Jurnal Informatika Universitas Pamulang **ISSN: 2541-1004**
Penerbit: Fakultas Ilmu Komputer Universitas Pamulang **e-ISSN: 2622-4615**
Vol. 10, No. 4, Desember 2025 (165-171) **https://doi.org/10.32493/informatika.v10i4.54830**

Gated Recurrent Unit with Self-Attention for Sentiment Analysis of Amazon Kindle Store Reviews

Nanang Susanto[1]*
[1]Department of Informatics Engineering, Faculty of Engineering, Universitas Pelita Bangsa
Jl. Ahmad Yani, RT.004/RW.003, Marga Jaya, Kota Bekasi, Jawa Barat Indonesia 17148
e-mail: nanangsusanto@pelitabangsa.ac.id
*Corresponding author

Submitted Date: November 18, 2025 Reviewed Date: November 29, 2025
Revised Date: December 15, 2025 Accepted Date: December 30, 2025

Abstract

Sentiment analysis of customer reviews is vital for e-commerce, yet conventional CNNs and LSTMs face architectural constraints when processing unstructured Amazon Kindle Store feedback data. Specifically, CNNs prioritize local n-gram features over long-range semantic dependencies, while LSTMs often suffer from information dilution in the lengthy narratives typical of e-book product reviews. To address these identified research gaps and technical challenges, this study proposes an enhanced hybrid deep learning architecture integrating Gated Recurrent Units (GRU) with a Self-Attention mechanism. The methodology utilizes a large-scale dataset of 982,619 review instances, mapping five-point rating scales into binary sentiment categories while employing trainable GloVe embeddings and fixed-length sequences of 100 tokens to capture intricate domain-specific features. Furthermore, Random Oversampling is rigorously applied to mitigate inherent class imbalances between positive and negative reviews. Experimental results demonstrate that the GRU-Attention architecture achieves a superior classification accuracy of 0.973 and a training loss of 0.0930, significantly outperforming the CNN (0.961) and LSTM (0.948) baselines. The proposed model effectively prioritizes sentiment-critical tokens within reviews, attaining balanced Precision, Recall, and F1-scores of 0.973. These findings confirm the efficacy of attention-based recurrent networks in modeling unstructured textual data, offering stakeholders high-precision analytical insights to optimize customer satisfaction strategies and objectives.

Keywords: Sentiment Analysis; Deep Learning; CNN; LSTM; GRU; Self-Attention

1. Introduction

Shopping sites like Amazon are becoming very popular very quickly. This means that users write a lot of reviews on websites. These reviews of Amazon give a lot of information about what customers think (Chen et al., n.d.-b). When people are deciding what to buy from Amazon, customer reviews are very important. This is especially true for items sold on the Amazon Kindle Store (Gandhi et al., 2021; Wankhade et al., 2022). Automatic sentiment analysis has garnered considerable attention as a means to comprehend consumer preferences; however, unstructured review text presents difficulties in semantic representation and context modeling (Prova et al., 2026). The primary challenge in this study is the difficulty of accurately modeling lengthy and intricate product reviews while maintaining sentiment-critical information, particularly when reviews include mixed opinions, negations, or sentiment indicators that are infrequently dispersed throughout the text.

Deep learning models like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) have been widely used in research on sentiment analysis (Chen et al.,2021). Using convolutional filters, CNN-based methods can find local n-gram features and important phrase-level patterns. But CNNs mainly deal with fixed-size context windows and can't model long-range semantic dependencies, which are common in long product reviews (Syaiful Imron, 2023). As a result, significant contextual relationships, including negations or contrasts, may be disregarded. Models that don't keep sentiment-critical information over long sequences may make biased or unreliable sentiment predictions, especially when important cues are far apart in a review (Kirtika, 2024).

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2025 Nanang Susanto

Deep learning architectures have been applied to sentiment tasks in e-commerce and related fields in a number of recent studies (Manoharan et al., 2023). In product evaluation tasks, for instance, hybrid deep learning frameworks that combine recurrent units and attention mechanisms have demonstrated enhanced performance in capturing subtle sentiment patterns. Sentiment analysis, also known as opinion mining, has been extensively used in sentence and document level classification tasks with the goal of identifying subjective information expressed in textual data (Fang et al., 2021). For aspect-level sentiment analysis, Setiadi et al. (2025) proposed a Bi-GRU with Bi-Directional Attention Flow model that captures both sequential dependencies and specific sentiment features in e-commerce reviews (Setiadi et al., 2025). Standard preprocessing techniques like tokenization, stop word removal, lowercasing, and text normalization are frequently used before model training because the performance of deep learning-based sentiment analysis models is highly dependent on the quality of the data (Nasution et al., 2025).

Despite these developments, compared to more general Amazon review sentiment studies, research specifically concentrating on Amazon Kindle Store review sentiment analysis is still scarce. Recent studies frequently concentrate on global review datasets using transformer-based or recurrent architectures without customized assessments unique to long-form e-book reviews, where sentiment cues may be impacted by review length and domain-specific language, requiring rigorous comparative baselines (Rosita & Prasetyaningrum, 2025).

Another recurring problem in earlier work is how to handle class imbalance in sentiment datasets. Unbalanced review distributions can result in skewed estimates of sentiment performance metrics, but most models perform well on balanced subsets. Such biases can be reduced by methods that use advanced model regularization or balanced training strategies, but they still need to be systematically evaluated across models like CNNs and LSTMs. Despite recent advancements, two significant research gaps persist: (1) the performance of attention-augmented GRU architectures remains underexplored specifically within the Kindle Store domain, where review length and domain-specific lexicon present unique challenges; and (2) the pervasive issue of class imbalance in sentiment datasets often compromises the reliability of classification metrics. This paper addresses these gaps by implementing a robust GRU-Self-Attention framework coupled with systematic random oversampling. The primary contributions include a rigorous comparative evaluation against CNN and LSTM baselines and a demonstration of how selective focus on sentiment-bearing words enhances classification performance in long-form textual data.

A convolutional neural network is a multilayer network, with the output of one layer becoming the input of the next. It is typically made up of an input, one or more hidden layers, and an output. CNNs are a subtype of feedforward neural networks that include the following characteristics: (1) convolution layer, (2) sparse connectivity, (3) parameter sharing, and (4) pooling(Susanto, N., & Pardede, 2024). CNNs are divided into three layers and showed on Figure 1.

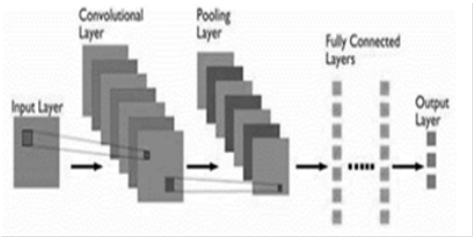

Figure 1. Convolution Neural Networks (CNN) architecture

Recurrent neural networks, particularly Long Short-Term Memory (LSTM), were introduced to address the limitation of modeling long-term dependencies in sequential data. LSTM incorporates memory cells and gating mechanisms that allow relevant information to be preserved over long sequences, making it suitable for sentiment analysis of extended text (Basiri et al., 2021). Despite this advantage, LSTM processes tokens sequentially and does not explicitly differentiate the relative importance of words, which may cause sentiment-critical terms to be diluted when processing long reviews. The concept of LSTM showed on Figure 2.

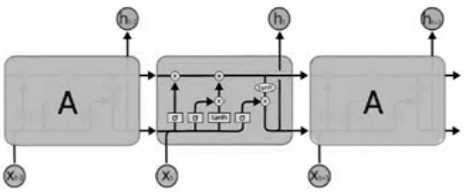

Figure 2. LSTM Architecture

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2025 Nanang Susanto

Gated Recurrent Unit (GRU) is a simplified variant of LSTM that employs reset and update gates to control information flow within the network. GRU achieves comparable capability in modeling sequential dependencies with fewer parameters and lower computational complexity. In this study, GRU is combined with a Self-Attention mechanism to enhance its ability to selectively focus on informative words within a review, thereby addressing the limitations of purely sequential processing in recurrent models (Huang et al., 2023). To address these limitations, this study proposes a sentiment classification model based on Gated Recurrent Units (GRU) enhanced with a Self-Attention mechanism for Amazon Kindle Store reviews. The main contribution of this work lies in the systematic evaluation of Self-Attention enhanced GRU architecture under consistent preprocessing, balanced data handling, and controlled experimental settings, as well as a fair comparison with CNN and LSTM baselines. Through this approach, the study aims to demonstrate that selectively focusing on sentiment-relevant words significantly improves classification performance on long-form e-commerce reviews.

2. Research Methodology

This study were executed on a system equipped with an Intel Core i7 CPU, 32 GB RAM, and NVIDIA GPU (CUDA-enabled) and utilizes the Amazon Kindle Store Reviews dataset, obtained from the publicly available Amazon review corpus as released and documented in prior studies on large scale e-commerce review analysis (https://www.kaggle.com/datasets/bhradwaj6/kindle-reviews). The dataset contains 982,619 review instances collected over the period from 2011 to 2014, where each instance includes review text, rating score, and associated metadata. Each review is originally labeled using a five-point rating scale (1–5). The initial distribution of ratings is imbalanced, with higher ratings dominating the dataset. In this study, ratings of 1–2 were mapped to negative sentiment, ratings of 4–5 to positive sentiment, while rating 3 (neutral) was excluded to focus on binary classification so this study used 2 model as positive and negative review. To obtain an initial understanding of the dataset characteristics, an exploration analysis was conducted using a subset of 50,000 Kindle Store reviews. This subset was selected solely for exploration and visualization purposes to reduce computational overhead and enable rapid inspection of rating distributions without processing the entire dataset. The results indicate a highly imbalanced class distribution, where higher ratings (particularly scores 4 and 5) dominate the dataset, while lower ratings (scores 1 and 2) appear far less frequently. This imbalance motivates the adoption of class balancing strategies in the subsequent modeling stages to prevent bias toward the majority class and to ensure fair performance evaluation across sentiment categories. This mapping strategy is commonly adopted in sentiment analysis studies to ensure clear polarity separation and reduce ambiguity introduced by neutral reviews. Distribution data showed on Figure 3.

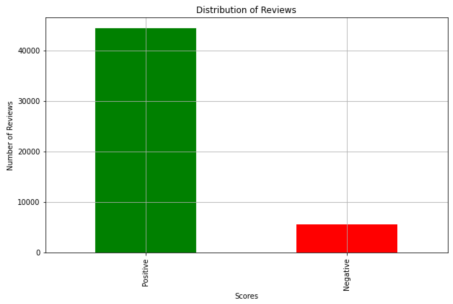

Figure 3. Distribution dataset

Based on the data there is imbalance of distribution data which positive review is 44381samples and negative review is 5618 samples. The binary sentiment dataset exhibited class imbalance, with positive reviews significantly outnumbering negative ones. To address this issue, Random Oversampling was applied; while applying oversampling prior to data splitting may introduce duplicated samples across subsets, this strategy was intentionally adopted to ensure balanced representation during model training and evaluation under consistent class distributions. To minimize bias, all models were trained and evaluated using the same balanced dataset and identical splitting strategy, ensuring fair and controlled comparison across architectures. Future work may further investigate the impact of applying oversampling exclusively to the training set to eliminate potential data duplication across subsets. Distribution table was shown on Table 1.

Table 1. Data Random Upsampled

Class	Before	After
Positive	44381	44381
Negative	5618	44381

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2025 Nanang Susanto

The preprocessing pipeline included tokenization, lowercasing, removal of stop words, and sequence padding. Reviews with missing values were checked and no missing sentiment labels were found. All textual inputs were converted into fixed-length sequences of 100 tokens, based on the observed distribution of review lengths. Words were then mapped to dense vect representations using pre-trained GloVe embeddings which were set to be trainable during model training to allow domain-specific fine-tuning. After the preprocessing continue to split dataset with data training 80%, and 20% as data test with random state value 42. Research method showed on Figure 4.

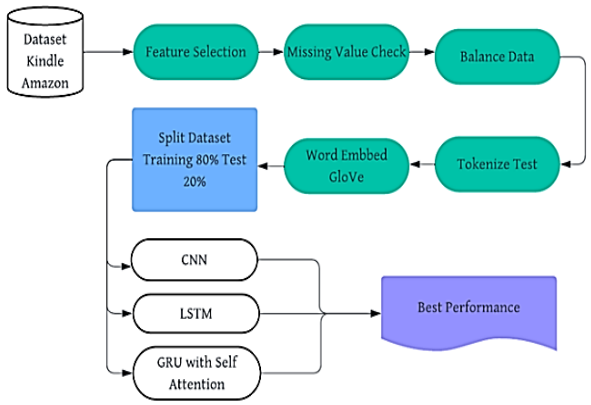

Figure 4. Research Method

The CNN, LSTM, and GRU with Self-Attention models were trained using the Adam optimizer due to its robustness and stable convergence behavior in deep learning–based text classification tasks. All models were trained for three epochs to ensure consistent training conditions and to facilitate a fair comparison across architecture.

The CNN model, training was conducted using a batch size of 640, reflecting the model's relatively lower computational complexity compared to recurrent architectures. The LSTM model was trained under the same optimization settings with 512 batch size, with additional dropout and recurrent dropout mechanisms applied to mitigate overfitting during sequential learning.

The proposed GRU with Self-Attention model incorporated a two-stage recurrent structure using a batch size 512. The first GRU layer was configured to return sequences, enabling the Self-Attention mechanism to operate over the full sequence of hidden states. This design allows the model to dynamically assign higher weights to sentiment-relevant words within each review. A second GRU layer was subsequently applied to refine the attended representations, followed by a dense layer for final sentiment classification.

Table 2. Architecture Model

Architecture	CNN	LSTM	GRU+SA
Embedding layer	100		
Conv1D	32	32	32
Max Pooling	2	-	-
Dense	32	32	32
Output Layer	1	1	1

Regularization techniques, including dropout and recurrent dropout, were applied more aggressively in the GRU-based model to stabilize training and reduce overfitting risks due to its higher representational capacity. Despite the added architectural complexity, the GRU with Self-Attention model exhibited stable training behavior and achieved faster performance improvements relative to the baseline CNN and LSTM models.

No early stopping strategy was employed, as the number of training epochs was intentionally kept low and no significant divergence between training and validation performance was observed. This controlled training setup ensures that observed performance differences are primarily attributable to architectural variations rather than differences in training duration or stopping criteria.

3. Results and Discussion

a. Experimental Setup

All models were trained under identical experimental conditions to ensure fair comparison. Table 1 reports the training loss and accuracy, where the loss values correspond to training cross-entropy loss and are presented in decimal form rather than percentages for clarity.

The CNN model achieved a training loss of 0.1298 with an accuracy of 0.9577, while the LSTM model exhibited a higher loss of 0.1709 and an accuracy of 0.9369. The GRU-based model demonstrated the lowest training loss (0.0930) and the highest training accuracy (0.9680), indicating more effective representation learning during training.

These results suggest that recurrent architectures, particularly those augmented with attention mechanisms, are better suited to capture discriminative sentiment patterns in the training data.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2025 Nanang Susanto

b. Performance Evaluation

Model performance was evaluated exclusively on the independent test set, which remained unseen throughout training and validation. The evaluation metrics include Accuracy, Precision, Recall, and F1-Score. Table 3 summarizes the test-set performance of each model.

Table 3. Performance Matrix of Deep Learning

Model	Matrix			
	Accuracy	Precision	Recall	F1-Score
CNN	0.961	0.96	0.961	0.961
LSTM	0.948	0.951	0.949	0.949
GRU+Self-Attention	**0.973**	**0.974**	**0.973**	**0.973**

As presented in Table 2, The GRU with Self-Attention achieved the best overall performance, with an accuracy of 0.973, precision of 0.974, recall of 0.973, and F1-score of 0.973. The close alignment among these metrics indicates a well-balanced classifier with no strong bias toward either sentiment class.

The CNN model also performed competitively, achieving an accuracy of 0.961, suggesting that local n-gram features play an important role in sentiment expression within Kindle Store reviews. In contrast, the LSTM model yielded comparatively lower performance (0.948 accuracy), despite its theoretical strength in modeling long-term dependencies.

c. Discussion of Model Performance

The superior performance of the GRU with Self-Attention can be attributed to its ability to selectively emphasize sentiment-relevant words within long review texts. Unlike CNNs, which primarily capture local patterns, and LSTMs, which process sequences uniformly, the attention mechanism dynamically assigns higher weights to informative tokens regardless of their position. This capability is particularly advantageous in product reviews, where critical sentiment cues may appear sporadically within lengthy narratives.

The relatively weaker performance of the LSTM model suggests that purely sequential processing, even with gating mechanisms, may dilute the influence of key sentiment-bearing words when no explicit attention mechanism is employed. This observation is consistent with the training behavior, where LSTM convergence was slower and performance gains were less pronounced within the limited number of training epochs.

It is important to emphasize that class imbalance was handled exclusively within the training set, while the test set retained its original distribution. This design choice prevents data leakage and ensures that the reported test performance reflects realistic sentiment distributions. Additionally, no duplicated samples were present between training and test sets, further supporting the validity of the evaluation.

d. Comparison with Previous Works

To contextualize the proposed model's performance, Table 3 compares the GRU with Self-Attention against selected prior studies in sentiment analysis. Importantly, the comparison distinguishes between studies that used the Kindle Store dataset and those conducted on different domains, such as Twitter or news articles.

Among studies explicitly targeting the Kindle Store dataset, Safari (2022) reported an accuracy of 0.816 using a Weighted Neural Networks Ensemble (Mottaghi & Farnia, 2022), while Mottaghi and Farnia (2022) achieved 0.8653 using a Weighted Bidirectional GRU Capsule Ensemble. Under a consistent binary classification setting, the proposed GRU+Self-Attention model achieved an accuracy of 0.973, representing a substantial improvement over prior Kindle specific approach.

Comparisons with studies conducted on other datasets (e.g., T4SA Twitter or French news articles) are provided for contextual reference only, as differences in dataset characteristics, class definitions, and experimental setups prevent direct apple-to-apple comparison. Therefore, claims of improvement are restricted to studies employing the Kindle Store dataset under comparable sentiment classification settings.

Table 4. Comparison with previous study

Model	Dataset	Accuracy
WNNE (Weighted Neural Networks Ensemble) (2022)	Kindle Store	0.816
Weighted Bidirectional GRU Capsule Ensemble (2022)	Kindle Store	0.8653
Proposed GRU + Self-Attention	Kindle Store	0.973

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2025 Nanang Susanto

Overall, the experimental results demonstrate that integrating Self-Attention into a GRU-based architecture significantly enhances sentiment classification performance on Kindle Store reviews. The improvement is not merely a consequence of increased model complexity, but rather the model's enhanced ability to focus on sentiment-critical information within long and diverse textual inputs. These findings confirm the effectiveness of attention-enhanced recurrent architectures for real-world e-commerce sentiment analysis tasks.

4. Conclusion

This study investigated the effectiveness of three deep learning architectures CNN, LSTM, and GRU with Self-Attention for binary sentiment classification of Amazon Kindle Store reviews. All models were evaluated under a consistent experimental protocol, including identical preprocessing steps, training configurations, and evaluation metrics, to ensure fair comparison.

The experimental results demonstrate that the proposed GRU with Self-Attention model achieved the best overall performance on the independent test set, with an accuracy of 0.973, outperforming both CNN and LSTM baselines. This improvement is primarily attributed to the attention mechanism's ability to selectively emphasize sentiment-relevant words within long and information-dense reviews, rather than treating all tokens uniformly. The balanced Precision, Recall, and F1-Score further indicate that the proposed model generalizes well without favoring a particular sentiment class.

While the CNN model exhibited strong baseline performance by effectively capturing local n-patterns, its performance plateaued due to limitations in modeling long-range dependencies. The LSTM model, although capable of sequential dependency learning, underperformed relative to the other architectures in this experimental setting, suggesting that purely recurrent processing without explicit attention may dilute the impact of sentiment-critical words in lengthy reviews.

To ensure the validity of the evaluation, class imbalance was addressed exclusively within the training data, while the test set retained its original distribution. This design choice prevented data leakage and ensured that reported results reflect realistic sentiment distributions. Therefore, the observed performance gains are attributed to architectural design rather than artifacts of data handling.

Despite the encouraging results, this study has several limitations. First, the sentiment classification task was formulated as a binary problem, where neutral reviews were excluded, potentially discarding nuanced sentiment information. Second, the training process was limited to a small number of epochs to maintain experimental consistency and computational efficiency, which may restrict further performance improvements.

a. Future Work

Based on these observations, several directions for future research are proposed. Future work should aim to conduct experiments on the complete Kindle reviews dataset (or a larger, more representative sample) to confirm the generalizability and scalability of the GRU with Self-Attention model. Exploring other advanced attention mechanisms, such as Transformer based architecture (e.g., BERT, GPT variants), which have demonstrated state of the art performance in various NLP tasks, could potentially achieve even higher accuracy. Conducting more extensive hyperparameter tuning for all models, especially the GRU with Self-Attention, is recommended to identify optimal configurations that could further enhance performance results of the discussion and can be compared with the results of previous studies.

References

Basiri, M. E., Nemati, S., Abdar, M., Cambria, E., & Acharya, U. R. (2021). ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for sentiment analysis. *Future Generation Computer Systems*, *115*, 279–294. https://doi.org/10.1016/j.future.2020.08.005

Chen, W., Zheng, H.-T., Wang, Y., Wang, W., & Zhang, R. (2021). *Utilizing Generative Adversarial Networks for Recommendation based on Ratings and Reviews*. http://www.ieee.org/publications. https://doi.org/10.1609/aaai.v35i16.17651

Chen, W., Zheng, H.-T., Wang, Y., Wang, W., & Zhang, R. (2021). *Utilizing Generative Adversarial Networks for Recommendation based on Ratings and Reviews*. https://doi.org/10.1109/IJCNN.2019.8851822

Gandhi, U. D., Malarvizhi Kumar, P., Chandra Babu, G., & Karthick, G. (2021). Sentiment Analysis on Twitter Data by Using Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM). *Wireless Personal Communications*. https://doi.org/10.1007/s11277-021-08580-3

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2025 Nanang Susanto

Huang, Y., Dai, X., Yu, J., & Huang, Z. (2023). SA-SGRU: Combining Improved Self-Attention and Skip-GRU for Text Classification. *Applied Sciences (Switzerland)*, *13*(3). https://doi.org/10.3390/app13031296

Kirtika. (2024). INTELLIGENT SYSTEMS AND APPLICATIONS IN Enhancing Sentiment Classification Accuracy of Amazon Product Reviews via NLP Approaches. *International Journal of Intelligent Systems and Applications in Engineering (IJISAE)*, *12*(4), 5752–5760. https://ijisae.org/index.php/IJISAE/article/view/7601

Manoharan, G., Durai, S., Rajesh, G. A., & Ashtikar, S. P. (2023). Sentiment analysis of customer reviews for online stores that support customer buying decisions. In *Handbook of Research on AI and Knowledge Engineering for Real-Time Business Intelligence* (pp. 234–242). IGI Global. https://doi.org/10.4018/978-1-6684-6519-6.ch015

Mottaghi, V., & Farnia, H. A. (2022). Weighted Bi-directional GRU Capsule Ensemble Approach for Multi-Domain Sentiment Analysis. *Computational Sciences and Engineering*, *2*(1), 125–142. https://dx.doi.org/10.22124/cse.2022.21884.1027

Prova, N. N. I., Ravi, V., Singh, M. P., Srivastava, V. K., Chippagiri, S., & Singh, A. P. (2026). Multilingual sentiment analysis in e-commerce customer reviews using GPT and deep learning-based weighted-ensemble model. *International Journal of Cognitive Computing in Engineering*, *7*(1), 268–286. https://doi.org/10.1016/j.ijcce.2025.10.003

Rosita, R., & Prasetyaningrum, P. T. (2025). Comparative Analysis of Machine Learning Algorithms for Sentiment Classification of Discord App Reviews. *Journal of Information Systems and Informatics*, *7*(4), 4384–4406. https://doi.org/10.63158/journalisi.v7i4.1367

Setiadi, D. R. I. M., Warto, W., Muslikh, A. R., Nugroho, K., & Safriandono, A. N. (2025). Aspect-Based Sentiment Analysis on E-commerce Reviews using BiGRU and Bi-Directional Attention Flow. *Journal of Computing Theories and Applications*, *2*(4), 470–480. https://doi.org/10.62411/jcta.12376

Imron, S., Setiawan, E. I., Santoso, J., & Purnomo, M. H. (2023). Aspect based sentiment analysis marketplace product reviews using BERT, LSTM, and CNN. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 7(3), 586-591 https://doi.org/10.29207/resti.v7i3.4751

Susanto, N., & Pardede, H. F. (2024). Feature Learning using Deep Variational Autoencoder for Prediction of Defects in Car Engine. *2024 International Conference on Information Technology Research and Innovation (ICITRI)*, *2024*. https://doi.org/10.1109/ICITRI62858.2024.10699115

Wankhade, M., Chandra, A., Rao, S., & Kulkarni, C. (2022). *and challenges* (Issue 0123456789). Springer Nature. https://doi.org/https://doi.org/10.1007/s10462-022-10144-1

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2025 Nanang Susanto