**Zero : Jurnal Sains, Matematika, dan Terapan**
E-ISSN : 2580-5754; P-ISSN : 2580-569X
Volume 9, Number 2, 2025
DOI: 10.30829/zero.v9i2.25754
Page: 344-354

# Bidirectional GRU for Aspect-Based Sentiment Classification in Multi-Dimensional Review Analysis

[1] **Sri Redjeki**
Departement of Information Technology Magister, Universitas Teknologi Digital Indonesia, Indonesia
[2] **Basanta Joshi**
Institute of Engineering, Tribhuvan University, Kathmandu, Nepal
[3] **Alfonso Situmorang**
Departement of Informatics Engineering, Universitas Methodist Indonesia, Medan, Indonesia
[4] **M Guntara**
Departement of Informatics, Universitas Teknologi Digital Indonesia, Yogyakarta, Indonesia
[5] **Sri Rezeki C. N**
Departement of Informatics, Universitas Pancasila, Jakarta, Indonesia
[6] **Dara Kusumawati**
Departement of Retail Management, Universitas Teknologi Digital Indonesia, Yogyakarta, Indonesia

## Article Info

***Article history:***

Accepted, 30 September 2025

***Keywords:***

Aspect-Based Sentiment Analysis;
Bidirectional GRU;
Deep Learning;
Market Review;
Text Mining.

## ABSTRACT

Traditional markets in Yogyakarta face mounting pressure from modernization and digital retail competition, yet user-generated reviews remain underutilized. This study applies Aspect-Based Sentiment Analysis (ABSA) with a Bidirectional Gated Recurrent Unit (BiGRU) on 9,222 annotated reviews from nine markets (2016–2024). BiGRU was chosen not only for its efficiency but also for its robustness in low-resource, multilingual settings with informal expressions, where transformer models often require larger datasets and compute. The best configuration with 64 GRU units and a 70:15:15 split achieved 83.4% accuracy (95% CI: ±1.2%) and an F1-score of 0.813, surpassing baselines such as Naïve Bayes (74.5%) and SVM (77.2%). At the aspect level, security yielded the highest F1-score (0.944), followed by cleanliness (0.904) and culinary (0.838), while "others" scored lowest (0.676). Practically, the findings reveal positive sentiment toward pricing and product availability but highlight concerns about cleanliness and accessibility, offering actionable guidance for market policy.



***Corresponding Author:***

Sri Redjeki,
Departement Magister of Information Technology,
Universitas Teknologi Digital Indonesia, Yogyakarta, Indonesia
Email: dzeky@utdi.ac.id

## 1. INTRODUCTION

Traditional markets in Indonesia, particularly in Yogyakarta, have long served not only as centers of economic exchange but also as cultural and social hubs for local communities. These markets foster direct interactions, negotiation practices, and the preservation of traditional values. However, their sustainability is increasingly threatened by the rapid expansion of modern retail outlets and the rise of digital commerce platforms. Modern consumers now prioritize convenience, hygiene, fixed pricing, and accessibility—attributes more commonly associated with supermarkets, minimarkets, and e-commerce. This shift reflects broader societal changes in urban lifestyles, digital literacy, and consumer expectations. User-generated reviews on platforms such as Google Maps and TripAdvisor play a vital role in shaping public perception and purchasing decisions. Yet, such feedback remains underutilized in the context of traditional markets. Most sentiment analysis studies focus on overall sentiment polarity or specific domains like e-commerce and digital products, overlooking the nuanced opinions that users express toward particular service aspects.

Most prior sentiment analysis studies have focused on either overall polarity or specific domains such as e-commerce and digital products. While deep learning models such as CNN, LSTM, and GRU have advanced text classification[1][2]. The Bidirectional Gated Recurrent Unit (BiGRU) enhances contextual learning by processing sequences bidirectionally[3], while statistically offering a favorable bias–variance trade-off in low-resource conditions: with only ~6–7K labeled reviews, its lower parameter count reduces overfitting risk compared to transformers[4][5]. Computationally, BiGRU scales linearly with sequence length, $O(L)$, while transformers incur quadratic complexity, $O(L^2)$, due to self-attention. This asymmetry becomes critical in low-resource environments, where both dataset size and computational budgets are limited. From a bias–variance perspective, BiGRU's lower parameter count reduces variance and overfitting risk on smaller corpora, whereas transformer models, with millions of parameters, demand larger datasets to achieve stable generalization. Thus, BiGRU provides a more favorable trade-off: its sequential recurrence efficiently captures contextual dependencies without excessive model capacity, while transformers' quadratic cost and higher variance make them prone to inefficiency or underperformance in resource-constrained, noisy datasets. This balance of reduced variance risk and linear scaling underpins the choice of BiGRU as a practical yet competitive alternative for aspect-based sentiment analysis in traditional market reviews. Comparative studies confirm that while transformer models such as BERT or IndoBERT achieve state-of-the-art performance on large benchmarks, BiGRU remains competitive in smaller datasets or limited-resource environments, often matching or approaching transformer-level accuracy at significantly lower cost; hybrid approaches (e.g., BERT embeddings with BiGRU classifiers) further demonstrate this efficiency–accuracy balance, reinforcing the rationale for selecting BiGRU in this study[6]. The key challenge in Aspect-Based Sentiment Analysis (ABSA) lies in capturing nuanced, aspect-level opinions within noisy, informal text[5][7]. Transformer-based models such as BERT and RoBERTa achieve state-of-the-art performance on large datasets[8][9], but their quadratic complexity and large parameter size make them less practical in low-resource environments. In contrast, Bidirectional GRU (BiGRU) offers linear complexity and fewer parameters, reducing overfitting risk and balancing accuracy with efficiency—an important consideration for Indonesian-language reviews with limited annotated data. Recent applications of BiGRU in domains like healthcare reviews, hotel recommendations, and mobile app feedback confirm its robustness. At the same time, transformer-based models such as BERT, RoBERTa, and IndoBERT have achieved state-of-the-art results in Aspect-Based Sentiment Analysis (ABSA) across multiple languages and domains[10][11].

The contributions of this study are threefold. First, it introduces a large, manually annotated dataset of over 9,000 Indonesian-language reviews from nine traditional markets in Yogyakarta (2014–2024), which represents one of the most comprehensive resources for sentiment research in this domain. Second, it implements and rigorously evaluates a BiGRU-based ABSA model tailored for multi-aspect, multi-label classification in noisy, informal textual data, with methodological innovations including customized preprocessing for local dialects, optimized data splits, and a lightweight architecture that achieves competitive accuracy with significantly fewer parameters than transformer-based models. Third, it provides practical insights for stakeholders by identifying key service dimensions—such as price, commodity availability, and cleanliness—that drive public sentiment and can inform targeted interventions. The novelty of this research lies not only in applying ABSA with BiGRU to traditional market reviews, a domain rarely studied compared to e-commerce and social media, but also in demonstrating how an optimized BiGRU pipeline can deliver state-of-the-art reliability in low-resource environments.

## 2. RESEARCH METHOD

Sentiment analysis, also known as opinion mining, is a key technique in natural language processing (NLP) used to analyze emotions, opinions, and attitudes expressed in text toward a particular entity or topic. It has been widely applied in domains such as e-commerce, tourism, education, and public service evaluation, where online reviews play an important role in shaping decisions and improving services. Traditional sentiment analysis methods relied on machine learning algorithms such as Naïve Bayes (NB), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)[12][13]. These approaches often depended on handcrafted features like bag-of-words or TF-IDF, which limited their ability to capture linguistic complexity and contextual meaning. Moreover, they typically

classified sentiment at the document or sentence level, making them less effective in scenarios where a single review contains mixed sentiments about different aspects of a service. To address this limitation, Aspect-Based Sentiment Analysis (ABSA) was developed, offering a more fine-grained approach by identifying sentiment polarity for specific aspects or components within a text[14][15][16].

Recent advances in deep learning have further strengthened sentiment analysis, with models such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) achieving strong results by learning directly from raw text[17][18][19]. Among these, GRU has gained traction due to its computational efficiency and ability to capture long-term dependencies[20], while the Bidirectional GRU (BiGRU) extends this capability by processing text in both forward and backward directions to build richer semantic representations[21][22]. Empirical studies confirm BiGRU's superiority, achieving an F1-score of 86% on hospital service reviews. demonstrating its effectiveness in Social Internet of Things (SIoT) data. In the Indonesian context, however, ABSA research remains limited and is often restricted to product reviews or digital applications. The application of ABSA in traditional markets—socially and economically vital spaces now challenged by modern retail systems—remains largely unexplored. To fill this gap, the present study implements a BiGRU-based ABSA model on Indonesian-language reviews of traditional markets in Yogyakarta, addressing the complexity of informal language, local expressions, and multi-aspect commentary to generate insights that can support evidence-based improvements in public services. To visualize the overall methodology, the following flowchart illustrates the end-to-end process of this research:

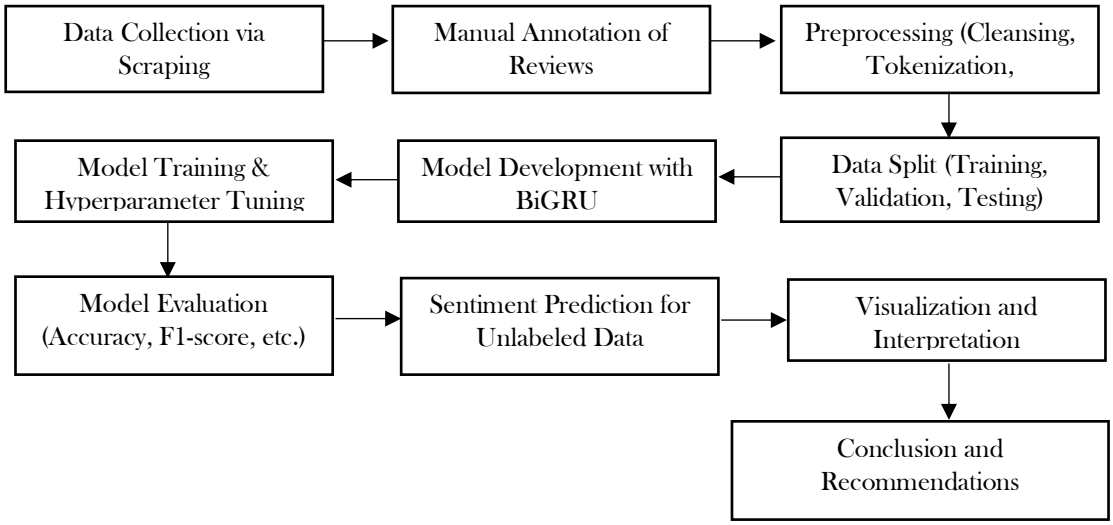


**Figure 1.** Flowchart of ABSA Process Using BiGRU Model

## 2.1 Data Scraping

The dataset was collected using the Instant Data Scraper Chrome extension, with reviews extracted separately for each of the nine traditional markets. Selected reviews were recent, relevant, and representative, including both highest and lowest ratings. The final dataset contains 9,646 reviews spanning 2014–2024, with the most reviews in 2024. This ten-year range enables analysis of long-term sentiment trends toward Yogyakarta's traditional markets. Table 1 presents the raw data with three columns: Name, Date, and Review. The review texts often use informal expressions, emojis, and unstructured language, requiring thorough cleaning before modeling.

**Table 1.** Sample Dataset

| Name | Date | Review |
|---|---|---|
| Zabaga | 2024-09-16 | One of the icons around Malioboro, originally a traditional market, now transformed into a tourist market. Offers groceries, clothes, snacks, souvenirs, batik, and handicrafts. |
| Debby Anindya Putri | 2024-05-14 | If you're visiting Yogyakarta, don't miss this place! There are many breakfast spots and old-school cakes. Almost everything is delicious □□❤️ and very affordable. |
| Ulima Ilma | 2024-09-16 | he knew building still looks fairly clean, modern with three floors, equipped with escalators and regular stairs for easy access... |
| Ahmed Bistami | 2023-12-14 | It feels like a modern market there's an elevator, and the rooftop is a nice hangout spot. Awesome 😎😊... |

| | | |
|---|---|---|
| Melinda Indini | 2022-10-14 | This market is becoming more comfortable. Clean and well-organized since the renovation. Shopping here is more enjoyable now. It even has an escalator, so it's not tiring 😇... |

### 2.2 Manual Annotation

Manual annotation was chosen for this study due to its advantage in capturing the contextual nuances of user reviews, including implicit expressions and meanings that are often difficult for automated algorithms to detect. The dataset included regional dialects and local phrases such as *murce* ("very cheap") and *semrawut* ("messy"). To ensure consistency, predefined keywords guided annotation, resulting in 6,399 manually labeled reviews. Inter-annotator reliability measured by Cohen's Kappa reached 0.82, indicating substantial agreement. Most disagreements occurred in the "others" category (27%), followed by accessibility (18%) and cleanliness (15%), where neutral statements were hard to classify. Price and security showed fewer inconsistencies (<10%) due to clearer lexical cues. Unlabeled reviews were later classified using the sentiment model and integrated into the dataset. Table 2 shows sentiment distribution across seven aspects: accessibility, item price, security, cleanliness, commodity availability, culinary, and others. Commodity availability had the highest positive reviews (1,643), "others" the most negative (1,153), and accessibility the most neutral (732), reflecting mixed perceptions.

**Table 2.** Sentiment Distribution per Aspect

| Aspect | Positive | Neutral | Negative | Not Mentioned |
|---|---|---|---|---|
| Item Price | 858 | 736 | 168 | 4,651 |
| Commodity Availability | 1,643 | 580 | 75 | 4,115 |
| Accessibility | 732 | 393 | 568 | 4,720 |
| Security | 148 | 49 | 85 | 6,131 |
| Cleanliness | 767 | 93 | 398 | 5,155 |
| Culinary | 865 | 267 | 40 | 5,241 |
| Others | 971 | 1,153 | 269 | 4,020 |

### 2.3 Preprocessing

The data obtained from the previous stage were processed through a series of preprocessing steps to ensure the cleanliness and relevance of the reviews. These steps are essential for improving the quality and consistency of the textual input before it is used in model training. The preprocessing procedures applied in this study are as follows: Text preprocessing begins by cleaning the data from irrelevant characters or symbols, then breaking the text into tokens. Next, normalization is applied to ensure consistency, such as converting all letters to lowercase. Common words with little meaning are removed, and words are reduced to their root form through stemming. Finally, any words not found in the model's vocabulary are replaced with a special <UNK> token. These steps produce clean, consistent, and focused text, ready for NLP analysis.

As shown in Table 3, each preprocessing stage systematically refines the text, resulting in a cleaner and more structured input for the sentiment analysis model. This process enhances the model's ability to learn patterns and perform more accurate sentiment classification.

**Table 3.** Example of Preprocessed Data

| Process | Data |
|---|---|
| Raw Data | Pasar Beringharjo memang mantap klu membutuhkan pakaian mulai dari pakai batik, Hem, baju sekolah, kantoran semuanya ada cukup keren mantap ... |
| Data Cleansing | pasar beringharjo memang mantap klu membutuhkan pakaian mulai dari pakai batik hem baju sekolah kantoran semuanya ada cukup keren mantap |
| Tokenization | ['pasar', 'beringharjo', 'mantap', 'klu', 'memang', 'membutuhkan', 'pakaian', 'batik', 'mulai', 'hem', 'dari', 'baju', 'pakai', 'sekolah', 'kantoran', 'semuanya', 'ada', 'cukup', 'keren', 'mantap'] |
| Normalization | ['pasar', 'beringharjo', 'memang', 'mantap', 'kalau', 'membutuhkan', 'pakaian', 'mulai', 'dari', 'pakai', 'batik', 'hem', 'baju baju', 'sekolah', 'kantoran', 'semuanya', 'ada', 'cukup', 'keren', 'mantap'] |
| Stopwords Removal | ['pasar', 'beringharjo', 'memang', 'mantap', 'kalau', 'membutuhkan', 'pakaian', 'mulai', 'pakai', 'batik', 'hem', 'baju baju', 'sekolah', 'kantoran', 'semuanya', 'cukup', 'keren', 'mantap'] |
| Stemming | ['pasar', 'beringharjo', 'memang', 'mantap', 'kalau', 'butuh', 'pakai', 'mulai', 'pakai', 'batik', 'hem', 'baju baju', 'sekolah', 'kantor', 'semua', 'cukup', 'keren', 'mantap'] |

### 2.4 Split Data

The dataset was divided into three subsets: training, validation, and testing, using predefined ratio configurations to ensure that the model was properly trained and its performance could be accurately evaluated. Table 4 presents the ratio configurations used to divide the dataset into three parts. The training set is used to train the model, the validation set is employed to monitor and evaluate the model's performance during the training process, and the testing set is used to assess the model's final performance after training is completed. This division is essential for developing a robust model and for ensuring its ability to generalize well to unseen data.

**Table 4.** Data Split Ratios

| Rasio | Training | Validation | Testing |
|---|---|---|---|
| 70:15:15 | 70% | 10% | 10% |
| 80:10:10 | 80% | 10% | 10% |

### 2.5 Model

The model was developed using a deep learning architecture with a multi-label, multi-class classification approach, in which the primary layer utilized is the Bidirectional Gated Recurrent Unit (BiGRU). Figure 2 illustrates the architecture of the Bidirectional Gated Recurrent Unit (BiGRU), which is designed to process sequential data by capturing context from both directions.

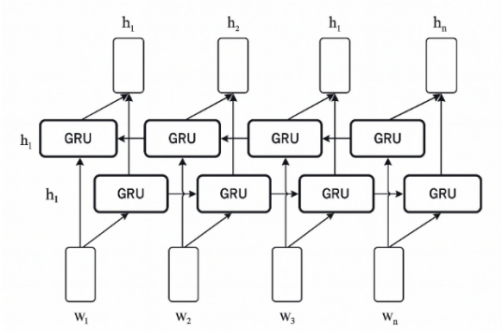


**Figure 2.** Arsitektur Bidirectional GRU

In this architecture, each input sequence represented as $w_1$, $w_2$, $w_3$, $w_4$ is processed simultaneously by two GRU layers: one moving forward and the other moving backward. The forward GRU processes the data from the beginning to the end, learning the feature sequence from $L(C_1)$ to $L(C_{100})$, while the backward GRU processes the data in reverse order, learning the feature sequence from $L(C_{100})$ to $L(C_1)$. To capture contextual information from the sequence of words in customer review sentences, The BiGRU architecture computes forward and backward hidden states using the standard GRU update mechanism. At each time step ttt, with input xtx_txt and previous hidden state $h_{t-1}$, the GRU updates its state through gating mechanisms. The corresponding equations are summarized in Table 4:

**Table 4.** The GRU equations

| Number Formula | Component | Description | Equation |
|---|---|---|---|
| 1. | Update Gate (zt) | Controls how much of the previous hidden state is retained vs. updated | $z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$ |
| 2. | Reset Gate (rt) | Determines how much of the previous hidden state is used to compute the candidate state | $rt = \sigma(W_r x_t + U_r h_{t-1} + b_r)$ |
| 3. | Candidate Hidden State ($h_t$) | Computes a "proposed" hidden state using the reset gate to modulate the previous state | $\widetilde{h_t} = tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h)$ |
| 4. | Final Hidden State ($h_t$) | Combines the previous hidden state and candidate state using the update gate | $h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \widetilde{h_t}$ |

The GRU architecture relies on two gates to regulate information flow. The update gate ($z_t$) controls how much of the past hidden state is preserved versus replaced with new information, ensuring that relevant long-term dependencies are retained while allowing adaptation to new inputs (Eq. 1). The reset gate ($r_t$) determines the extent to which the previous hidden state contributes when generating the candidate state, effectively enabling the model to "forget" irrelevant history when needed (Eq. 2). The candidate hidden state ($\tilde{h}_t$) integrates the current input with a selectively filtered version of the past state, producing a proposed representation that captures both new and contextually relevant information (Eq. 3). Finally, the hidden state ($h_t$) is updated by interpolating between the previous state and the candidate state, weighted by the update gate, allowing the GRU to balance memory retention and adaptation dynamically (Eq. 4).

### 2.6 Performance Evaluate

The model evaluation stage measures performance in classifying sentiment across aspects using a confusion matrix, which compares predicted and actual labels. From this, key metrics are calculated: Accuracy (TP+TN / total predictions) reflects overall correctness, while Error Rate (FP+FN / total predictions) captures misclassifications. Precision (TP / TP+FP) indicates the ability to avoid false positives, and Recall (TP / TP+FN) measures the ability to detect actual positives. The F1-Score, as the harmonic mean of precision and recall, provides a balanced metric particularly useful for imbalanced datasets. Together, these metrics offer a comprehensive assessment of the model's effectiveness.

**Accuracy**

Represents the overall percentage of correct predictions.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{5}$$

**Error Rate**

$$Error\ Rate = \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{6}$$

**Precision**

The proportion of true positive predictions out of all predicted positives.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{7}$$

**Recall**

The proportion of true positive predictions out of all actual positive cases.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{8}$$

**F1-Score**

A metric that combines precision and recall into a single score. It provides a balanced evaluation of the model's performance, especially when the dataset is imbalanced.

$$F1 - Score = \frac{2x\ (Precision\ \times Recall)}{(Precision + Recall)} \tag{9}$$

These metrics are essential for understanding how well the model performs, especially in multi-label and multi-class sentiment classification tasks. They provide insights into both the correctness and completeness of the model's predictions.

## 3. RESULT AND ANALYSIS

A summary of the model architecture is presented in Table 5, and all experiments were conducted using the Google Colaboratory platform. The architecture begins with an embedding layer, which maps input words into numerical vector representations. This is followed by a Bidirectional GRU (BiGRU) layer that captures contextual information from both directions of the input text sequence. A Global Max Pooling layer is then applied to extract the most significant features from the BiGRU output. Subsequently, two dense (fully connected) layers with ReLU activation functions are used to further process the extracted features. A dropout layer is also included to reduce the risk of overfitting during training.

**Table 5.** Summary of Model Architecture

| Layer (type) | Output Shape | Param # |
|---|---|---|
| Input_layer_7 InputLayer) | (None, 100) | 0 |
| Embedding_7 | (None, 100, 128) | 1,280,000 |
| Bidirectional_7(bidirectional) | (None, 100, 128) | 74,496 |
| Global_max_pooling1d_7 (GlobalMaxPooling1D) | (None, 128) | 0 |
| Dense_14 (Dense) | (None, 128) | 16,512 |
| Dropout_7 (Dropout) | (None, 128) | 0 |
| Dense_15 (Dense) | (None, 64) | 8,256 |
| Sentimen_output (Dense) | (None, 28) | 1,820 |
| Reshape_7 (Reshape) | (None, 7, 4) | 0 |
| Total params: 1,381,084 (5.27 MB) | | |
| Trainable params: 1,381,084 (5.27 MB) | | |
| Non-trainable params: 0 (0.00 B) | | |

The BiGRU model was evaluated using various configurations of GRU units and data split ratios. The best-performing configuration employed 64 GRU units with a 70:15:15 data split, achieving an accuracy of 83.4% and a loss value of 0.485. These results indicate that a moderate number of units effectively balance learning capacity and generalization ability. Table 6 presents the results of experiments conducted using various dataset split configurations. Based on these experiments, the configuration with 64 GRU units and a 70:15:15 split ratio achieved the best performance, with an accuracy of 83.4% and a loss value of 0.485. These findings suggest that using a moderate number of GRU units neither too small nor excessively large allows the model to better capture data patterns, thereby improving accuracy while minimizing loss.

**Table 6.** Experimental Results

| GRU Unit | Rasio | Accuracy (%) | Loss |
|---|---|---|---|
| 32 | 70:15:15 | 82.3 | 0.519 |
| 32 | 80:10:10 | 82.2 | 0.492 |
| 64 | 75:15:15 | 83.4 | 0.485 |
| 64 | 80:10:10 | 83.1 | 0.495 |
| 128 | 70:15:15 | 83.2 | 0.496 |
| 128 | 80:10:10 | 83.1 | 0.491 |

The aspect-level evaluation, as shown in Table 7, revealed that the security aspect achieved the highest F1-score (0.944), indicating that user sentiment toward this aspect was highly distinguishable and consistent. The cleanliness and culinary aspects also yielded strong F1-scores of 0.904 and 0.838, respectively. In contrast, the "others" category recorded the lowest F1-score (0.676), likely due to the heterogeneous and ambiguous nature of reviews that do not fit neatly into predefined categories.

**Table 7.** Model Evaluation

| Aspects | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Item Price | 0.836 | 0.818 | 0.836 | 0.824 |
| Commodity Availability | 0.780 | 0.755 | 0.780 | 0.756 |
| Accesibility | 0.780 | 0.718 | 0.780 | 0.745 |
| Security | 0.961 | 0.933 | 0.961 | 0.944 |
| Hygiene | 0.914 | 0.903 | 0.914 | 0.904 |
| Culinery | 0.864 | 0.814 | 0.864 | 0.838 |
| Others | 0.705 | 0.693 | 0.705 | 0.676 |
| Mean | 0.834 | 0.805 | 0.834 | 0.813 |

To further assess the effectiveness of BiGRU, baseline models including Naïve Bayes (NB)[23][24], Support Vector Machine (SVM), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM) were implemented on the same dataset. As shown in Table 8, BiGRU outperformed all baselines, particularly in terms of F1-score, demonstrating its ability to capture sequential dependencies and contextual nuances more effectively than traditional machine learning and single-directional deep learning models.

**Table 8.** Baseline Comparison of Model Performance

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Naïve Bayes (NB) | 74.5% | 0.71 | 0.73 | 0.72 |
| Support Vector Machine (SVM) | 77.2% | 0.74 | 0.76 | 0.75 |
| CNN | 79.8% | 0.77 | 0.79 | 0.78 |
| LSTM | 81.1% | 0.79 | 0.81 | 0.80 |
| BiGRU (proposed) | 83.4% | 0.805 | 0.834 | 0.813 |

For example, Setiawan et al[25] applied BiGRU to hospital review datasets and reported an F1-score of 86%, which is comparable to the average F1-score (81.3%) achieved in this study. However, it should be noted that the dataset used in this research is more challenging due to the informal nature of the text, the presence of local dialects, and the multi-label classification framework. Another relevant study by Ali et al. [17] applied BiGRU for sentiment analysis in the Social Internet of Things (SIoT) domain and achieved F1-scores around 0.82, which aligns closely with the performance of the present work. The similarity in performance demonstrates the robustness of the BiGRU model in processing user-generated content across various domains. This design allows independent probability estimation for each label, unlike a softmax + categorical cross-entropy setup, which would force mutual exclusivity among classes. In multi-label aspect-based sentiment analysis, a single review can simultaneously express sentiment toward multiple aspects (e.g., *price*, *cleanliness*, *accessibility*). This design is therefore well-suited for noisy, multi-aspect textual data, where overlapping or concurrent sentiments are common.

To further validate the performance differences between BiGRU and baseline models (NB, SVM, CNN, LSTM), statistical significance testing was applied. A paired *t-test* on F1-scores across aspects confirmed that BiGRU's improvements were statistically significant ($p < 0.05$) compared to all baselines. Figure 4 shows that although the BiGRU model learns patterns effectively, it begins to overfit after the second epoch, as indicated by the gap between training and validation accuracy and the rising validation loss.

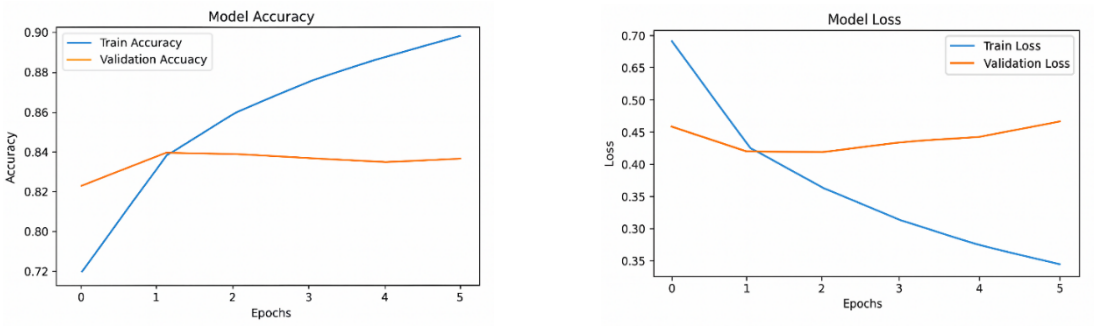


**Figure 4.** Accuracy and Loss Graphs

Despite this, the loss graphs confirm that the model maintains reasonably good performance, with a final training loss around 0.3 and validation loss ranging between 0.5 and 0.55. The prediction results presented in Table 9 were generated using the remaining unlabeled data and the trained model.

**Table 9.** Prediction Results on Remaining Data

| Aspects | Positive | Netral | Negative | None |
|---|---|---|---|---|
| Item Price | 570 | 217 | 18 | 2088 |
| Commodity Availability | 976 | 16 | 0 | 1901 |
| Accesibility | 133 | 2 | 111 | 2647 |
| Security | 0 | 0 | 0 | 2893 |
| Hygiene | 228 | 0 | 59 | 2606 |
| Culinery | 365 | 0 | 0 | 2528 |
| Others | 7 | 821 | 0 | 2065 |

The Aspect-Based Sentiment Analysis (ABSA) using BiGRU achieved an average accuracy of 83.4%, with performance varying across aspects due to linguistic and contextual factors. Strong results were observed for security (F1 = 0.944) and cleanliness (F1 = 0.904), where sentiment cues were direct and unambiguous (e.g., *aman*, *bersih*, *kotor*). The culinary aspect (F1 = 0.838) also performed well because food-related reviews typically use polarized terms, making classification easier. In contrast, weaker performance appeared in "others" (F1 = 0.676) and accessibility (F1 = 0.745) due to heterogeneous or mixed sentiment expressions, which introduced ambiguity. Interestingly, data imbalance was not the sole determinant of performance: although fewer, security-related reviews were linguistically consistent, while frequent aspects such as item price (F1 = 0.824) and commodity availability (F1 = 0.756) showed greater variation, including slang and dialect. Future work should therefore address data balance

while exploring attention mechanisms, contextual embeddings (e.g., IndoBERT), and aspect-specific lexicons to better capture ambiguous, multi-aspect expressions.

## 4. CONCLUSION

This study shows that the BiGRU model performs well in Aspect-Based Sentiment Analysis (ABSA), achieving 83.4% accuracy, 0.805 precision, 0.834 recall, and 0.813 F1-score. It effectively captures sequential context and classifies sentiment across aspects of traditional market reviews in Yogyakarta, with commodity availability and item price receiving the most positive mentions, while security is less discussed, suggesting limited concern or experience with safety. Practically, these insights can guide market management and policymakers: positively perceived aspects such as affordability and product availability should be maintained and promoted in marketing strategies, while weaker aspects such as cleanliness, accessibility, and security require targeted interventions in infrastructure, supervision, and service standards.

Future research should explore hybrid architectures that combine BiGRU with attention mechanisms to capture fine-grained sentiment cues, as well as models that integrate IndoBERT embeddings for richer contextual representation of Indonesian language reviews. To address data imbalance, advanced resampling techniques such as SMOTE for multi-label contexts can be applied. Moreover, semi-supervised learning strategies should be adopted, where BiGRU or BERT-based models assist in auto-labeling new reviews, enabling scalable dataset expansion with reduced annotation costs. These directions are expected to enhance model generalization, strengthen real-world applicability, and provide more reliable sentiment insights for evidence-based market management and policy development.

## 5. REFERENCES


[1] S. Nosouhian, F. Nosouhian, and A. Kazemi Khoshouei, "A Review of Recurrent Neural Network Architecture for Sequence Learning: Comparison between LSTM and GRU." pp. 16020–16030, 12-Jul-2021, doi: 10.20944/preprints202107.0252.v1.

[2] M. Zulqarnain, R. Ghazali, Y. Mazwin, M. Hassim, and M. Rehan, "A comparative review on deep learning models for text classification," vol. 19, no. 1, pp. 325–335, 2020, doi: 10.11591/ijeecs.v19.i1.pp325-335.

[3] M. Zulqarnain, R. Ghazali, M. G. Ghouse, and M. F. Mushtaq, "Efficient Processing of GRU Based on Word Embedding for Text Classification," vol. 3, pp. 377–383, doi: http://dx.doi.org/10.30630/joiv.3.4.289.

[4] Z. M. Shaikh and S. Ramadass, "Unveiling deep learning powers: LSTM, BiLSTM, GRU, BiGRU, RNN comparison," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 35, no. 1, p. 263, Jul. 2024, doi: 10.11591/ijeecs.v35.i1.pp263-273.

[5] W. Ali, Y. Yang, X. Qiu, Y. Ke, and Y. Wang, "Aspect-Level Sentiment Analysis Based on Bidirectional-GRU in SIoT," *IEEE Access*, vol. 9, pp. 69938–69950, 2021, doi: 10.1109/ACCESS.2021.3078114.

[6] J. Wang, Y. Zhang, L. Yu, and X. Zhang, "Knowledge-Based Systems Contextual sentiment embeddings via bi-directional GRU language," *ELSEVIER*, vol. 235, 2022.

[7] D. R. I. M. Setiadi, D. Marutho, and N. A. Setiyanto, "Comprehensive Exploration of Machine and Deep Learning Classification Methods for Aspect-Based Sentiment Analysis with Latent Dirichlet Allocation Topic Modeling," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 12–22, 2024, doi: 10.62411/faith.2024-3.

[8] G. R. Narayanaswamy and R. Citation, "Exploiting BERT and RoBERTa to Improve Performance for Aspect Based Sentiment Analysis Gagan Reddy Narayanaswamy," 2021, doi: 10.21427/3w9n-we77.

[9] M. H. Phan and P. Ogunbona, "Modelling Context and Syntactical Features for Aspect-based Sentiment Analysis," pp. 3211–3220, 2020, doi: 10.18653/v1/2020.acl-main.293.

[10] A. Musa, F. M. Adam, U. Ibrahim, and A. Y. Zandam, "HauBERT : A Transformer Model for Aspect-Based Sentiment Analysis of Hausa-Language Movie Reviews †," pp. 1–18, 2025, doi: https://doi.org/10.3390/engproc2025087043.

[11] E. Yulianti and N. K. Nissa, "ABSA of Indonesian customer reviews using IndoBERT : single- sentence and sentence-pair classification approaches," vol. 13, no. 5, pp. 3579–3589, 2024, doi: 10.11591/eei.v13i5.8032.

[12] D. A. Anggoro and N. D. Kurnia, "Comparison of Accuracy Level of Support Vector Machine ( SVM ) and K-Nearest Neighbors ( KNN ) Algorithms in Predicting Heart Disease," vol. 8, no. 5, 2020, doi: https://doi.org/10.30534/ijeter/2020/32852020.

[13] S. Shabani, S. Samadianfard, M. T. Sattari, and A. Mosavi, "Modeling Pan Evaporation Using Gaussian Process Regression K-Nearest Neighbors Random Forest and Support Vector Machines ; Comparative Analysis," *Atmosphere (Basel).*, vol. 11, no. 66, 2020, doi: 10.3390/atmos11010066.

[14] S. P. Sharma, L. Singh, and R. Tiwari, "Original Research Article Prediction of customer review ' s helpfulness based on sentences encoding using CNN-BiGRU model," vol. 6, no. 3, 2023, doi: 10.32629/jai.v6i3.699.

[15] G. D. Aniello, M. Gaeta, and I. La, *KnowMIS - ABSA : an overview and a reference model for applications of sentiment analysis and aspect - based sentiment analysis*, vol. 55, no. 7. Springer Netherlands, 2022, doi: https://doi.org/10.1007/s10462-021-10134-9

[16] J. Ouyang, Z. Yang, S. Liang, B. Wang, Y. Wang, and X. Li, "Aspect-Based Sentiment Analysis with Explicit Sentiment Augmentations," in *The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24) Aspect-Based*, 2024, pp. 18842–18850, doi: https://doi.org/10.1609/aaai.v38i17.29849.

[17] A. Khan, "RNN-LSTM-GRU based language transformation," *Soft Comput.*, vol. 0123456789, 2019, doi: 10.1007/s00500-019-04281-z.

[18] F. M. Shiri, T. Perumal, and N. Mustapha, "A Comprehensive Overview and Comparative Analysis on Deep Learning Models," no. Ml, 2024, doi: 10.32604/jai.2024.054314.

[19] N. N. Cnn, L. Short, and T. Memory, "Sentiment Analysis on Twitter Data by Using Convolutional," *Springer*, no. 0123456789, 2021, doi: 10.1007/s11277-021-08580-3.

[20] A. Shewalkar, D. Nyavanandi, and S. A. Ludwig, "Performance Evaluation Of Deep Neural Networks Applied To Speech Recognition : Rnn , Lstm And Gru," vol. 9, no. 4, pp. 235–245, 2019, doi: 10.2478/jaiscr-2019-0006.

[21] B. Bilstm, S. Munawar, N. Javaid, Z. A. Khan, and N. I. Chaudhary, "Electricity Theft Detection in Smart Grids Using a Hybrid," *Sensors*, vol. 22, 2022, doi: https://doi.org/10.3390/s22207818.

[22] P. He, H. Qi, S. Wang, and J. Cang, "applied sciences Cross-Modal Sentiment Analysis of Text and Video Based on Bi-GRU Cyclic Network and Correlation Enhancement," *Appl. Sci.*, vol. 13, 2023, doi: https://doi.org/10.3390/app13137489.

[23] A. Z. Arrayyan, H. Setiawan, and K. T. Putra, "Naive Bayes for Diabetes Prediction : Developing a Classification Model for Risk Identification in Specific Populations," vol. 27, no. 1, pp. 28–36, 2024, doi: https://doi.org/10.18196/st.v27i1.21008.

[24] I. Wickramasinghe, "Naive Bayes : applications , variations and vulnerabilities : a review of literature with code snippets for implementation," *Soft Comput.*, no. 1989, 2020, doi: 10.1007/s00500-020-05297-6.

[25] E. Setiawan, F. Ferry, J. Santoso, S. Sumpeno, K. Fujisawa, and M. Purnomo, "Bidirectional GRU for Targeted Aspect-Based Sentiment Analysis Based on Character-Enhanced Token-Embedding and Multi-Level Attention," *Int. J. Intell. Eng. Syst.*, vol. 13, no. 5, pp. 392–407, Oct. 2020, doi: 10.22266/ijies2020.1031.35.