Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2026. Vol. 5. No. 2; e-ISSN: 2808-4519

Mitigating Data Sparsity in Code-Mixed Text through Back-Translation Augmentation for Aspect-Based Sentiment Analysis in Tokopedia Reviews

Abdul Hakim Prima Yuniarto[1]*, Aini Shofi Achsanti[2], Rizky Agil Singgih Susanto[3], Ardena Afif Pratama[4], Fandy Setyo Utomo[5]

[1,2,3,4,5] *Magister Ilmu Komputer, Fakultas Ilmu Komputer, Universitas Amikom Purwokerto*
[1]*Prodi Teknik Elektro, Sekolah Tinggi Teknik Wiworotomo Purwokerto*
a.hakim.py@gmail.com[1*]

Abstract

Aspect-Based Sentiment Analysis (ABSA) in e-commerce reviews in Indonesia faces significant challenges, including the use of mixed language or code-mixed language and limited labeled data, or data sparsity. This study proposes the use of Back-Translation data augmentation techniques to enrich Tokopedia's mixed Indonesian-English or South Jakarta language review dataset. Using the IndoBERT model, experimental results show a 3% increase in accuracy for both aspect and sentiment classification. These findings demonstrate that artificial data augmentation is effective in addressing data sparsity constraints in informal texts and improving the reliability of macro analysis for strategic platform recommendations.

Keywords: *Aspect-Based Sentiment Analysis; Back-Translation; Code-Mixed; Data Sparsity; IndoBERT; Tokopedia*

1. Introduction

The rapid growth of the e-commerce industry in Indonesia has resulted in a surge in user reviews across various digital platforms. These reviews are valuable assets for companies to gain a deeper understanding of customer experiences [1]. As one of the e-commerce champions with the highest web traffic in Indonesia, Tokopedia faces a massive yet unstructured volume of review data [2]. Traditional sentiment analysis, which only classifies text into general polarities, often fails to capture customer opinions on specific elements such as price, product, or delivery. Therefore, Aspect-Based Sentiment Analysis (ABSA) has become a crucial tool because it can identify sentiment towards specific aspects of a review at a granular level [3].

A major challenge in implementing ABSA in Indonesia is the widespread use of mixed language, also known as code-mixing [4]. Consumers tend to combine Indonesian and English, often referred to as the "Jaksel" language, when providing feedback. This communication pattern has been widely found on various social media platforms such as X [5] and Facebook [6], where the insertion of English vocabulary is often used to demonstrate prestige or simply attract attention. For natural language processing (NLP) models, code-mixed text creates high syntactic complexity and triggers data sparsity issues, or the scarcity of high-quality training data to capture dynamic linguistic variations. Furthermore, NLP analysis in this domain often involves very large but sparse word matrices, which require efficient computational techniques [7].

Previous research has evaluated the performance of traditional algorithms such as Naive Bayes to classify Tokopedia customer comments [2]. While traditional methods can provide adequate results for simple classification, Transformer-based models like IndoBERT demonstrate significant advantages in understanding sentence context bidirectionally [8]. However, training deep learning models on small datasets risks predictive overfitting, where the model performs well on the training data but fails to generalize to real-world data [9].

To mitigate the scarcity of labeled data on mixed-language text, data augmentation techniques are crucial [10]. One effective method is Back-Translation, which involves translating text into an intermediary language and then translating it back into the source language to generate sentence variations without changing semantic meaning [11]. This study combines Back-Translation augmentation techniques with the IndoBERT architecture to handle the ABSA task on mixed-language Tokopedia reviews, with the aim of improving model accuracy and providing more accurate strategic insights for the Indonesian e-commerce industry.

2. Research Methods

This research methodology is systematically designed to transform unstructured text data into strategic insights through several key technical steps. The architecture prioritizes the handling of high-dimensional sparse matrices and the nuances of colloquial language patterns found in Indonesian digital commerce.

2.1. Raw Data Acquisition and Filtration

The dataset utilized in this study is a public dataset available on Kaggle, containing 65,000 Tokopedia product reviews submitted by customers throughout the year 2025. The selection of Tokopedia as the primary data source is justified by its status as one of the largest e-commerce platforms in Indonesia, generating a massive surge in user reviews that serve as a critical repository for customer experience data [1]. To focus the study on mixed-language challenges, a specialized keyword filtering method was employed. This process aimed to isolate reviews that exhibit "Jaksel" language patterns a sociolinguistic phenomenon where Indonesian is mixed with English terms to draw attention or demonstrate prestige.

The filtration stage is necessary because most available pre-trained language models struggle when dealing with code-mixed data characterized by inconsistent linguistic forms and colloquialisms [4]. By filtering for specific English transactional terms (e.g., quality, worth it, delivery, fast response), the dataset was narrowed down to 11,000 reviews verified as mixed-language text. This targeted approach ensures that the model is trained on the specific patterns of entertainment-based and transactional communication often observed on social platforms like X [5].

2.2. Manual Extraction and Labeling

From the 11,000 filtered reviews, a random sample of 378 reviews was extracted to serve as the initial "seed data" for manual annotation. This stage is critical as ABSA requires identifying specific aspect terms and their corresponding sentiment polarities [3]. The manual labeling process categorized each review into four primary aspects: Product (P), Logistics (L), Service (S), and Price (H), while simultaneously assigning a Positive or Negative sentiment label.

While manual annotation is considered the gold standard for creating high-quality datasets, it is often limited by the time and expertise required, leading to the "data sparsity" problem. Utilizing a small but high-quality seed dataset represents a common scenario in low-resource NLP tasks, where advanced augmentation techniques must be deployed to simulate larger datasets and improve model performance [10]. This granular labeling allows for a more detailed understanding of user satisfaction than traditional document-level analysis.

2.3. Back-Translation Data Augmentation

To address the data sparsity issue identified in the seed dataset, a Back-Translation augmentation technique was applied. This method utilizes Neural Machine Translation (NMT) to translate the original Indonesian-English code-mixed reviews into a pivot language (English) and then translate them back into the source language (Indonesian). Back-translation has been empirically proven to be a highly effective means of generating diverse linguistic variations while preserving the core semantic meaning and labels of the original text [11].

By expanding the dataset from 378 to 756 rows of training data, this study creates a more robust training environment that forces the model to learn abstract semantic features rather than over-fitting on specific keyword-label pairs. Preventing predictive overfitting is essential when working with deep learning models and limited data, as it ensures the model can generalize to new, unseen variations of user reviews [9]. This strategy of artificial data expansion is particularly crucial for low-resource languages like Indonesian, which remain underrepresented in standard NLP corpora.

2.4. IndoBERT Model Training

The core classification engine for this study is the indobert-base-p2 model. The IndoBERT architecture was selected over traditional machine learning methods, such as Naïve Bayes which often lacks the context-modeling capabilities required for complex sentences [2]. IndoBERT's primary advantage lies in its bidirectional encoder representations and its WordPiece Tokenizer, which can break down informal slang or out-of-vocabulary "Jaksel" terms into meaningful subwords [8].

The training process involved a comparative analysis between models trained on the original dataset versus those trained on the augmented dataset to quantify the impact of increased data volume. Furthermore, the methodology accounted for the high-dimensional nature of word embedding matrices, which are often sparse (containing many zero elements). Efficient handling of these sparse matrices is vital to maintain calculation speed and model interpretability in large-scale data mining tasks [7]. By fine-tuning IndoBERT on the enriched code-mixed dataset, the system achieves a state-of-the-art balance between performance and robustness.

3. Result and Discussion

3.1. Model Performance Evaluation on Blind Test Set

Testing was conducted on a blind test set comprising 20% of the manually labeled data. This approach ensures that the evaluation reflects the model's ability to generalize to new, unseen instances of code-mixed language. The experimental results, as summarized in the following tables, demonstrate that the integration of augmented data through back-translation yields consistent improvements across all primary evaluation metrics.

Table 1. Comparison of Aspect Classification Evaluation Metrics

Model	Precision	Recall	F1-Score	Accuracy
Original	0.74	0.71	0.72	0.72
Augmented	0.78	0.74	0.75	0.75

As shown in Table 1, the use of augmented data improved aspect classification accuracy by 3%. This improvement is significant considering the high-dimensional and sparse nature of word embeddings often encountered in natural language processing. The linguistic variation introduced by back-translation helps the model better capture the nuances of mixed languages, which are characterized by inconsistent linguistic forms and colloquialisms [4]. Compared to traditional machine learning baselines such as Naïve Bayes, which previously achieved high accuracy in simpler binary sentiment tasks on Tokopedia [2], the IndoBERT-based model provides a more granular understanding of specific service attributes.

Table 2. Comparison of Sentiment Classification Evaluation Metrics

Model	Precision	Recall	F1-Score	Accuracy
Original	0.94	0.93	0.93	0.93
Augmented	0.96	0.96	0.96	0.96

The results for sentiment classification as shown in Table 2 further validate the effectiveness of the proposed method. A 3% increase in sentiment accuracy confirms that the augmentation technique effectively fill the gaps in the feature space, which is often a challenge in low-resource settings. This data enrichment is crucial in preventing predictive overfitting, a common pitfall when training deep learning models on limited labeled datasets [9]. By exposing the IndoBERT model to various paraphrased versions of the same sentiment, the model is forced to learn more abstract and robust semantic features rather than simply memorizing specific keywords [10].

3.2. Large-Scale Analysis and Statistical Validation

The final model was implemented on 10,641 new reviews (unseen data) for automated information extraction. Prediction reliability was validated by calculating the average confidence score. To test the real-world utility of the model, it was implemented on 10,641 new, unseen reviews. This large-scale inference provides a "stress test" for the model's generalizability across diverse user writing styles.

Table 3. Confident Score

Category	Confident Score average
Aspect	0.8598
Sentiment	0.9954

The confidence scores presented in Table 3 serve as a proxy for prediction reliability. The near-perfect score for sentiment (0.9954) suggests that the polarities of Indonesian-English mixed reviews are highly discernible for the IndoBERT model. The slightly lower score for aspect (0.8598) reflects the inherent ambiguity of aspect terms in unstructured text, where a single word might refer to multiple attributes depending on the context [3]. Nevertheless, these scores prove that the model maintains high reliability even when dealing with real-world data outside the original seed data.

3.3. Strategic Findings and Recommendations

A macro-level analysis of the 10,641 reviews provides a comprehensive overview of customer behavior and satisfaction on Tokopedia, which can be visualized through the following figures.

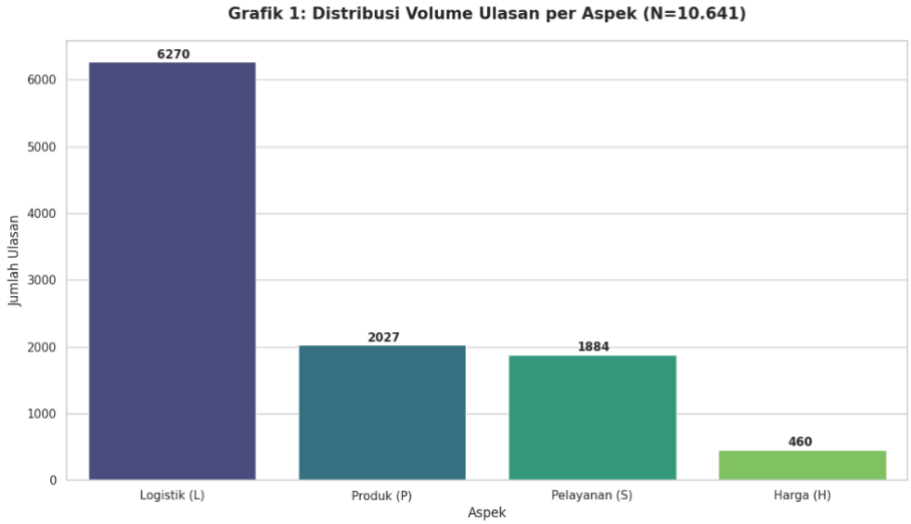

Fig. 1: Distribution of Review Volume per Aspect

As illustrated in Figures 1, the dominant issue discussed by users is the Logistics (L) aspect. This finding aligns with broader e-commerce trends in Indonesia, where delivery speed and courier reliability are paramount to the customer experience. The high volume of discussions in this category suggests that Tokopedia's logistical infrastructure is a primary touchpoint for user engagement, both positive and negative.

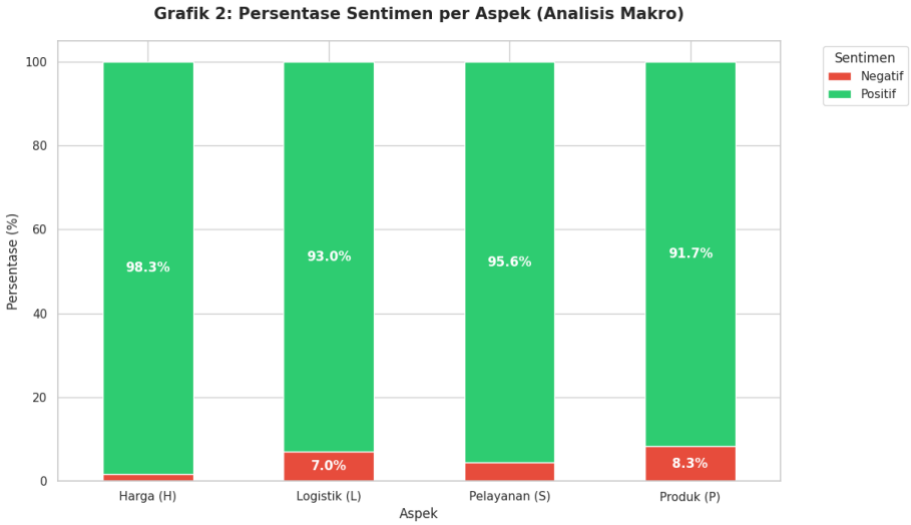

Fig. 2. Percentage of Sentiment for each Aspect (Macro Analysis)

The sentiment distribution across aspects, as shown in Gambar 2, offers critical business insights. The exceptionally high satisfaction in the Price (98.3% positive) and Service (95.6% positive) aspects indicates the success of current promotional strategies, discounts, and customer support protocols. However, the Product (P) aspect recorded the highest negative sentiment at 8.3%. This disparity highlights a "quality gap" where users are satisfied with the transaction and delivery but sometimes disappointed with the physical item received. For management, this serves as a clear signal to strengthen merchant quality control and product verification processes to protect the platform's reputation.

3.4. Advanced Optimization and Overfitting Prevention

Despite the 3% improvement, achieving even higher accuracy in aspect classification (currently 75%) remains a challenge. Future optimizations should consider even lower learning rates to preserve the deep linguistic knowledge captured in the pre-trained IndoBERT weights. Additionally, as datasets scale, the use of advanced mathematical tools like Sparse Principal Component Analysis could help in managing the computational efficiency of high-dimensional matrices.

Finally, the role of data augmentation as a regularizer cannot be understated. By preventing the model from over-fitting on specific lexical cues, the combination of back-translation and Early Stopping ensures that the system remains adaptable to the ever-evolving slang and code-mixed patterns of Indonesian netizens. This robustness is essential for maintaining the accuracy of ABSA as a long-term strategic tool for e-commerce platforms.

4. Conclusion

This study concludes that data augmentation using the Back-Translation method empirically successfully addresses the data sparsity issue in Tokopedia's mixed-language reviews. The implementation of the IndoBERT model demonstrates significant superiority in understanding informal linguistic patterns without a complex dictionary normalization stage. A 3% accuracy improvement on the ABSA task demonstrates that the addition of relevant synthetic data strengthens the model's robustness and prevents overfitting. Strategically, the macro analysis results provide a strong basis for recommendations for e-commerce platforms to maintain price competitiveness while prioritizing product quality control improvements to maintain long-term customer loyalty.

Acknowledgement

Thanks are due to the Master of Computer Science Study Program, Faculty of Computer Science, Amikom University, Purwokerto, which has facilitated and provided support, so that this research can be completed well and smoothly.

References

[1] M. M. Pakpahan, M. Halmi Dar, and M. Nirmala Sari Hasibuan, "Performance Evaluation of Machine Learning Algorithms in Aspect-Based Sentiment Analysis on E-Commerce User Reviews," *Int. J. Sci. Technol. Manag.*, vol. 6, no. 4, pp. 958–965, 2025.

[2] S. M. Salsabila, A. Alim Murtopo, and N. Fadhilah, "Analisis Sentimen Pelanggan Tokopedia Menggunakan Metode Naïve Bayes Classifier," *J. Minfo Polgan*, vol. 11, no. 2, pp. 30–35, 2022.

[3] G. Tripathy and A. Sharaff, "Traversing the landscape of aspect-based sentiment analysis: Delving deeper into techniques, trends, and future directions," *Comput. Sci. Rev.*, vol. 60, no. November 2025, 2026.

[4] A. F. Hidayatullah, R. A. Apong, D. T. C. Lai, and A. Qazi, "Pre-trained language model for code-mixed text in Indonesian, Javanese, and English using transformer," *Soc. Netw. Anal. Min.*, vol. 15, no. 1, pp. 1–17, 2025.

[5] L. A. Ridhawati, A. R. Firdhani, and M. N. Assyddyq, "Indonesian-English Code-Mixing in Entertainment-Based Communication on X," *Lang. Teach. Learn. Linguist. Lit.*, vol. 13, no. 2, pp. 5582–5592, 2025.

[6] A. A. F. Zalukhu, R. E. Laiya, and M. Y. Laia, "ANALYSIS OF INDONESIAN-ENGLISH CODE SWITCHING AND CODE MIXING ON FACEBOOK," *Res. English Lang. Educ.*, vol. 3, no. 2, pp. 1–10, 2021.

[7] R. Drikvandi and O. Lawal, "Sparse Principal Component Analysis for Natural Language Processing," *Ann. Data Sci.*, vol. 10, no. 1, pp. 25–41, 2023.

[8] E. Yulianti and N. K. Nissa, "ABSA of Indonesian customer reviews using IndoBERT: single-sentence and sentence-pair classification approaches," *Bull. Electr. Eng. Informatics*, vol. 13, no. 5, pp. 3579–3589, 2024.

[9] J. P. Gygi, S. H. Kleinstein, and L. Guan, "Predictive overfitting in immunological applications: Pitfalls and solutions," *Hum. Vaccines Immunother.*, vol. 19, no. 2, 2023.

[10] J. Chen, D. Tam, C. Raffel, M. Bansal, and D. Yang, "An Empirical Survey of Data Augmentation for Limited Data Learning in NLP," *Trans. Assoc. Comput. Linguist.*, vol. 11, pp. 191–211, 2023.

[11] S. Ranathunga, E. S. A. Lee, M. Prifti Skenduli, R. Shekhar, M. Alam, and R. Kaur, "Neural Machine Translation for Low-resource Languages: A Survey," *ACM Comput. Surv.*, vol. 55, no. 11, 2023.