Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 10, No. 1, February, Pp. 91-100

Aspect-based multilabel classification of e-commerce reviews using fine-tuned IndoBERT

Fahrendra Khoirul Ihtada[1], Rizha Alfianita[1], Okta Qomaruddin Aziz[*1]
Department of Informatics, Universitas Islam Negeri Maulana Malik Ibrahim, Indonesia[1]

Article Info
Keywords:
Aspect category, Multilabel classification, E-commerce reviews, IndoBERT

Article history:
Received: August 02, 2024
Accepted: December 03, 2024
Published: February 01, 2025

Cite:
F. K. Ihtada, R. Alfianita, and O. Q. Aziz, "Aspect-based Multilabel Classification of E-commerce Reviews Using Fine-tuned IndoBERT", *KINETIK*, vol. 10, no. 1, Feb. 2025.
https://doi.org/10.22219/kinetik.v10i1.2088

*Corresponding author.
Okta Qomaruddin Aziz
E-mail address:
okta.qomaruddin@uin-malang.ac.id

Abstract
In recent years, e-commerce has experienced rapid growth. A significant change in consumer behavior is marked by the ease of access and time flexibility offered by e-commerce platforms, as well as the existence of the review feature to assess products and services. However, with the ever-increasing number of reviews, consumers and store owners face challenges in sorting out relevant information. This research focuses on the multilabel classification of Indonesian e-commerce reviews. This research was undertaken because the application of multilabel classification, especially for e-commerce reviews in Indonesia, has received little attention. This research compares three classification models: end-to-end IndoBERT, IndoBERT-CNN, and IndoBERT-LSTM, to determine the most effective model for multilabel aspect classification of customer reviews. The multilabel classification method was applied to determine the aspect categories of the reviews, such as product, customer service, and delivery, using different thresholds for evaluation. Results show that 0.6 threshold is optimal, with the IndoBERT-LSTM model as the best-performing model for the multilabel aspect classification of these e-commerce reviews. Optimal classification of the model enables more precise information extraction from customer reviews. This can be useful for e-commerce businesses to gain insight from the reviews they get from customers. This insight can be used to find out which aspects need to be improved from the e-commerce business which leads to increased customer satisfaction and trust.

1. Introduction

In recent years, e-commerce has become increasingly popular for buying and selling activities among internet users. Based on data from Statistics Indonesia, survey shows that 80.2% of businesses in Indonesia have used the Internet to sell their product in 2021. Along with this, e-commerce sales have increased 54% since 2019 to IDR 266.3 trillion in 2021 [1]. With this ever-growing e-commerce trend, the number of reviews is also growing. This growth reflects that customers are increasingly relying on reviews to make purchasing decisions. According to Chen et al. [2], almost 60% of consumers browse customer reviews at least once a week. On one hand, a large number of reviews can provide valuable information, but customers may find it difficult to sort out reviews that can help them choose a reliable product [3]. On the other hand, it is also important for online store owners to read reviews so that they can recommend their best products or find out complaints from customers to improve their services. It would take a lot of effort and time to read all types of reviews and separate them manually. As a result, there might be some information that the online shop owners did not capture which made them unable to provide fully efficient services to the customers [4].

Several studies have explored the classification of Indonesian e-commerce reviews using various methods. Imron et al. [5] examined IndoBERT for detecting aspects in reviews from Bukalapak, identifying key categories like service, packaging, quality, price, and accuracy in a dataset of 3,114 reviews. The results showed IndoBERT's effectiveness, achieving significant accuracy, and compared its performance with two neural network models: CNN and LSTM. The CNN model outperformed LSTM, with accuracies of 94.86% and 88.92%, respectively. However, the authors noted challenges with mixed-language reviews and high computational demands. Nasiri and Budi [6] analyzed aspect detection in reviews from mobile applications, using a dataset of 64,113 unduplicated reviews from the Google Play Store and Apple App Store, with 3,748 annotated for analysis. They proposed a model combining GRU and CNN, which achieved an accuracy of 71.3%, with over half of the aspects obtaining F1-score higher than 80%. Nevertheless, they highlighted the need for further analysis to improve individual aspect identification.

Most classification systems for reviews tend to rely on binary classification, which often oversimplifies the content by categorizing reviews into only one aspect [7]. This method is ineffective, as many reviews can discuss multiple aspects, such as product quality and delivery service, simultaneously [8]. Multilabel classification enables each review to be categorized into multiple aspects, reflecting the reality that reviews often discuss more than one topic [9]. This

approach ensures that each aspect mentioned in the review is accurately identified. Although multilabel classification is essential for analyzing customer feedback, its application to Indonesian e-commerce reviews remains relatively unexplored, despite the substantial growth of the e-commerce industry in Indonesia. This is particularly noticeable when classifying reviews based on multiple aspects, which could provide deeper insights into customer opinions and experiences.

To address these issues, we propose a multilabel classification method for Indonesian e-commerce reviews, categorizing them into three key aspects: “product”, “customer service", and “shipping”. Product reviews highlight the qualities and features that customers appreciate or dislike [10], customer service reviews aim to enhance the customer interactions [11], and delivery reviews provide insights into reliability and timeliness [12]. For this purpose, we utilize a variant of Bidirectional Encoder Representation from Transformers (BERT) known as IndoBERT. While general transformer models like RoBERTa and XLNet have shown success in various natural language processing (NLP) tasks, they are primarily optimized for multilingual or English-based datasets [13][14]. IndoBERT, specifically trained on Indonesian language corpora, offers distinct advantages for Indonesian e-commerce datasets. This model makes it effective for multilabel classification tasks, enabling it to analyze customer reviews more accurately rather than those that are not designed for this language. IndoBERT has been used by Nissa and Yulianti [15] to perform multilabel aspect classification on hotel customer reviews in Indonesia. They used the CNN-XGBoost classification method, and the results showed that the model was effective with a micro F1-score reaching 0.899 for the case. They also compared IndoBERT with the multilingual BERT (m-BERT, distil-BERT, and XLM-RoBERTa), and they found that IndoBERT is slightly more accurate than the multilingual BERT. Its understanding of Indonesian text helps identify different aspects of reviews, which is important for classifying the varied feedback on e-commerce platforms.

In our research, the IndoBERT model is fine-tuned using Indonesian review datasets from various well-known e-commerce platforms. In the classification process, we focus on comparing the performance of CNN, LSTM, and the neural network architecture inherent to IndoBERT itself. By systematically evaluating end-to-end IndoBERT, IndoBERT-CNN, and IndoBERT-LSTM, we aim to identify the most effective approach for multilabel aspect classification. Despite some limitations, such as difficulties with mixed languages and substantial computational requirements for larger datasets [5], the model could contribute significantly to enhancing the aspect detection in e-commerce reviews, leading to improved customer insights and service quality.

2. Research Method

This research applies a multilabel classification technique to determine aspect categories from customer reviews. Multilabel classification is different from single-label classification. Single-label classification assigns only one label to a sample from a set of labels. However, in multilabel classification, a sample can have more than one label [16]. For example, for the multilabel aspect classification in this case, one review can be labeled as “product” and “customer service”, even “product”, “customer service”, and “shipping”. The system design of this study is depicted in Figure 1.

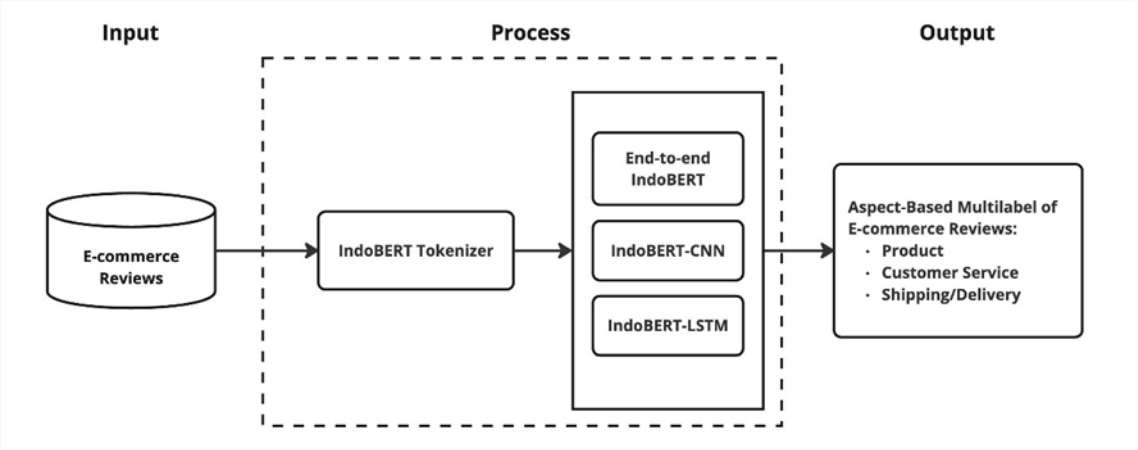


Figure 1. System Design

The system takes e-commerce reviews as input, which are then processed using the IndoBERT tokenizer to convert text into numerical representations. These numerical representations are subsequently fed into three different classification models: End-to-end IndoBERT, IndoBERT-CNN, and IndoBERT-LSTM. The output of each model is a set of predicted labels indicating the relevant aspects present in the review, such as "product," "customer service," and "shipping " or a combination of them.

2.1 Data Collection

The data in this study is a collection of reviews from various e-commerce products available on the Kaggle platform. However, the dataset still does not have a label or ground truth. To address this, we conducted a survey involving 81 respondents with diverse backgrounds who are active e-commerce users. In this crowdsourcing approach, each review was labeled by multiple respondents. To determine the final label for each review, we implemented a voting system where the label selected by the majority of respondents was considered the ground truth. This method leverages the collective wisdom of multiple individuals to ensure the accuracy and reliability of the labels. The attributes of the review aspects used in this study are shown in Table 1.

Table 1. Attributes of the E-commerce Review Aspect

Aspect	Description
Product	Customer satisfaction with the quality, performance, and conformity of the product to the description given
Customer service	Interaction between customers and sellers, friendliness and speed of response from sellers, and handling complaints.
Shipping	Shipping speed, condition of goods when received, and timeliness of shipping

2.2 IndoBERT

IndoBERT is a BERT model trained using Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) pre-training techniques on Indonesian datasets. The existence of IndoBERT has opened up new possibilities for understanding the complex Indonesian language, allowing various NLP applications such as sentiment analysis, text processing, and information extraction to become more sophisticated and reliable [17]. This research uses the IndoBERT pre-trained model with a new benchmark called Indonesian NLP (IndoNLU) proposed by [18]. The model was trained using the Indo4B dataset on more than 4 billion words with 250 million sentences. Besides being trained on a wider dataset and benchmark, this model also provides better performance than the previous IndoBERT model that used IndoLEM as a benchmark [19]. This study implements fine-tuned IndoBERT from the IndoNLU benchmark.

In BERT models including IndoBERT, before providing input, the raw text will go through a tokenization process using IndoBERT Tokenizer. In this tokenization process, special marks such as “[CLS]” and “[SEP]” will be added [20]. These special cues such as instructions to start reading (for [CLS]) and stop reading (for [SEP]) allow the BERT model to focus on each sentence. Furthermore, the sentences will be converted into word tokens, such as breaking a long sentence into individual words [5]. To understand the meaning of the words, IndoBERT has a specialized dictionary which is generated using the Sentence-Piece method. For words that have never been encountered before or new words, they will be broken down into smaller parts to understand their meaning. Next, the labeled review data will be converted into numeric vectors using word embedding from IndoBERT [21].

Then, these numerical vectors will be input into the IndoBERT model for further processing. In this stage, the model will perform a forward pass through the Transformer layers consisting of self-attention and feed-forward neural network. After passing through all Transformer layers, the model will produce two main outputs that are often used in classification tasks, namely the last hidden state and the pooled output.

2.3 Classifier

For the aspect classification stage, three different model architectures were implemented: the end-to-end IndoBERT, IndoBERT-CNN, and IndoBERT-LSTM models. The IndoBERT model was utilized in an end-to-end, serving as both the embedding generator and the classifier. These models were applied to a multilabel aspect classification dataset of e-commerce reviews. Specifically, the model used was indobert-base-p1 [18] from the Indo Benchmark project that can be accessed here [22]. The fine-tuning process adapted the model to the specific dataset, enabling it to effectively capture the unique linguistic features present in the data.

In the second model, the embedding from IndoBERT is combined with a Convolutional Neural Network (CNN) as the classifier. CNN is a type of neural network that is particularly well-suited for processing data that has a grid-like structure, such as images [23]. In this case, the CNN is used to process the embedding from IndoBERT, which is a fixed-size vector representation of the input text. The CNN can learn complex patterns in the embedding that are relevant to multilabel text classification. As in the first model, the embedding used comes from the same IndoBERT model, called indobert-base-p1. The result of embedding which is pooled output is then used as input to the CNN. The architecture of the IndoBERT and CNN model in this study is shown in Figure 2.

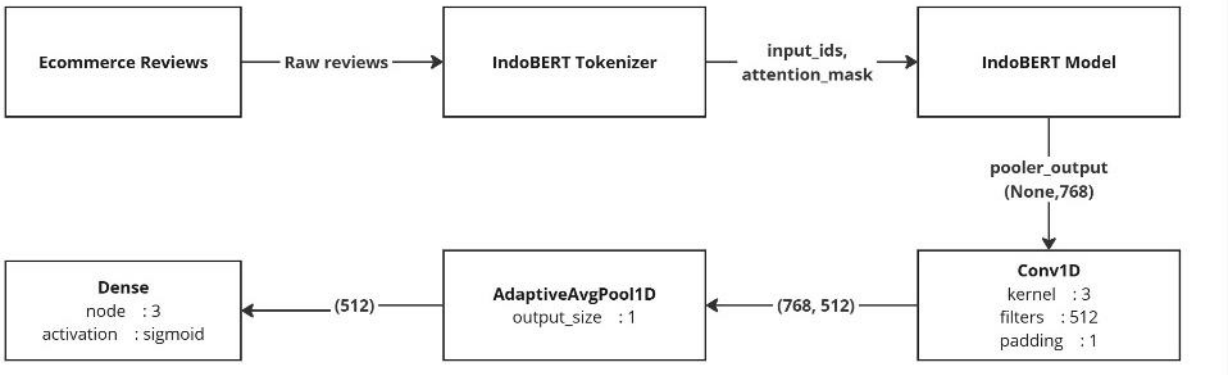


Figure 2. The Architecture of IndoBERT-CNN Model

The CNN method used in this model is set with certain parameters detailed in Table 2, including kernel size, number of filters, batch size, and learning rate.

Table 2. Setting the Parameters of the CNN

Parameter	Value
Kernel Size	3
Filter Size	512
Batch Size	32
Learning Rate	3e-5
Optimizer	Adam

Finally, in the third model, the same IndoBERT embedding is used, but this time with Long Short-Term Memory (LSTM) as the classifier. LSTM is a type of Recurrent Neural Network (RNN) that is effective in handling sequence data and can remember long-term information, which is very useful in text analysis [24]. IndoBERT provides a deep semantic understanding of the text, while LSTM captures the order and relationships between words. This combination allows our model to accurately classify text based on both its meaning and structure. The pooled output from IndoBERT is used as input to the LSTM, which then learns to classify aspects based on the order and context of the words. The architecture of the IndoBERT and LSTM model in this study is shown in Figure 3.

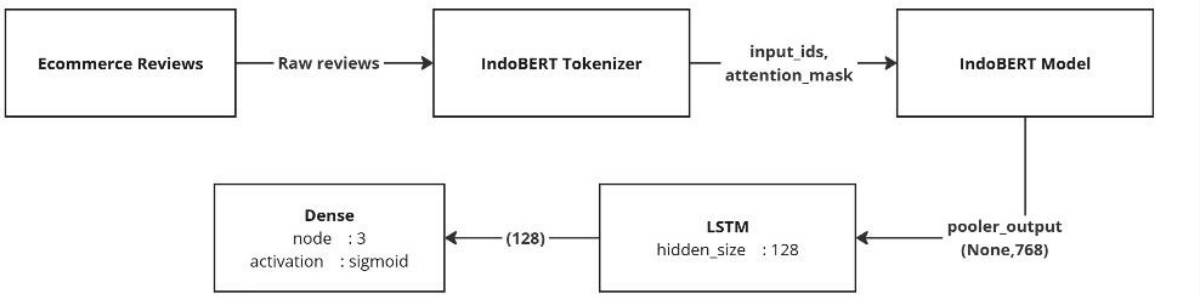


Figure 3. The Architecture of IndoBERT-LSTM Model

The parameters for this LSTM model are set according to those listed in Table 3, including the number of LSTM units, batch size, learning rate, and other relevant parameters to optimize the model performance.

Table 3. Setting Parameters of LSTM

Parameter	Value
Node	128
Batch Size	32
Learning Rate	3e-5
Optimizer	Adam

This classification model produces an output with three main labels: product, customer service, and shipping. To handle the multilabel case, thresholding was used on the three class outputs. The threshold set for each class output

is 0.3 to 0.99. This means that if the output value of each class exceeds the threshold value, it is considered 1 otherwise it is considered 0. The choice of this range is based on the need to balance the precision and recall, which is critical in multilabel classification tasks. Lower threshold values, like 0.3, help the model find more examples of a label, which can improve recall. This is helpful for imbalanced labels because it ensures that less common labels, like customer service or shipping, are not missed. On the other hand, higher thresholds (such as 0.99) only count very confident predictions, which improves precision. This range was chosen to strike a balance, allowing the model to detect both common and rare labels without being too strict or too loose in the classification.

The application of thresholding is important to address the possible imbalance in the number of labels observed [25]. The selection of this threshold is based on experiments with a threshold value close to 0 and a threshold value close to 1. A lower threshold value allows the model to accept more true values, whereas a higher threshold value makes the model more stringent [26]. With this approach, the model can more accurately identify the relevant classes from the prediction results, ensuring a more precise interpretation of the classification result data from our model.

3. Results and Discussion

The data used in this study consists of 4,002 customer review texts that have been labeled. As explained in Section 2, labeling is done using crowdsourcing techniques. Following the multilabel concept, each review can have only one label, two labels, and up to three labels at once. Examples of reviews that have one label can be seen in Table 4, while reviews with two and three labels are shown in Table 5 and Table 6, respectively. Each table consists of three columns of labels, where a value of 1 indicates that the review sentence belongs to that label, while a value of 0 indicates that it does not.

Table 4. Sample of E-commerce Review with 1 Label

Customer Reviews	Label		
	Product	Customer service	Shipping
"Barang bagus tapi analog kiri sama kanan kalo di setting jadi sama terus selain itu bagus" *("Good stuff but the left and right analog in the setting is always the same other than that it's good.")*	1	0	0
"Penjual yang direkomendasikan respons cepat" *("Recommended seller fast response")*	0	1	0
Barang sudah sampai tujuan dalam kondisi baik proses Shipping sangat cepat terima kasih *("The item has arrived in good condition Shipping process is very fast thank you")*	0	0	1

Table 5. Sample of E-commerce Reviews with 2 Labels

Customer Reviews	Label		
	Product	Customer service	Shipping
"Pelayanan ok barang ok berfungsi baik thanks" *("Service ok goods ok functioning well thanks")*	1	1	0
bagus Shipping cepat dan fast respon sukses teruz gan *("good fast shipping and fast response success always bro")*	0	1	1
"Barang bagus dan barang cepat sampai" *("Good item and fast delivery")*	1	0	1

Table 6. Sample of E-commerce Reviews with 3 Labels

Customer Reviews	Label		
	Product	Customer service	Shipping
"Product bagus tombol2 empuk Respon cepat Shipping cepat & packing bagus Recommended seller." *("Good product soft buttons Quick response Fast shipping & good packaging Recommended seller.")*	1	1	1

From the labeling results, it is found that as many as 4,002 review data have product labels, 1,173 review data have customer service labels, and as many as 2,262 review data have shipping labels. The distribution of each label as a whole is presented in Table 7.

Table 7. Distribution of Overall Aspect

Label	Total
Product	2.902
Customer service	1.173
Shipping	1.740

As for the multilabel aspect, e-commerce customer review data consists of 1,576 data including product labels, 297 data including customer service labels, 582 data including shipping labels, 288 data including “product” and “customer service”, 670 data including “product” and “shipping”, 220 data including “customer service” and “shipping”, and 268 data including “product”, “customer service”, and “shipping”. Details of the multilabel distribution can be seen in Table 8.

Table 8. Distribution of Each Aspect Labels

Label	Total
Product	1.576
Customer service	297
Shipping	582
Product, Customer Service	288
Product, Shipping	670
Customer service, Shipping	220
Product, Customer service, Shipping	268

Based on the distribution of the dataset in Table 8, it is clear that the data for each label is imbalanced, with some labels showing significant differences. This study evaluates the model's performance on this imbalanced data. Additionally, this study examines model performance on balanced data by applying undersampling, resulting in a total of 1500 data points, with 214 instances per label. This approach was used to assess the impact of data imbalance on model performance. The balanced data results are included to provide additional insight into how the model behaves when trained on imbalanced versus balanced datasets. This comparison helps illustrate the extent to which data imbalance affects model performance, supporting the main goal of comparing the effectiveness of the classification methods.

As explained in the research methods section, three models were used: end-to-end IndoBERT, IndoBERT-CNN, and IndoBERT-LSTM. This study used data with a proportion of 80% for training and 20% for testing. The performance of the three models was evaluated using precision, recall, and F1-score metrics. Given the accuracy limitations in multilabel classification, this research does not employ the accuracy metric. Additionally, it explores various threshold value schemes that have been previously described. The performance of all models on the imbalanced dataset is shown in Table 9.

Table 9. Model Evaluation End-to-end IndoBERT, IndoBERT-CNN, IndoBERT-LSTM

Model	Metric	Threshold							
		0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.99
End-to-end IndoBERT	Precision	0.545	0.545	0.545	0.964	0.972	0.0	0.0	0.0
	Recall	1.0	1.0	1.0	0.935	0.921	0.0	0.0	0.0
	F1-score	0.689	0.689	0.689	0.949	0.946	0.0	0.0	0.0
IndoBERT-CNN	Precision	0.559	0.559	0.559	0.963	0.967	0.0	0.0	0.0
	Recall	1.0	1.0	1.0	0.937	0.933	0.0	0.0	0.0
	F1-score	0.697	0.697	0.697	0.949	0.949	0.0	0.0	0.0
IndoBERT-LSTM	Precision	0.546	0.547	0.546	0.962	0.772	0.0	0.0	0.0
	Recall	1.0	1.0	1.0	0.952	0.506	0.0	0.0	0.0
	F1-score	0.690	0.690	0.690	0.957	0.605	0.0	0.0	0.0

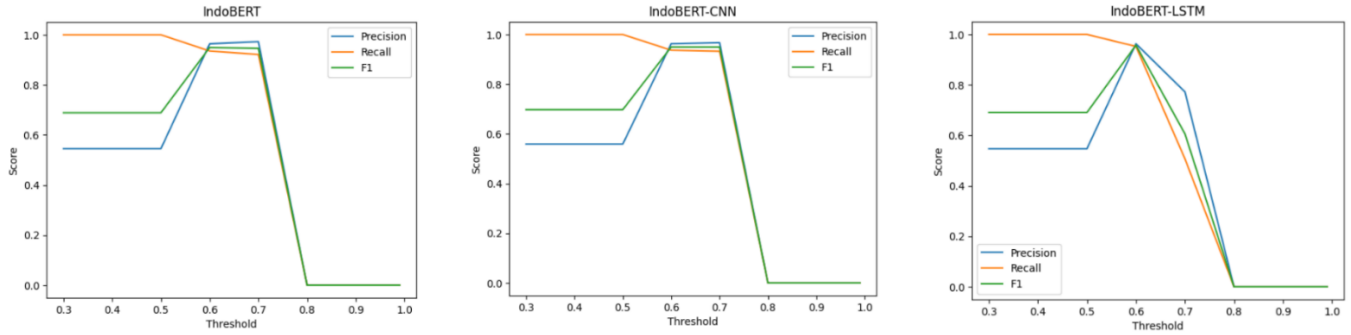
According to Table 9, all models show poor performance at thresholds between 0.3 and 0.5, with recall consistently at 1.0 while precision hovers around 0.5. This suggests that the models are over-predicting labels, classifying many instances as positive even when the corresponding label is not present in the review. In contrast, at higher thresholds (0.8 to 0.99), precision, recall, and F1-score drop sharply to 0.0, as the models fail to predict any labels. The imbalance in label distribution and frequency makes it increasingly difficult for the models to maintain both precision and recall, as they tend to predict the most frequent labels seen during training. In this case, the "product" label dominates the dataset, causing the model to over-predict it, while rarely occurring labels are left unpredicted. This results in no true positives (TP), causing all metrics to fall to 0. At thresholds between 0.6 and 0.7, the models perform

more evenly in terms of recall and precision, as reflected in F1-scores, which are mostly above 0.9. At these moderate thresholds, the balance between precision and recall improves, allowing the model to capture a sufficient number of true positives (TP) without over-predicting classes.

Table 10. Evaluation of End-to-end IndoBERT, IndoBERT-CNN, IndoBERT-LSTM on Balance Data (Undersampling)

Model	Metric	Threshold							
		0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.99
End-to-end IndoBERT	Precision	0.513	0.513	0.514	0.904	0.925	0.0	0.0	0.0
	Recall	1.0	1.0	1.0	0.912	0.876	0.0	0.0	0.0
	F1-score	0.669	0.669	0.669	0.906	0.898	0.0	0.0	0.0
IndoBERT-CNN	Precision	0.537	0.537	0.537	0.906	0.913	0.0	0.0	0.0
	Recall	1.0	1.0	1.0	0.921	0.891	0.0	0.0	0.0
	F1-score	0.687	0.687	0.687	0.913	0.901	0.0	0.0	0.0
IndoBERT-LSTM	Precision	0.517	0.517	0.517	0.919	0.972	0.0	0.0	0.0
	Recall	1.0	1.0	1.0	0.938	0.390	0.0	0.0	0.0
	F1-score	0.669	0.669	0.669	0.928	0.526	0.0	0.0	0.0

Due to the imbalanced label distribution in the dataset, an experiment was conducted to balance the labels by applying an undersampling technique. This was done to observe the model's behavior on a balanced dataset. Table 10 presents the model's performance on the balanced data, which shows no significant improvement compared to its performance on the original data. The model's performance appears to be slightly lower. This occurs because undersampling reduces the amount of training data, which can limit the model's ability to learn from a diverse range of examples. Although undersampling helps balance label distribution, it also removes potentially valuable information, resulting in decreased overall model performance.



(a) (b) (c)

Figure 4. Model Performance by Threshold (a) End-to-end IndoBERT, (b) IndoBERT-CNN, (c) IndoBERT-LSTM

To gain a better understanding of the performance trends of all models across different thresholds, refer to Figure 4. The majority of models demonstrate optimal performance when the threshold values are set to 0.6 and 0.7. The end-to-end IndoBERT in Figure 4 (a), IndoBERT-CNN in Figure 4 (b), and IndoBERT-LSTM in Figure 4 (c) all show a significant increase in performance at 0.6 threshold, followed by stable performance at 0.7 threshold. However, IndoBERT-LSTM behaves differently, as seen in Figure 4 (c), where its performance declines at the 0.7 threshold. Several factors may explain this decline, including IndoBERT-LSTM's high sensitivity to threshold changes, overfitting on training data, and the complexity of handling data sequences, which makes decision-making more challenging. In contrast, the end-to-end IndoBERT and IndoBERT-CNN models perform better at the 0.7 threshold. IndoBERT, with its end-to-end architecture, and IndoBERT-CNN, utilizing convolutions to capture patterns, tend to be more stable and effective at higher thresholds.

Looking at the overall performance at each threshold shown in Figure 4, it can be concluded that 0.6 is the best threshold value for all models. At 0.6 threshold, the end-to-end IndoBERT model achieved an F1-score of 94.9%. The IndoBERT-CNN model also achieved its best performance with the same F1-score. Likewise, the IndoBERT-LSTM model provides significant results at an F1-score of 95.7%. Using a smaller threshold value or below 0.3 the model will be looser in performing class selection, allowing more false predictions to be generated. On the contrary, using a larger threshold value, such as 0.8 and above, allows the model to not easily pass the class so that fewer classes can be predicted. The performance of the models at 0.6 threshold is shown more specifically in Table 11.

Table 11. Model Performance at Threshold Value 0.6

Model	F1-score
End-to-end IndoBERT	0.949
IndoBERT-CNN	0.949
IndoBERT-LSTM	**0.957**

Based on the performance of all models at 0.6 threshold, the model with the best performance is IndoBERT-LSTM. IndoBERT-LSTM is recognized as the best-performing model, achieving high F1-score of 95.7%. This success may come from its ability to handle the sequential nature of review texts effectively. LSTM networks are good at remembering information from previous inputs, which helps them better understand text patterns. This capability allows the model to make more accurate predictions about the aspects of reviews. However, the performance of all metrics decreases when the threshold value gets higher than 0.6. This happens because the higher the threshold value, the tighter the model selection makes it difficult to pass a class. On the contrary, the lower the threshold value, the looser the selection process becomes. While IndoBERT-LSTM stands out, it is important to note that all models perform well in multilabel aspect classification tasks. Although IndoBERT-LSTM has the highest F1-score, it is only slightly better than the performances of end-to-end IndoBERT and IndoBERT-CNN.

In this study, IndoBERT was utilized to effectively address multilabel aspect classification on an Indonesian e-commerce review dataset. By leveraging several deep learning methods as classifiers, it was found that LSTM yielded the most optimal results with the following hyperparameters: 128 nodes, a batch size of 32, a learning rate of 3e-5, and the Adam optimizer. The performance comparison of the methods in this study with other research on multilabel classification using Indonesian datasets is presented in Table 12.

Table 12. Performance Comparison Across Previous Studies

Study	Method	F1-score
In [15]	End-to-end IndoBERT	0.928
In [20]	IndoBERT embedding + MBERT	0.903
In [27]	FastText + CNN + Bi-LSTM	0.733
Proposed Method	**IndoBERT embedding + LSTM**	**0.957**

As shown in Table 12, the most optimal method proposed in this study is IndoBERT combined with LSTM, achieving the highest performance with an F1-score of 95.7%. This demonstrates that IndoBERT is particularly effective in handling multilabel classification tasks on Indonesian datasets.

4. Conclusion and Future Work

This research focuses on classifying e-commerce reviews into multilabel aspect classes using the fine-tuned IndoBERT model. By combining IndoBERT with various classifiers, including Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM), this study explores the effectiveness of different approaches for analyzing customer reviews. The dataset comprises 4,002 labeled entries, covering aspects such as product quality, customer service, shipping, and their combinations.

The experiments evaluated the model performance on both imbalanced original data and balanced data, which was created through an undersampling process. With 1,500 instances in the balanced dataset, the research demonstrates that larger imbalanced datasets tend to yield better performance compared to smaller balanced ones. The results indicate that the models perform optimally at 0.6 threshold, with the IndoBERT-LSTM achieving the highest F1-score of 95.7%, while both end-to-end IndoBERT and IndoBERT-CNN models reached F1-scores of 94.9%. Although all models show satisfactory performance, especially at higher thresholds, their effectiveness decline at lower thresholds.

The classification performance of these models plays a vital role in extracting precise information from customer reviews, offering valuable insights for e-commerce businesses. This capability can help identify aspects that require improvement, ultimately enhancing customer satisfaction and trust. In future research, improvements can be carried out on the exploration of more specific labels to deepen the analysis of customer reviews, such as utilizing more detailed aspects like product quality, response speed, item condition, price, or delivery accuracy. In addition, the analysis can be improved into sentiments, such as positive, negative, and neutral to more accurately understand customer perceptions of various aspects on the services provided. Therefore, the analysis results obtained can be deeper and broader, and capture all information contained in the reviews.

Acknowledgement

We are sincerely grateful for the valuable contributions of Universitas Islam Negeri Maulana Malik Ibrahim, the respondents who participated, as well as other parties who have supported and encouraged this research.

References

[1] T. K. Lestari, A. L. Kusumatrisna, and A. Syakilah, *Statistik E-commerce 2021*. BPS-Statistics Indonesia, 2021.

[2] T. Chen, P. Samaranayake, X. Y. Cen, M. Qi, and Y. C. Lan, "The Impact of Online Reviews on Consumers' Purchasing Decisions: Evidence from an Eye-Tracking Study," *Front Psychol*, vol. 13, Jun. 2022. https://doi.org/10.3389/fpsyg.2022.865702

[3] D. Yanti, N. Kristya Ningsih, J. G. Ony, and S. P. Suhalim, "The Influence of Online Customer Reviews and Online Customer Ratings on Product Purchase Decisions on The Tokopedia Marketplace," *Jurnal Pemasaran Kompetitif*, vol. 07, no. 2, p. 2024. https://doi.org/10.32493/jpkpk.v7i2.40083

[4] Y. Gurav, V. Ingawale, and A. Yadav, "A Study on The Impact Of Online Product Reviews On Consumers' Buying Intentions," *The Online Journal of Distance Education and e-Learning*, vol. 11, no. 2, pp. 1924–1928, 2023.

[5] S. Imron, E. I. Setiawan, and J. Santoso, "Deteksi Aspek Review E-Commerce Menggunakan IndoBERT Embedding dan CNN," *Journal of Intelligent System and Computation*, vol. 5, no. 1, pp. 10–16, Apr. 2023. https://doi.org/10.52985/insyst.v5i1.267

[6] D. F. Nasiri and I. Budi, "Aspect Category Detection on Indonesian E-commerce Mobile Application Review," in *2019 International Conference on Data and Software Engineering (ICoDSE)*, Institute of Electrical and Electronics Engineers Inc., Nov. 2019. https://doi.org/10.1109/ICoDSE48700.2019.9092619

[7] A. Lunardi, J. Viterbo, C. Boscarioli, F. Bernardini, and C. Maciel, "Domain-tailored Multiclass Classification of User Reviews based on Binary Splits," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 2016, pp. 298–309. https://doi.org/10.1007/978-3-319-39910-2_28

[8] E. Deniz, H. Erbay, and M. Coşar, "Multi-Label Classification of E-Commerce Customer Reviews via Machine Learning," *Axioms*, vol. 11, no. 9, p. 436, Sep. 2022. https://doi.org/10.3390/axioms11090436

[9] G. Khlifi, I. Jenhani, M. Ben Messaoud, and M. W. Mkaouer, "Multi-label Classification of Mobile Application User Reviews Using Neural Language Models," in *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, Z. Bouraoui and S. Vesic, Eds., Cham: Springer Nature Switzerland, 2024, pp. 417–426. https://doi.org/10.1007/978-3-031-45608-4_31

[10] M. Abubakar, A. Shahzad, H. Abbasi, and I. Abbottabad Campus Pakistan, "Aspect-Based Sentiment Analysis on Amazon Product Reviews," *International Journal of Informatics Information System and Computer Engineering*, vol. 2, no. 2, pp. 206–211, 2021. https://doi.org/10.34010/injiiscom.v2i2.7455

[11] G. Liu, S. Fei, Z. Yan, C. H. Wu, and S. B. Tsai, "An Empirical Study on Response to Online Customer Reviews and E-Commerce Sales: From the Mobile Information System Perspective," *Mobile Information Systems*, vol. 2020, no. 1, 2020. https://doi.org/10.1155/2020/8864764

[12] P. Ravula, "Impact of Delivery Performance on Online Review Ratings: The Role of Temporal Distance of Ratings," *Journal of Marketing Analytics*, vol. 11, no. 2, pp. 149–159, Jun. 2023. https://doi.org/10.1057/s41270-022-00168-5

[13] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019. https://doi.org/10.48550/arXiv.1907.11692

[14] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," Jun. 2019. https://doi.org/10.48550/arXiv.1906.08237

[15] N. K. Nissa and E. Yulianti, "Multi-label Text Classification of Indonesian Customer Reviews Using Bidirectional Encoder Representations from Transformers Language Model," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 5, pp. 5641–5652, Oct. 2023. https://doi.org/10.11591/ijece.v13i5.pp5641-5652

[16] I. Akbar, M. Faisal, and T. Chamidy, "Multi-label Classification of Indonesian Qur'an Translation using Long Short-Term Memory Model," *Computer Network, Computing, Electronics, and Control Journal*, vol. 9, no. 2, pp. 119–128, 2019. https://doi.org/10.22219/kinetik.v9i2.1901

[17] J. Forry Kusuma and A. Chowanda, "Indonesian Hate Speech Detection Using IndoBERTweet and BiLSTM on Twitter," *JOIV : International Journal on Informatic Visualization*, vol. 7, no. 3, pp. 773–780, 2023. https://dx.doi.org/10.30630/joiv.7.3.1035

[18] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," Sep. 2020. https://doi.org/10.18653/v1/2020.aacl-main.85

[19] B. V. Kartika, M. J. Alfredo, and G. P. Kusuma, "Fine-Tuned IndoBERT based model and data augmentation for indonesian language paraphrase identification," *Revue d'Intelligence Artificielle*, vol. 37, no. 3, pp. 733–743, Jun. 2023. https://doi.org/10.18280/ria.370322

[20] G. Z. Nabiilah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, "Indonesian multilabel classification using IndoBERT embedding and MBERT classification," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 1, pp. 1071–1078, Feb. 2024. https://doi.org/10.11591/ijece.v14i1.pp1071-1078

[21] H. Tanaka, H. Shinnou, R. Cao, J. Bai, and W. Ma, "Document Classification by Word Embeddings of BERT," in *Communications in Computer and Information Science*, Springer, 2020, pp. 145–154. https://doi.org/10.1007/978-981-15-6168-9_13

[22] Hugging Face, "IndoBERT: indobenchmark/indobert-base-p1." Accessed: Oct. 14, 2024.

[23] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, "A Review of Convolutional Neural Networks in Computer Vision," *Artif Intell Rev*, vol. 57, no. 4, Apr. 2024. https://doi.org/10.1007/s10462-024-10721-6

[24] W. K. Sari, D. P. Rini, and R. F. Malik, "Text Classification Using Long Short-Term Memory With GloVe Features," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 5, no. 2, p. 85, Feb. 2020. https://doi.org/10.26555/jiteki.v5i2.15021

[25] Z. Xingfu, H. Gweon, and S. Provost, "Threshold Moving Approaches for Addressing the Class Imbalance Problem and their Application to Multi-label Classification," in *Proceedings of 2020 the 4th International Conference on Advances in Image Processing (ICAIP 2020)*, 2020, pp. 72–75. https://doi.org/10.1145/3441250.3441274

[26] C. Murphy, J. A. Tawn, and Z. Varty, "Automated Threshold Selection and Associated Inference Uncertainty for Univariate Extremes," Oct. 2023. https://doi.org/10.1080/00401706.2024.2421744

[27] A. Rofiqul Muslikh, I. Akbar, D. Rosal Ignatius Moses Setiadi, and H. Md Mehedul Islam, "Multi-label Classification of Indonesian Al-Quran Translation based CNN, BiLSTM, and FastText," *Techno.COM*, vol. 23, no. 1, pp. 37–50, 2024. https://doi.org/10.62411/tc.v23i1.9925