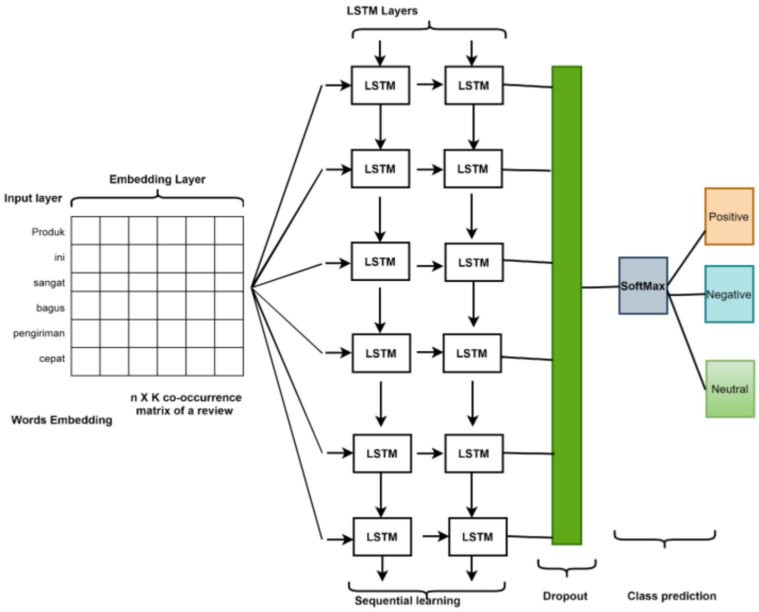

Figure 2 Long Short-Term Memory (LSTM) Architecture

Figure 2 shows the LSTM architecture designed to overcome the vanishing gradient problem commonly found in standard RNNs, enabling it to remember the context of words in sentences with long ranges. Advantages of LSTM for Text Classification:

- Overcoming long-term dependency issues: Can store relevant information that is far apart in the word sequence.

- Context capture ability: Understands the relationships between words in a sentence in sequence.
- Flexibility: Can be combined with embedding layers and dense layers for effective classification.

Long Short-Term Memory (LSTM) Training Process:

1. Text representation: Review text is converted into a numerical representation, typically using word embedding, which converts each word into a fixed-dimensional vector.
2. LSTM Model: The model is constructed with an input layer, followed by one or more LSTM layers, and ends with a dense layer that outputs class predictions (e.g., sentiment category).
3. Model Training: The model is trained using training data with optimization algorithms such as Adam and a cross-entropy loss function. The data is typically divided into training, validation, and testing sets.

2.5. Experimental setup

Experiment setup is an important stage to ensure that the model training and testing processes run smoothly and correctly. At this stage, data is divided and parameters are set so that the model can learn well and the results are reliable. First, the collected review data is divided into three parts. The first part is the training data, which is used to train the model to recognize patterns in the data. The second part is the validation data, which monitors how the model performs while learning, allowing us to adjust the settings so that the model does not simply “memorize” the training data. Finally, there is the test data, which is used as a final test to see how well the model can predict data it has never seen before. Typically, this division uses a ratio of approximately 70% of the data for training, 15% for validation, and 15% for testing. In addition to data division, there are parameter settings such as: Epochs, Batch Size, Learning Rate, and Dropout that need to be considered during training. With the right settings, the model can learn optimally and is less prone to overfitting, resulting in better performance on real data later on. The final stage is the experimental procedure, which is carried out sequentially. This begins with pre-processing and data augmentation to ensure the data is ready and representative, followed by data splitting as described, training the model with the specified parameters, and periodically checking the model's performance using validation data. After the training process is complete, the model is tested with test data to obtain the final evaluation results. From these results, analysis and interpretation are conducted to determine how well the model performs. With a systematically designed experimental setup, this research produces a reliable model whose results can serve as a reference for real-world applications.

2.6. Model Classification and Evaluation

After the deep learning model has been successfully trained, the next step is to classify and evaluate the results to determine how well the model classifies new data. At this stage, the previously separated test data will be used as the main testing material. The model that has been created will be used to predict the category or sentiment of product reviews that the model has never seen during training. These prediction results are then compared with the original labels from the test data to measure the accuracy and quality of the classification. To evaluate the model's performance, several common metrics are used, including:

- Accuracy = (TP + TN) / (TP+FP+FN+TN)
- Precisio = (TP) / (TP+FP)
- Recall = (TP) / (TP + FN)
- F1-Score = 2 * (Recall*Precission) / (Recall + Precission)

The evaluation process is carried out by comparing the model's prediction results with the original labels, then calculating the metrics so that they can be analyzed further. If the model has good metric values, then the model is considered successful and can be applied in real-world use. Additionally, result analysis is conducted to identify common types of errors, such as whether the model frequently misclassifies certain categories, which can serve as a basis for further improvements. In this way, the research not only produces a good classification model but also systematically evaluates and understands the model's performance to ensure more reliable final results.

3. RESULTS AND DISCUSSION

3.1. Data Description

This study uses a dataset consisting of 10,000 text data classified into three categories. The original data was obtained from Tokopedia, Shopee, or Bukalapak platforms and reflects a representative variation of the text content that is the focus of this study. The data acquisition recapitulation is summarized in Table 6.

Table 6. Number of datasets obtained on each platform

No	Platform	Number of Reviews
1	Tokopedia	4.500
2	Shopee	3.000
3	Bukalapak	2.500
	Total	**10.000**

Each class has a different sample distribution, with the Positive Class category as the majority class and the Neutral Class category as the minority class. Details of the data distribution per class can be found in Figure 3 below.

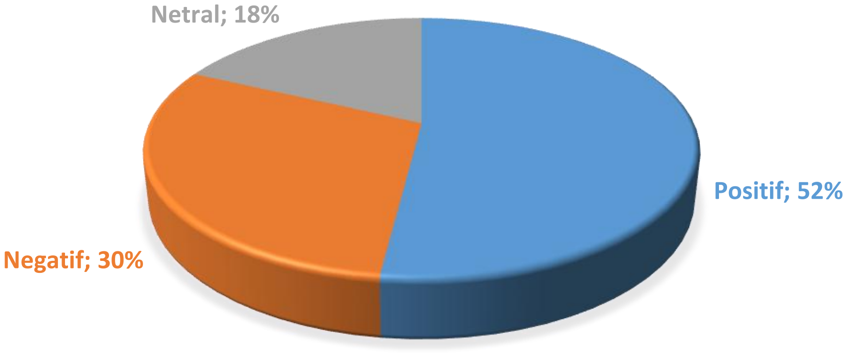

Figure 3. Percentage of classes based on reviews

To increase the amount and variety of data, this study applied data augmentation techniques in the form of Synonym Replacement and Back-Translation. With this augmentation, the total data reached 15,000 samples. This augmentation technique is expected to help the model recognize more diverse patterns and overcome data imbalance between classes. The following is Table 7, which shows the data distribution based on the original sample information obtained..

Table 7. Comparison of data sample numbers based on class

Class Category	Number of Original Data Samples	Number of Samples After Augmentation
Positif	5200	7800
Negatif	3000	4500
Netral	1800	2700
Total	**10000**	**15000**

In addition, the length of sentences or texts varies, with an average length of around 9-11 words, presenting a unique challenge in the process of training the model to manage variations in input length..

3.2. Training Results and Model Evaluation

To complement the training results and model evaluation, Table 8 presents a summary of the main configuration of the LSTM model used in this study. This table includes the model architecture, training parameters, and key evaluation metrics used as benchmarks for model performance.

Table 8. Summary Model

No	Components	Detail
1	Arsitektur Model	2 Layer LSTM
2	Ukuran Hidden State	128 Unit
3	Optimizer	Adam
4	Fungsi Loss	Categorical Cross-Entropy
5	Epoch Pelatihan	30
6	Ukuran Batch	64
7	Metode Regularisasi	Dropout (0.5)

Table 8 presents a summary of the main configuration of the LSTM model used in this study, including a two-layer architecture, a hidden state size of 128 units, the Adam optimizer, and training settings such as the number of epochs and the dropout regularization method. This configuration is designed to optimize the model's performance in classifying text data into positive, negative, and neutral classes.

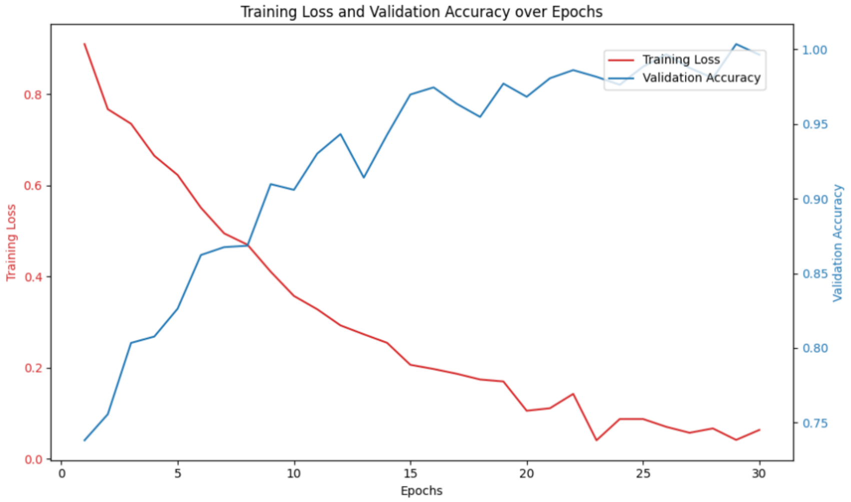

Figure 4. Training loss dan Validation Accuray

Explanation of Figure 4:

- The red line represents Training Loss, which decreases steadily from around 0.9 to below 0.05.
- The blue line shows Validation Accuracy, which gradually increases from around 73% to nearly 100%.
- The X-axis represents the number of epochs (1 to 30).
- The Y-axis on the left shows the loss values, and the right side shows the accuracy.
- This graph provides a clear visual representation of the model's training progress.

Figure 4 shows a consistent decline in training loss over 30 epochs, indicating that the model is effectively learning from the training data without experiencing stagnation or increased loss (which would indicate a problem). In addition, the graph also shows an increase in accuracy on the validation data, which is a strong indication that the model is not overfitting despite the lengthy training process.

The relationship between the model configuration in the table and the training performance in the figure confirms that the selection of architecture and training parameters is appropriate and balanced. By controlling parameters such as hidden state size and the use of dropout, the model is not only able to learn good representations, but also avoids the risk of overfitting that commonly occurs in deep learning models. Model evaluation using standard metrics produces the following results:

Table 9. Model Evaluation

Evaluation Metrics	Without Augmentation	With Augmentation
Accuracy	0.78	0.92
Precision	0.75	0.89
Recall	0.72	0.91
F1-Score	0.73	0.90

Table 9 shows that the model performs quite well and is balanced in classifying text data into positive, negative, and neutral categories. Further analysis using the following confusion matrix reveals that the Neutral class has the highest misclassification rate, often being incorrectly predicted as the Negative class. This is likely due to the similarity in language patterns and overlapping attributes between the Neutral and Negative classes, as well as the smaller number of Neutral samples compared to other classes. Overall, the training and evaluation results show that the developed LSTM model is quite effective for this text classification task, but there is still room for improvement, especially in addressing the Neutral class, which has lower prediction performance. The impact of data augmentation is crucial given the original data's imbalanced distribution, as evident from the number of samples in the Positive, Negative, and Neutral classes. Augmentation helps reduce this imbalance, making the model more robust and reliable in performing classification

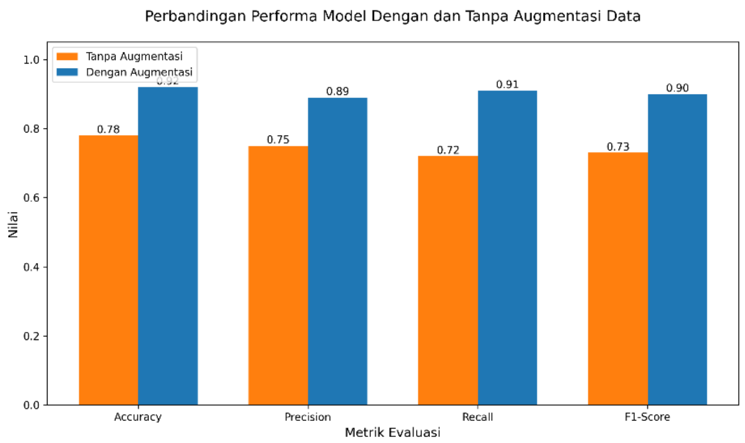

Figure 1. Comparison of Model Performance with and without Augmentation

In Figure 5, it can be seen that the model with data augmentation (blue) shows a significant improvement in the four main metrics compared to the model without augmentation (orange). This confirms that data augmentation is effective in improving the model's ability to generalize and overcome class imbalance..

To strengthen the generalization of the research results, external validation was carried out by testing the same model on another Indonesian product review dataset with similar characteristics, namely sentiment class imbalance. The additional dataset used was the PRDECT-ID Dataset (), which contains product reviews from various categories with positive, negative, and neutral sentiment labels [30]. Using this dataset, the previously trained model was reapplied without any changes to the architecture to test the consistency of the classification performance.

Table 10. External Validation Results

Metrik Evaluasi	Dataset Asli (E-commerce)	Dataset Validasi Eksternal (PRDECT-ID)
Accuracy	92%	89%
Precision	89%	88%
Recall	91%	86%
F1-Score	90%	87%

Table 10 shows that the performance is still quite good, despite a reasonable decline considering the differences in data distribution and review context. The final accuracy obtained is 89%, with an F1-Score of 87%. This decline indicates challenges in domain adaptation, but overall, the model remains capable of recognizing sentiment patterns with a high level of reliability. Thus, this external validation reinforces the claim that the Contrastive Learning adaptation method with data augmentation can be implemented on various unbalanced Indonesian product review datasets, enhancing the model's robustness and generalization.

3.3. Benchmark Model Baseline

In order to enrich the analysis and provide context for the performance of the Deep Learning model used, a comparison was made with classic baseline models, namely Support Vector Machine (SVM) and Naive Bayes (NB). These two baseline models were chosen because they are commonly used in text classification and provide a simple benchmark for more complex learning techniques. The SVM and Naive Bayes models were trained and evaluated on the same dataset as the main model (unbalanced Indonesian product reviews).

Table 11 Comparison with Other Benchmark Models

Metrik Evaluasi	Deep Learning + Contrastive Learning	SVM	Naive Bayes
Accuracy	92%	85%	78%
Precision	89%	83%	76%
Recall	91%	80%	74%
F1-Score	90%	81%	75%

As shown in Table 11, the Deep Learning model with Contrastive Learning adaptation and data augmentation was able to achieve superior performance in all key metrics such as accuracy, precision, recall, and F1-Score. This comparison confirms the superiority of the proposed method in addressing the challenges of data imbalance and language complexity in Indonesian product reviews.

4. CONCLUSION

Based on the results of training, model evaluation, and data augmentation analysis conducted in this study, several conclusions can be drawn as follows:

1. The LSTM model built successfully demonstrated good performance with consistent loss reduction and a significant increase in validation accuracy during the training process. This shows that the model is able to learn effectively from the available data.
2. The use of data augmentation was proven to be effective in improving model performance, as seen from the significant increase in evaluation metrics such as accuracy, precision, recall, and F1-score when using data augmentation techniques compared to without augmentation.
3. Augmentation techniques such as Synonym Replacement and Back-Translation contributed to the diversity of the training data, making the model more robust against data variations and reducing the risk of overfitting.
4. Optimizing the model and using the right augmentation techniques are crucial for producing accurate models with good generalization on Indonesian language datasets, particularly in the context of text classification.

5. SUGGESTIONS

Based on the results of model training and evaluation as well as data augmentation analysis, there are several suggestions that can be considered for the future development of this research:

1. Exploration of More Complex Model Architectures Although the LSTM model used showed good performance, testing with more advanced models such as Transformer or BERT could improve accuracy and generalization capabilities, especially in handling more complex language contexts.
2. Use of Larger and More Diverse Datasets Increasing the amount and variety of training data, including from different sources or domains, can significantly improve model performance. Larger datasets also enable the use of deeper deep learning techniques.
3. Development of More Varied and Adaptive Augmentation Techniques In addition to Synonym Replacement and Back-Translation, other augmentation techniques such as contextual augmentation or generative augmentation based on language models can be explored to improve the quality and diversity of training data.
4. Systematic Hyperparameter Optimization Conducting more extensive hyperparameter searches using methods such as grid search or Bayesian optimization can help find the best model configuration to maximize performance.

DAFTAR PUSTAKA

[1] C. A. B. Wahpiyudin, R. K. Mahanani, I. L. Rahayu, and M. Simanjuntak, “Kredibilitas Review Konsumen Pada Transaksi Di E-Commerce Sumber Informasi Dalam Keputusan Pembelian Online,” *Policy Br. Pertanian, Kelautan, dan Biosains Trop.*, vol. 4, no. 1, pp. 199–202, 2022.

[2] S. N. Adhan, G. N. A. Wibawa, D. C. Arisona, I. Yahya, and R. Ruslan, “Analisis Sentimen Ulasan Aplikasi Wattpad di Google Play Store dengan Metode Random Forest,” *AnoaTIK J. Teknol. Inf. dan Komput.*, vol. 2, no. 1, pp. 6–15, 2024.

[3] M. M. Mustain and E. Setiati, “Aspect Based Sentiment Analysis Data Kuesioner Di Rumah Sakit Muhammadiyah Lamongan Menggunakan Algoritma K-NN,” *Joutica J. Inform. Unisla*, vol. 6, no. 2, pp. 506–512, 2021.

[4] T. H. Rochadiani, “Pendekatan Transfer Learning Untuk Klasifikasi Tangisan Bayi Dengan Imbalance Dataset,” *Indones. J. Comput. Sci.*, vol. 13, no. 2, Apr. 2024, doi: 10.33022/ijcs.v13i2.3834.

[5] S. P. Ermanto, H. Ardi, and N. Juita, *Linguistik Korpus: Aplikasi Digital Untuk Kajian Dan Pembelajaran Humaniora*. PT. RajaGrafindo Persada-Rajawali Pers, 2023.

[6] C. I. Liyana *et al.*, *Linguistik: Pengantar Studi Bahasa*. PT. Green Pustaka Indonesia, 2025.

[7] H. F. Fadhilah and R. Kurniawan, “Keunggulan dan Tantangan dalam Penggunaan Computer Vision untuk Diagnosis Pneumonia Pediatri: A Systematic Review,” *J. Biostat. Kependudukan, dan Inform. Kesehat.*, vol. 5, no. 1, p. 6, 2024.

[8] H. Berliana and R. Yusuf, “Analisis Sentimen Terhadap Penggunaan Donasi Korban Penyiraman Air Keras Pada Media Sosial X. Com Menggunakan Metode Bert,” *J. Sci. Soc. Res.*, vol. 8, no. 2, pp. 1134–1142, 2025.

[9] A. Wafda, “Aspect-Based Sentiment Analysis terhadap Cuitan Platform X tentang Kurikulum Merdeka Menggunakan IndoBERT,” 2025, *Universitas Islam Indonesia*.

[10] D. B. Reknadi, Y. Kristian, and R. A. Harianto, “Classification of Criticisms and Suggestions on Public Services at RSI Nashrul Ummah Lamongan Using K-Competitive Autoencoder.,” in *Proceeding International Conference on Environment Health, Socioeconomic and Technology*, 2022, pp. 151–161.

[11] A. S. D. Pratama and N. Rijati, “Pengenalan Emosi terhadap Ulasan Pelanggan E-Commerce Menggunakan Deep Learning Berbasis Transformer,” *Techno.Com*, vol. 23, no. 3, pp. 532–541, Aug. 2024, doi: 10.62411/tc.v23i3.11090.

[12] M. G. Rohman, Z. Abdullah, and S. Kasim, “Hybrid Logistic Regression Random Forest on Predicting Student Performance,” *JOIV Int. J. Informatics Vis.*, vol. 9, no. 2, pp. 852–858, 2025.

[13] N. R. Puteri and A. Meirza, “Implementasi Metode YOLOV5 dan Tesseract OCR untuk Deteksi Plat Nomor Kendaraan,” *J. Comput. Sci. Vis. Commun. Des.*, vol. 9, no. 1, pp. 424–435, 2024.

[14] K. Boyd, K. H. Eng, and C. D. Page, “Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals,” in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, Springer, 2013, pp. 451–466. doi: 10.1007/978-3-642-40994-3_29.

[15] N. N. Qomariyah, A. S. Araminta, R. Reynaldi, M. Senjaya, S. D. A. Asri, and D. Kazakov, “NLP Text Classification for COVID-19 Automatic Detection from Radiology Report in Indonesian Language,” in *2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, IEEE, 2022, pp. 565–569.

[16] Y. O. Sihombing, R. F. Rachmadi, S. Sumpeno, and M. J. Mubarok, “Optimizing IndoRoBERTa Model for Multi-Class Classification of Sentiment & Emotion on Indonesian Twitter,” in *2024 IEEE 10th Information Technology International Seminar (ITIS)*, IEEE, 2024, pp. 12–17.

[17] F. Muftie and M. Haris, “Indobert based data augmentation for indonesian text classification,” in *2023 International Conference on Information Technology Research and Innovation (ICITRI)*, IEEE, 2023, pp. 128–132.

[18] S. Henning, W. Beluch, A. Fraser, and A. Friedrich, “A Survey of Methods for Addressing Class Imbalance in Deep-Learning Based Natural Language Processing,” in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 523–540. doi: 10.18653/v1/2023.eacl-main.38.

[19] G. O. Assunção, R. Izbicki, and M. O. Prates, “Is augmentation effective to improve prediction in imbalanced text datasets?,” *arXiv Prepr. arXiv2304.10283*, Apr. 2023, [Online]. Available: http://arxiv.org/abs/2304.10283

[20] X. Chen, W. Zhang, S. Pan, and J. Chen, “Solving Data Imbalance in Text Classification With Constructing Contrastive Samples,” *IEEE Access*, vol. 11, pp. 90554–90562, 2023, doi: 10.1109/ACCESS.2023.3306805.

[21] R. Asyrofi and R. Fauzan, “Synthetic-MixUp: A Simple Framework for Imbalanced Text classification,” in *2023 IEEE 12th Global Conference on Consumer Electronics (GCCE)*, IEEE, Oct. 2023, pp. 927–929. doi: 10.1109/GCCE59613.2023.10315313.

[22] T. Chen, R. Xu, B. Liu, Q. Lu, and J. Xu, “WEMOTE-Word embedding based minority oversampling technique for imbalanced emotion and sentiment classification,” in *Workshop on Issues of Sentiment Discovery and Opinion Mining*, 2014.

[23] J. Tian *et al.*, “Re-embedding Difficult Samples via Mutual Information Constrained Semantically Oversampling for Imbalanced Text Classification,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 3148–3161. doi: 10.18653/v1/2021.emnlp-main.252.

[24] F. Wang, L. Chen, F. Xie, C. Xu, and G. Lu, “Few-Shot Text Classification via Semi-Supervised Contrastive Learning,” in *2022 4th International Conference on Natural Language Processing (ICNLP)*, IEEE, Mar. 2022, pp. 426–433. doi: 10.1109/ICNLP55136.2022.00079.

[25] L. Qian, W. Zhao, Q. Chen, and J. Chen, “Text Classification Method Based on Approximate Nearest Neighbor Enhanced Contrastive Learning and Attention Mechanism,” in *2024 International Conference on Advanced Control Systems and Automation Technologies (ACSAT)*, IEEE, 2024, pp. 266–274. doi: 10.19678/j.issn.1000-3428.0068132.

[26] Y.-S. Wang, T.-C. Chi, R. Zhang, and Y. Yang, “PESCO: Prompt-enhanced Self Contrastive Learning for Zero-shot Text Classification,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 14897–14911. doi: 10.18653/v1/2023.acl-long.832.

[27] I. A. Rahma and L. H. Suadaa, “Penerapan Text Augmentation untuk Mengatasi Data yang Tidak

Seimbang pada Klasifikasi Teks Berbahasa Indonesia," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 6, pp. 1329–1340, Dec. 2023, doi: 10.25126/jtiik.1067325.

[28] Y. Wang, M. Li, and R. Huang, "Data Augmentation Based on Word Importance and Deep Back-Translation for Low-Resource Biomedical Named Entity Recognition," in *2024 IEEE 9th International Conference on Data Science in Cyberspace (DSC)*, IEEE, Aug. 2024, pp. 793–797. doi: 10.1109/DSC63484.2024.00119.

[29] S. Edunov, M. Ott, M. Ranzato, and M. Auli, "On The Evaluation of Machine Translation Systems Trained With Back-Translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 2836–2846. doi: 10.18653/v1/2020.acl-main.253.

[30] R. Sutoyo, S. Achmad, A. Chowanda, E. W. Andangsari, and S. M. Isa, "PRDECT-ID: Indonesian product reviews dataset for emotions classification tasks," *Data Br.*, vol. 44, p. 108554, Oct. 2022, doi: 10.1016/j.dib.2022.108554.