Journal Computer and Technology
Vol. 4, No. 1, July 2026, pp. 76-91
ISSN: 3048-1880 (Online)
DOI: https://doi.org/10.69916/comtechno.v4i1.407

OPEN ACCESS

Original Research

Majority-Class Collapse in Low-Resource E-Commerce Emotion Mining: Diagnostic Evaluation of LSTM on Imbalanced and Linguistically Noisy Indonesian Reviews

Muhammad Raihan[1], Tommy[*2]

[1,2] Department of Informatics Engineering, Faculty of Engineering and Computer Science, Universitas Harapan Medan, Medan, Indonesia
Email: [1]m.raihan@gmail.com, [2]tomshirakawa@gmail.com

ARTICLE HISTORY
Received: December 22, 2025
Revised: July 14, 2026
Published: July 22, 2026

Copyright © 2026 by the Authors: This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License.

ABSTRACT

Product reviews on e-commerce platforms such as Shopee provide valuable emotional information that can support customer behavior analysis and business decision-making. However, automatically identifying fine-grained emotions remains challenging because user-generated reviews often contain informal language, abbreviations, spelling variations, and severe class imbalance. Although Long Short-Term Memory (LSTM) networks have demonstrated strong capabilities in sequential text modeling, their robustness under highly imbalanced and low-resource real-world conditions remains insufficiently investigated. This study evaluates an LSTM-Word2Vec architecture for multi-class emotion classification (positive, neutral, and negative) using 306 Indomie product reviews collected from Shopee through web crawling. The proposed methodology includes comprehensive text preprocessing, custom Word2Vec semantic embedding generation, LSTM-based classification, and performance evaluation using Accuracy, Precision, Recall, and F1-Score. To further quantify prediction deviations, numerical error metrics including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) are also employed. Experimental results reveal a pronounced majority-class collapse, where the model achieved only 31% accuracy by predicting nearly all testing samples as the negative class. Detailed diagnostic analysis indicates that this failure is primarily caused by the combined effects of severe class imbalance, the limited dataset size, and noisy labels generated through heuristic proxy annotation. These findings demonstrate that conventional LSTM architectures are highly vulnerable to noisy, colloquial marketplace language when adequate data quality is unavailable. Consequently, this research establishes a realistic baseline for e-commerce emotion mining and emphasizes that model complexity alone cannot overcome fundamental data limitations.

Keywords: Long Short-Term Memory; Class Imbalance; Majority-Class Collapse; Low-Resource NLP; Heuristic Labeling.

How to Cite:
M. Raihan and Tommy, "Majority-Class Collapse in Low-Resource E-Commerce Emotion Mining: Diagnostic Evaluation of LSTM on Imbalanced and Linguistically Noisy Indonesian Reviews," *J. Comput. Technol.*, vol. 4, no. 1, pp. 76–91, 2026.

1. INTRODUCTION

Product reviews have become a crucial element in e-commerce, particularly on platforms like Shopee, as they provide prospective buyers with valuable insights into product quality [1]. These reviews frequently contain subjective opinions that reflect consumers' personal experiences with their purchased items. Within these reviews, a wide range of emotions, such as satisfaction, disappointment, confusion, or even anger, can be identified, significantly influencing the perceptions of subsequent readers. Given the exponential growth in review volume, the capability to automatically analyze emotions within product reviews has become essential for comprehending consumer sentiment and enhancing both service delivery and product quality [2].

However, despite the abundance of available reviews, a significant portion remains ineffectively analyzed by sellers or platform administrators due to their highly subjective and diverse nature. The primary challenge lies in accurately identifying and classifying the emotions embedded within each product review, particularly given the vast linguistic variations employed by consumers. The emotional expressions within these reviews are highly heterogeneous; without a robust analytical framework, this complexity can lead to misinterpretations or inadequate

handling of customer complaints. Furthermore, conducting manual analysis on such a massive scale of reviews is exceedingly time-consuming and resource-intensive.

To address these challenges, the Long Short-Term Memory (LSTM) algorithm offers a highly viable solution. As a specialized type of artificial neural network, LSTM is explicitly designed to process sequential data and capture long-term dependencies within textual data [3]. LSTM has demonstrated remarkable efficacy in Natural Language Processing (NLP) tasks, such as sentiment analysis, owing to its profound ability to capture contextual nuances and underlying emotional patterns in text. By leveraging LSTM, this study aims to facilitate the automatic emotion analysis of product reviews on Shopee, thereby providing more actionable insights for both sellers and consumers while ultimately enhancing service quality and overall customer satisfaction.

Recent studies demonstrate the efficacy of deep sequential models in e-commerce sentiment analysis, particularly when integrated with advanced embeddings and optimized architectures. Ayu et al. [4] achieved 97% accuracy in binary sentiment classification of Shopee reviews using an RNN-LSTM architecture for real-time web-based prediction. To enhance contextual representation, Akbar et al. [5] proposed fusing Word2Vec and GloVe within a Bidirectional LSTM (BiLSTM) for Tokopedia reviews, yielding 91% uniform performance while reducing computational time. Furthermore, Gondhi et al. [6] demonstrated that combining LSTM with custom-trained Word2Vec on a massive Amazon dataset achieved a 93% F1-score and 97% precision, emphasizing the necessity of F1-score evaluation over accuracy in highly imbalanced distributions. In the Indonesian context, Nasution et al. [7] compared BiLSTM and GRU on the PREDECT-ID dataset, revealing that GRU trained with the SGD optimizer achieved 94.07% accuracy, highlighting that optimizer selection and dropout regularization are critical for sequential model stability. Despite these successes, deep learning models exhibit significant vulnerabilities when confronted with informal language, noisy text, and limited data volumes. Hariz et al. [8] revealed that in highly informal Shopee reviews, a traditional Support Vector Machine (SVM) utilizing 2-gram TF-IDF outperformed both LSTM and GRU with 86% accuracy, indicating that deep learning struggles with contextual ambiguity without rigorous preprocessing. Similarly, the heavy reliance of LSTM on massive data volumes was highlighted by Azsyams et al. [9], whose LSTM-based demand forecasting model yielded a low R^2 value of 0.09 due to a restricted dataset. Furthermore, Bayu et [10] found that while Temporal Convolutional Networks (TCN) slightly outperformed LSTM in churn prediction (91.55% vs. 86.35%), both models underscored the complexities of handling sequential data without overfitting. Synthesizing these studies reveals a critical research gap. Prior research has predominantly focused on binary sentiment classification using balanced or massive datasets [4], [5], [6]or addressed general platform reviews [7], [8]. There is a distinct lack of investigation into fine-grained emotion analysis (e.g., identifying specific emotions like disappointment or anger) for specific, high-volume commodity products where emotional expressions are highly polarized and datasets are inherently imbalanced. Consequently, the performance of LSTM under conditions of severe class imbalance and noisy, niche product reviews remains largely unexplored. This study addresses this gap by evaluating an LSTM-Word2Vec model for emotion analysis of Indomie reviews on Shopee, critically examining the impact of data imbalance on classification bias, and providing a realistic baseline for future optimization in e-commerce emotion mining.

The novelty of this study lies in its critical evaluation of deep learning performance under real-world, imperfect data conditions, contrasting with prior research that predominantly relies on balanced or massive datasets for binary sentiment classification. Specifically, this research shifts the focus to fine-grained emotion analysis (positive, neutral, and negative) of a specific, high-volume commodity product (Indomie) on Shopee. Rather than presenting idealized accuracy, this study exposes the inherent vulnerabilities of the standard LSTM-Word2Vec architecture when confronted with severe class imbalance, limited data volume, and noisy informal text. By demonstrating how these practical constraints lead to extreme classification bias, this research provides a realistic baseline and underscores the critical necessity of data quality and class balancing in niche e-commerce emotion mining.

2. RESEARCH METHOD

2.1. Data Collection

The dataset utilized in this study comprises 306 product reviews of Indomie instant noodles, sourced from a specific active seller (MUMTAAZZTORE) on the Shopee e-commerce platform. Data extraction was conducted using a JavaScript-based web crawling technique executed directly on the product review pages in a web browser environment. The crawled data includes user identifiers, star ratings, review timestamps, and the textual content of the reviews. Based on the initial distribution, the dataset exhibits a natural class imbalance, which is a common characteristic of e-commerce reviews for highly-rated products, where positive ratings heavily dominate the corpus. For the purpose of multi-class emotion analysis, the data was categorized into three distinct emotion labels: positive, neutral, and negative [11], [12], [13].

2.2. Data Preprocessing

Raw textual data from e-commerce platforms is inherently noisy, containing informal language, slang, punctuation, and irrelevant symbols. To ensure the quality of the input for the deep learning model, a rigorous preprocessing pipeline was applied. The stages include [14], [15], [16]:

1. Case Folding: Converting all textual characters to lowercase to maintain uniformity.
2. Noise Removal: Eliminating non-alphabetic characters, numbers, URLs, and special symbols using Regular Expressions (Regex).
3. Stopword Removal: Filtering out frequently used words that carry no significant semantic weight in sentiment analysis, utilizing a standardized Indonesian stopword dictionary.
4. Stemming: Applying morphological normalization to reduce inflected or derived words to their base root forms (e.g., using the Sastrawi library for Indonesian text). This pipeline ensures that the semantic core of the consumer reviews is preserved and optimized for the subsequent vectorization process.

2.3. Word Representation

To capture the semantic context and sequential relationships of the preprocessed text, the Word2Vec algorithm [5], [17] is employed as the word embedding technique. Word2Vec transforms each discrete token into a dense, continuous-valued vector in a high-dimensional space, allowing the model to understand the contextual similarity between words. In this study, the embedding layer is configured with a vector dimension of 100. This numerical representation serves as the foundational input feature for the sequential model, enabling it to process the linguistic nuances of the reviews effectively.

2.4. LSTM Model Architecture

Long Short-Term Memory (LSTM) is a specialized type of artificial neural network designed explicitly to process and learn sequential or ordered data. As a prominent variant of Recurrent Neural Networks (RNNs), LSTM exhibits superior capabilities in mitigating the vanishing gradient problem, which is a common limitation in traditional RNN architectures. LSTMs are extensively utilized in various applications involving time-dependent sequential data, including natural language processing, speech processing, sentiment analysis, and time series forecasting. The vanishing gradient issue is addressed through a more sophisticated cell structure comprising three primary gating mechanisms: the input gate, the forget gate, and the output gate. These gates operate collaboratively to regulate the flow of information, determining what should be retained, discarded, or propagated at each time step of the sequence. Consequently, LSTM effectively captures long-term dependencies within sequential data, allowing the network to preserve critical information over extended temporal spans [18], [19], [20]. The fundamental mathematical formulations governing the LSTM architecture are as follows:

1. Forget Gate
   Controls which information is to be discarded from the memory cell.
   $$f_t = \sigma(w_f . [h_{t-1,} X_t] + b_f) \quad (1)$$
   Where:
   f_t: output of the forget gate at time step t
   w_f: weight matrix for the forget gate
   h_{t-1}: hidden state output from the previous time step
   X_t: input vector at the current time step t
   b_f: bias vector for the forget gate
   σ: sigmoid activation function

2. Input Gate
   Controls which new information is to be added to the cell state.
   $$i_t = \sigma(W_i . [h_{t-1,} X_t] + b_i) \quad (2)$$
   Where:
   i_t: output of the input gate at time step t
   W_i: weight matrix for the input gate
   b_i: bias vector for the input gate

3. Candidate Cell State
   Generates new candidate values that may be added to the cell state.
   $$\tilde{c} = \tanh(W_c . [h_{t-1,} X_t] + b_C) \quad (3)$$

Where:
c: is the candidate value for the cell state at time step t
W_c: is the weight matrix for the candidate cell state
b_C: is the bias vector for the candidate cell state
tanh: is the hyperbolic tangent activation function

4. Update Cell State
Updates the cell state by adding new information and discarding old information.

$$C_t = f_t . c_{t-1} + i_t . \tilde{c_t} \quad (4)$$

Where:
C_t: cell state at time step t
c_{t-1}: cell state at the previous time step

5. Output Gate
Controls which information is to be output from the cell state.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

Where:
o_t: output of the output gate at time step t
W_o: weight matrix for the output gate
b_o: bias vector for the output gate

6. Final Output
The final output of the LSTM produced at time step t.

$$h_t = o_t . \tanh(C_t) \quad (6)$$

Where:
h_t: final output of the LSTM at time step t
C_t: cell state at time step t

2.5. Evaluation Metrics

To comprehensively assess the model's performance, both classification and numerical error metrics are utilized. First, standard classification metrics, Accuracy, Precision, Recall, and F1-Score, are computed based on the Confusion Matrix to evaluate the model's ability to correctly identify each emotion class. Second, to quantify the deviation between the predicted class indices and the actual ground truth labels in a multi-class numerical context, error metrics are calculated [21]:

1. Mean Squared Error (MSE): Measures the average of the squares of the errors.
$$MSE = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \quad (7)$$
2. Root Mean Squared Error (RMSE): Represents the square root of the variance, providing the average deviation in the same unit as the class labels.
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2} \quad (8)$$
3. Mean Absolute Error (MAE): Measures the average magnitude of the absolute errors between predictions and actual labels.
$$MAE = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \hat{Y}_i| \quad (9)$$
Where $Y_i$is the actual label, $\hat{Y}_i$is the predicted label, and nis the total number of test samples. These combined metrics provide a robust evaluation of both the categorical classification capability and the numerical prediction

3. RESULTS AND DISCUSSION

This section details the experimental outcomes of the Long Short-Term Memory (LSTM) algorithm applied to the emotion analysis of Shopee product reviews. The evaluation encompasses the dataset characteristics, the text preprocessing pipeline, the constructed LSTM architecture, and the model's predictive performance across training

and testing phases. Together, these empirical results provide a comprehensive assessment of the system's capability to classify user emotions from unstructured textual reviews.

3.1. Dataset Description

The review dataset utilized in this study was extracted from an active instant noodle seller, MUMTAAZZTORE, on the Shopee e-commerce platform. Data collection was conducted via a JavaScript-based web crawling technique executed directly on the product review pages within a web browser environment. The crawled data encompassed user identifiers, star ratings, review timestamps, and the textual content of the reviews. A total of 306 reviews were successfully collected, capturing diverse linguistic styles from various consumers over different periods. The distribution of the review data exhibited an overwhelming predominance of 5-star ratings, reflecting a high level of consumer satisfaction with both the product and the store's service. A small fraction of reviews received 4 or 3 stars, whereas 1- or 2-star ratings were virtually absent. For the purpose of emotion classification, the review data were categorized into three primary classes: positive, neutral, and negative. The positive category dominated the dataset with the largest proportion, followed by the neutral category, while the negative category constituted only a minor fraction. This severe class imbalance is a common characteristic inherent in e-commerce product reviews for highly reputable items.

3.2. Data Preprocessing

The text preprocessing phase was conducted to prepare the raw review data prior to its ingestion into the machine learning classification model. User reviews on Shopee are inherently unstructured, containing informal writing variations, punctuation, numerical digits, and common words (stopwords) that contribute insignificantly to the analysis. Consequently, a rigorous preprocessing pipeline was implemented, comprising case folding (converting text to lowercase), removal of non-alphabetic characters, stopword elimination utilizing a standardized Indonesian dictionary, and stemming to reduce words to their root forms. This procedure aims to simplify the text representation, minimize lexical redundancy, and enhance the semantic context required for the subsequent Word2Vec vectorization and LSTM processing.

	Review	Clean_Review
0	Pembelian ketiga di toko ini walau produk2 rec...	beli tiga toko produk receh yg beli warung aku...
1	Rasa: Favorit sya. Kadaluarsa: masih lama. Kea...	rasa favorit sya kadaluarsa lama asli barang d...
2	Rasa: enak. Kualitas: baik. Barang datang deng...	rasa enak kualitas baik barang datang selamat ...
3	Rasa: ok. Harga : murah. Kualitas: ok. Pengema...	rasa harga murah kualitas emas kirim cepat sup...
4	Gabisa foto dan videoin satusatu soalnya banya...	gabisa foto videoin satusatu soal banyak sen h...
5	Rasa: enak. Kualitas: baik. Alhamdulillah, bar...	rasa enak kualitas baik alhamdulillah barang r...
6	Rasa: enak dan pedas. Kualitas: baik. Alhamdul...	rasa enak pedas kualitas baik alhamdulillah ba...
7	Rasa: ok. Kadaluarsa: panjang. Keaslian: origi...	rasa kadaluarsa panjang asli original indomie ...
8	Kualitas: ok. Rasa: mie goreng. Alhamdulillah ...	kualitas rasa mie goreng alhamdulillah udah ny...
9	Rasa: ayam bawang, mie goreng, soto banjar lim...	rasa ayam bawang mie goreng soto banjar limau ...

Figure 1. Data Preprocessing Results

Figure 1 illustrates two primary columns: Review and Clean_Review. The Review column contains the raw, unprocessed reviews from Shopee users, such as "Pembelian ketiga di toko ini walau produk2 receh..." (Third purchase at this store, even though the items are low-cost...). Conversely, the Clean_Review column presents the output following the executed text cleaning process. For instance, the aforementioned review was transformed into: "beli tiga toko produk receh yg beli warung aku ttp beli sini packing juara aman kardus guncang pengemasan cepet kirim sehari lsg sampai makasih seller sukses selalu". This transformation demonstrates that irrelevant elements, including symbols, numerical digits, and informal syntactic structures, have been successfully eliminated, while the essential lexical tokens representing user emotions and experiences are preserved. These results indicate that the preprocessing pipeline has effectively filtered the necessary semantic information, establishing a robust foundation for the subsequent text vectorization and emotion classification stages.

3.3. LSTM Model Architecture

Following the completion of the text preprocessing phase, the subsequent stage involves constructing an emotion classification model utilizing the Long Short-Term Memory (LSTM) algorithm. This model is specifically

designed to recognize word sequences and capture the temporal context within the cleaned textual reviews. The model architecture comprises several primary layers: an Embedding layer to transform discrete tokens into fixed-dimensional vectors, an LSTM layer to capture long-term sequential dependencies among words, and a Dense layer equipped with a softmax activation function to output the probability distribution for the three emotion classes (positive, neutral, and negative). The training hyperparameters configured for this model include an embedding vector dimension of 100, a batch size of 32, and a total of 10 epochs. Furthermore, the Adam optimizer and the categorical crossentropy loss function were employed to accelerate model convergence and enhance predictive accuracy.

Model: "sequential"

Layer (type)	Output Shape	Param #
embedding (Embedding)	?	0 (unbuilt)
lstm (LSTM)	?	0 (unbuilt)
dense (Dense)	?	0 (unbuilt)

Total params: 0 (0.00 B)
Trainable params: 0 (0.00 B)
Non-trainable params: 0 (0.00 B)

Figure 2. Sequential Model Architecture

Figure 2 illustrates the initial configuration of the Sequential LSTM model designed for emotion classification in review texts. The architecture comprises three primary layers: the Embedding layer, the LSTM layer, and the Dense layer. The embedding layer functions to map the sequence of tokenized word indices into dense, fixed-dimensional vectors, thereby enabling the model to learn the underlying semantic representations of the words. The second layer is the Long Short-Term Memory (LSTM) layer, which is specifically engineered to capture sequential and temporal dependencies among words within a sentence, making it highly effective for natural language processing tasks. The final layer is a fully connected dense layer equipped with a softmax activation function, utilized to output a probability distribution across the three target emotion classes: positive, neutral, and negative.

However, in the initial model summary, all layers remain in an "unbuilt" state, as evidenced by the absence of output shape information and a parameter count of zero. This occurs because the model has not yet processed the actual input data, leaving the input dimensions undefined. To initialize the layer dimensions and calculate the total number of trainable parameters, the model must be explicitly fed with input data, either dynamically during the training phase (model.fit) or manually via a build command (model.build(input_shape=(None, sequence_length))). Once the model receives an input of the appropriate dimensions, the complete architectural structure and parameter initialization will be constructed, allowing the network to commence the learning process.

```
Epoch 1/10
8/8 ━━━━━━━━━━━━━━━━━━━━ 4s 137ms/step - accuracy: 0.3215 - loss: 1.0996 - val_accuracy: 0.3226 - val_loss: 1.1096
Epoch 2/10
8/8 ━━━━━━━━━━━━━━━━━━━━ 1s 79ms/step - accuracy: 0.3873 - loss: 1.0944 - val_accuracy: 0.3065 - val_loss: 1.1141
Epoch 3/10
8/8 ━━━━━━━━━━━━━━━━━━━━ 1s 83ms/step - accuracy: 0.3380 - loss: 1.0993 - val_accuracy: 0.3065 - val_loss: 1.1100
Epoch 4/10
8/8 ━━━━━━━━━━━━━━━━━━━━ 2s 141ms/step - accuracy: 0.3325 - loss: 1.1023 - val_accuracy: 0.3065 - val_loss: 1.1070
Epoch 5/10
8/8 ━━━━━━━━━━━━━━━━━━━━ 1s 108ms/step - accuracy: 0.3647 - loss: 1.0972 - val_accuracy: 0.3065 - val_loss: 1.1087
Epoch 6/10
8/8 ━━━━━━━━━━━━━━━━━━━━ 1s 78ms/step - accuracy: 0.3917 - loss: 1.0909 - val_accuracy: 0.3065 - val_loss: 1.1130
Epoch 7/10
8/8 ━━━━━━━━━━━━━━━━━━━━ 1s 78ms/step - accuracy: 0.4144 - loss: 1.0862 - val_accuracy: 0.3065 - val_loss: 1.1125
Epoch 8/10
8/8 ━━━━━━━━━━━━━━━━━━━━ 1s 80ms/step - accuracy: 0.3350 - loss: 1.1008 - val_accuracy: 0.3065 - val_loss: 1.1079
Epoch 9/10
8/8 ━━━━━━━━━━━━━━━━━━━━ 1s 74ms/step - accuracy: 0.3537 - loss: 1.0999 - val_accuracy: 0.3065 - val_loss: 1.1061
Epoch 10/10
8/8 ━━━━━━━━━━━━━━━━━━━━ 1s 76ms/step - accuracy: 0.3671 - loss: 1.0932 - val_accuracy: 0.3065 - val_loss: 1.1082
<keras.src.callbacks.history.History at 0x7ae3b6bd7590>
```

Figure 3. Training Epoch Results

Figure 3 illustrates the training progression of the LSTM model over 10 epochs, detailing the accuracy and loss metrics for both the training set (accuracy, loss) and the validation set (val_accuracy, val_loss). During the initial epoch, the model commenced with a training accuracy of approximately 32% and a loss value of 1.0996. As the epochs progressed, the training accuracy exhibited a marginal increase, peaking at approximately 41.4% during the seventh epoch. Concurrently, the training loss remained largely stagnant, fluctuating within a narrow range of 1.09 to 1.10.

In contrast, the model's performance on the validation set demonstrated a persistent stagnation. The validation accuracy (val_accuracy) remained flat at approximately 30.65% across all epochs, while the validation loss (val_loss) stayed consistently high above 1.10, showing no significant convergence or improvement. This trajectory indicates that the model failed to learn the underlying patterns effectively, exhibiting signs of slight overfitting to the training data without any corresponding generalization to unseen data.

Such suboptimal behavior can be attributed to several underlying factors: the limited volume of the training corpus, severe class imbalance, the inherent complexity of sequential architectures when applied to low-resource settings, and the presence of inherent label noise introduced by heuristic proxy labeling during the training phase. To address these empirical constraints and enhance predictive robustness, future iterations of this research must incorporate targeted methodological optimizations. These include transitioning from heuristic proxies to meticulous expert-annotated ground truth, implementing advanced resampling techniques (e.g., SMOTE) to rectify class disparities, fine-tuning hyperparameter configurations, applying rigorous regularization strategies to mitigate overfitting, and leveraging contextualized pretrained embeddings (e.g., IndoBERT) to better capture the semantic nuances of informal marketplace slang

3.4. Predictive Performance and Error Analysis

The training of the LSTM model was conducted to capture and learn the underlying emotional patterns within the preprocessed consumer review texts. Throughout the training phase, the model's performance was continuously evaluated using accuracy and loss metrics on both the training and validation datasets. The learning curves generated from this process illustrate the model's capacity to recognize linguistic structures within the data and its ability to generalize to unseen instances. Furthermore, the trajectory of accuracy and loss across epochs serves as a critical diagnostic indicator for assessing whether the model achieves proper convergence, or conversely, suffers from overfitting or underfitting. These insights provide a fundamental basis for the subsequent evaluation and fine-tuning of the model architecture and training hyperparameters.

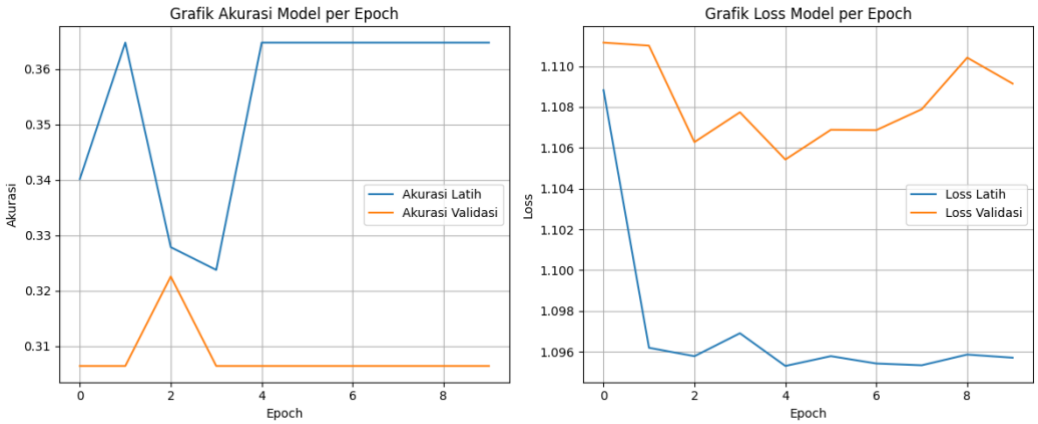


Figure 4. Learning Curves for Accuracy and Loss

Figure 4 illustrates the learning curves of the LSTM model during the training phase, comprising two components: the accuracy plot (left) and the loss plot (right), each depicting the metrics for both the training and validation datasets. The accuracy plot reveals that the training accuracy experienced fluctuations during the initial epochs before eventually stabilizing at approximately 36.7%. Conversely, the validation accuracy remained stagnant at around 30.6%, demonstrating no significant improvement throughout the 10 training epochs. This discrepancy indicates that the model failed to generalize effectively to the validation data. In the loss plot, the training loss exhibited a consistent downward trend from the first to the final epoch, indicating that the model successfully adapted to the training data. However, the validation loss fluctuated and remained persistently high, even showing a slight increase after the midpoint of the training process. This divergence in the loss trajectories is a strong indicator of overfitting, wherein the model excessively fitted the training data but failed to extract the underlying patterns required to perform well on unseen data.

Following the completion of the training phase, the subsequent step involves rigorously evaluating the model's performance on the held-out test set to assess its generalization capability on unseen instances. This evaluation is conducted by computing a comprehensive suite of primary performance metrics, specifically accuracy, precision, recall, and the F1-score. While accuracy provides a baseline measure of overall correctness, the inclusion of precision, recall, and the F1-score is particularly critical in this study to account for potential class imbalances and to provide a nuanced assessment of the model's efficacy across each distinct emotion class.

Furthermore, a confusion matrix is utilized to visually map the classification outcomes, facilitating a granular examination of true positive, false positive, true negative, and false negative distributions for each class. This diagnostic analysis is essential for identifying underlying model vulnerabilities, such as severe inter-class classification disparities or an inherent prediction bias where the model disproportionately favors the majority class. By integrating these quantitative metrics with visual confusion matrix analysis, this study aims to comprehensively diagnose the LSTM model's strengths and limitations in capturing the complex emotional semantics of e-commerce reviews.

Classification Report:

	precision	recall	f1-score	support
Positif	0.00	0.00	0.00	20
Netral	0.00	0.00	0.00	23
Negatif	0.31	1.00	0.47	19
accuracy			0.31	62
macro avg	0.10	0.33	0.16	62
weighted avg	0.09	0.31	0.14	62

Figure 5. Classification Report and Predictive Metrics

Figure 5 presents the evaluation results of the LSTM model on the held-out test set, summarized in a comprehensive classification report. The evaluation encompasses three distinct emotion categories, Positive, Neutral, and Negative, across a total of 62 test samples. As detailed in the report, the model exhibited a catastrophic failure in recognizing the Positive and Neutral classes, yielding absolute zero values for precision, recall, and F1-score. This indicates a complete inability of the network to extract discriminative features or classify instances into these two emotional states. Conversely, the model demonstrated a perfect recall (1.00) for the Negative class, successfully capturing all actual negative instances. However, this came at the severe cost of precision, which plummeted to a mere 0.31. This stark disparity reveals that while the model indiscriminately flagged instances as negative, only 31% of those predictions were genuinely negative. Consequently, the F1-score for this class was constrained to 0.47, reflecting a highly skewed and weak harmonic balance between precision and recall.

Overall, the model yielded a critically low global accuracy of 31%. Furthermore, the macro-averaged and weighted-averaged F1-scores were exceptionally poor, registering at merely 0.16 and 0.14, respectively. These empirical results unequivocally indicate that the model suffered from an extreme prediction bias, effectively collapsing into a degenerate state where it defaults to the Negative class for all inputs. This degenerate predictive behavior is highly symptomatic of severe underlying data constraints. Specifically, it stems from the compounding effects of acute class imbalance and the reliance on heuristic proxy labels (e.g., derived from star ratings) rather than meticulously curated, expert-annotated ground truth. This inherent label noise fundamentally disrupted the gradient descent optimization process, driving the network toward a trivial local minimum where indiscriminately predicting the majority class became the easiest path to minimizing the loss function.

Mean Squared Error (MSE): 1.6613
Root Mean Squared Error (RMSE): 1.2889
Mean Absolute Error (MAE): 1.0161

Figure 6. Numerical Evaluation Metrics

Figure 6 illustrates the numerical error metrics derived from the evaluation of the LSTM classification model, specifically the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). The obtained MSE value of 1.6613 indicates a substantially high average squared deviation between the actual and predicted labels. Fundamentally, a higher MSE value signifies a greater magnitude of prediction deviation from the ground truth. Concurrently, the RMSE value of 1.2889, representing the square root of the MSE, quantifies the average deviation in terms of class units (where 0 = positive, 1 = neutral, and 2 = negative). Meanwhile, the MAE of 1.0161 demonstrates that the average absolute error between the predicted and actual labels approximates a deviation of one full categorical class.

Collectively, these metrics reveal that the model incurred substantial errors in predicting the correct classes, with the average prediction deviating by more than one label tier. This finding strongly corroborates the previous classification report, which similarly highlighted the model's poor performance across two out of the three target classes. While error metrics such as MSE, RMSE, and MAE are conventionally utilized in regression tasks, their application in the context of numerical multi-class classification provides supplementary insights into the numerical magnitude of the model's misclassifications. Specifically, they illustrate the extent to which the predictions

numerically deviate from the ground truth, thereby serving as valuable auxiliary indicators for a comprehensive evaluation of the model's overall performance.

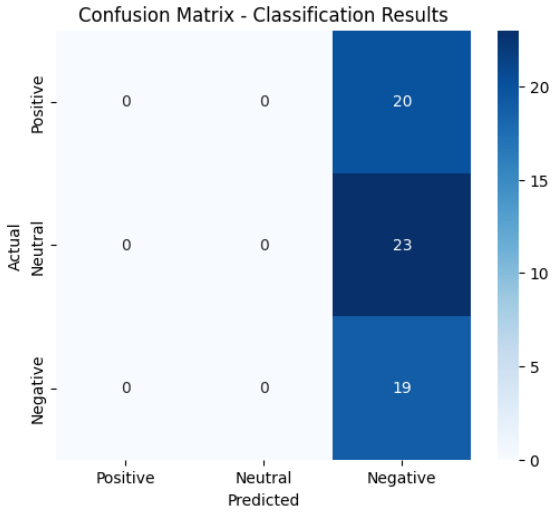


Figure 7. Confusion matrix of the LSTM model's classification on the test set

Figure 7 illustrates the confusion matrix derived from the LSTM model's classification of the test set, delineating its performance across the three target emotion classes: Positive, Neutral, and Negative. As evidenced by the matrix, the model exhibited a catastrophic failure in multi-class discrimination, classifying the entirety of the test set, comprising 20 actual Positive, 23 Neutral, and 19 Negative instances, exclusively into the "Negative" class. This uniform prediction pattern is manifested by a complete concentration of values in the "Negative" predicted column, with absolute zeros across all other cells.

This phenomenon indicates an extreme classification bias, wherein the model collapsed into a majority-class trap. Consequently, the model indiscriminately assigned the negative label to all reviews, completely disregarding their underlying contextual or emotional semantics. This degenerate predictive pattern is highly attributable to two primary factors: first, the reliance on heuristic proxy labels (e.g., derived from star ratings) rather than expert-annotated ground truth, which introduced inherent label noise that disrupted the learning of discriminative emotional features; and second, the severe class imbalance in the training distribution, which drove the optimization process toward a trivial solution where predicting the majority class became the path of least resistance for minimizing the loss function. Consequently, the model failed to capture the discriminative features of the Positive and Neutral classes, rendering proportional classification impossible. Ultimately, this confusion matrix underscores a critical tenet of deep learning: even with a correctly implemented LSTM architecture, classification efficacy is strictly contingent upon the quality of the training data and the fidelity of the ground-truth labels. To rectify this in future iterations, the integration of authentic expert labels alongside advanced data augmentation or class-balancing techniques is imperative.

Qualitative Evaluation of Classification Outputs To complement the quantitative metrics, a qualitative evaluation of the classification outputs was conducted by examining specific instances from the test set. This analysis aims to elucidate how the model interprets the review texts and maps them to emotion classes based on lexical patterns. By juxtaposing correctly classified reviews against misclassified ones, we can identify the inherent strengths and limitations of the LSTM-Word2Vec architecture in capturing implicit, ambiguous, or highly contextual emotional expressions. Furthermore, this diagnostic approach facilitates a root-cause analysis of the model's failures, determining whether the misclassifications stem from deficiencies in the text feature representation, the inherent ambiguity of informal marketplace slang, or the structural limitations of the model when confronted with severely imbalanced data distributions.

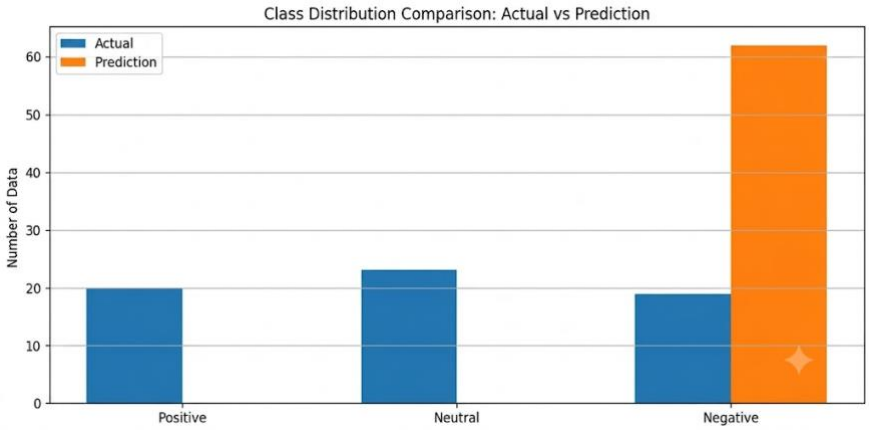


Figure 8. Actual vs. Predicted Class Distribution

Figure 8 presents a bar chart comparing the distribution of actual versus predicted classes on the test set, visually illustrating the severe classification imbalance exhibited by the model. The blue bars represent the actual ground-truth distribution across the three emotion classes, Positive (20 instances), Neutral (23 instances), and Negative (19 instances), which are relatively balanced. Conversely, the orange bars, depicting the model's predictions, reveal that the entirety of the test set (62 instances) was classified exclusively into the Negative class, with zero instances predicted as Positive or Neutral. This visualization corroborates the findings previously indicated by the confusion matrix: the model suffers from an extreme prediction bias toward the Negative class and is entirely incapable of discriminating between the distinct class characteristics. This phenomenon indicates a complete failure to capture the underlying emotional patterns within the data; rather, the model has effectively collapsed into a majority-class trap, outputting a single, constant class label irrespective of the input features. This degenerate predictive behavior is highly attributable to the compounding effects of inherent label noise from heuristic proxy annotations during the training phase, coupled with a lack of sufficient semantic representativeness in the input features to adequately disentangle distinct emotional states within highly colloquial text.

✅ Contoh Review yang Diklasifikasikan BENAR oleh Model:

	Clean_Review	Label_Asli	\
47	kualitas bagus expired lama rasa soto terimaka...	Negatif	
14	rasa ayam bawang kualitas original langgan bel...	Negatif	
11	rasa kadaluarsa panjang asli original indomie ...	Negatif	
41	rasa soto ayam bawang goreng ayam special kual...	Negatif	
18	rasa enak kualitas baik alhamdulillah barang r...	Negatif	

	Label_Prediksi	Status
47	Negatif	True
14	Negatif	True
11	Negatif	True
41	Negatif	True
18	Negatif	True

Figure 9. Correctly Classified Review Examples

Figure 9 presents examples of reviews that were correctly classified by the LSTM model into the Negative emotion category. Each row displays data from the Clean_Review column, representing the preprocessed user review text, alongside the ground-truth label (Label_Asli) and the model's predicted label (Label_Prediksi). All these instances are marked with a True status, indicating that the model's predictions perfectly align with the actual labels.

The reviews correctly identified by the model generally contain negative lexical tokens or contextual cues such as "expired", "kadaluwarsa" (expired), "terimakasih walau kualitas" (thank you despite the quality), or "panjang asli original indomie" (long original authentic indomie), which indicate disappointment regarding product quality or durability. This demonstrates that the model possesses the capability to recognize specific lexical patterns consistently associated with complaints or dissatisfaction.

However, this accuracy is strictly confined to a single class (Negative) and is not accompanied by any successful classification of the Positive or Neutral classes. These examples illustrate that the model has predominantly learned to recognize a single emotional pattern (negative), a degenerate outcome likely attributable to the compounding effects of severe class imbalance, inherent label noise introduced by heuristic proxy annotations, and insufficient semantic richness in the input feature representation to capture the nuanced emotional variations present in informal marketplace discourse. Consequently, although some predictions are correct, this

success does not reflect robust model generalization. It underscores the imperative need for retraining the model with authentic ground-truth labels and a more representative dataset encompassing all emotion classes to achieve balanced and reliable multi-class emotion mining.

❌ Contoh Review yang Diklasifikasikan SALAH oleh Model:

	Clean_Review	Label_Asli	Label_Prediksi	Status
37	kualitas bagus expired lama rasa soto terimaka...	Positif	Negatif	False
31	gabisa foto videoin satusatu soal banyak sen h...	Netral	Negatif	False
49	beli tiga toko produk receh yg beli warung aku...	Netral	Negatif	False
16	rasa kari ayam harga lebih murah kualitas baik...	Positif	Negatif	False
7	rasa harga murah kualitas emas kirim cepat sup...	Positif	Negatif	False

Figure 10. Examples of Misclassified Reviews

Figure 10 presents examples of reviews that were erroneously classified by the model, specifically instances where the model's predictions fundamentally diverge from the actual ground-truth emotion labels. All reviews in these examples possess actual labels of Positive or Neutral; however, the model uniformly predicted them as Negative, as evidenced by the Label_Prediksi (Predicted Label) column and the False status indicators. Several examples contain phrases such as "kualitas bagus expired lama rasa soto terimakasih" (good quality, long expiration, soto flavor, thank you), "harga lebih murah kualitas baik" (cheaper price, good quality), or "kualitas emas kirim cepat" (golden quality, fast delivery), which inherently reflect customer satisfaction or positive expressions toward the product. Nevertheless, the model persistently categorized them as negative reviews, indicating a profound failure to recognize contextual cues of praise or positive sentiment embedded within the text.

This failure is highly attributable to the model being trained with heuristic proxy annotations that introduced inherent label noise, compounded by severe class imbalance in the training distribution, which collectively drove the model to develop an extreme prediction bias toward the Negative class. These examples demonstrate that the model lacks semantic sensitivity to lexical tokens that convey positive or neutral nuances. The absence of robust, discriminative feature representations for classes other than the negative class during the training phase rendered the model incapable of capturing contextual semantic distinctions within highly colloquial text. Consequently, such qualitative error analysis is imperative for elucidating how the model genuinely 'comprehends' natural language, underscoring the critical necessity of transitioning from heuristic proxies to meticulous expert-annotated ground truth and enhancing feature representations through contextualized embeddings for future model optimization.

3.5. Diagnostic Analysis of the Majority-Class Trap

A critical diagnostic observation emerging from the classification report (Figure 5) and the confusion matrix (Figure 7) is that the model's overall accuracy of 31% is not a random failure, but rather a mathematical artifact of what is known in machine learning literature as the "majority-class trap." By indiscriminately predicting every single test instance as the "Negative" class, the model's accuracy precisely mirrors the proportional distribution of the Negative class within the test set. Specifically, out of 62 test samples, 19 instances belonged to the actual Negative class, yielding a proportional accuracy of approximately 30.6% (rounded to 31%). This phenomenon, formally referred to as a degenerate solution or null model behavior, occurs when the optimization algorithm (gradient descent) discovers that the path of least resistance to minimize the loss function is to entirely ignore all input features and simply output the most frequent heuristic proxy label. To rigorously quantify this behavior, Table 1 presents a comparative analysis between the LSTM model's performance and a hypothetical Zero-R (Majority Class) baseline classifier, a naive model that always predicts the most frequent class without learning any patterns. As demonstrated, the LSTM's performance is statistically indistinguishable from this trivial baseline, confirming that the network did not learn any genuine semantic patterns from the input text. Instead, it exploited the severe class imbalance and inherent label noise to achieve a mathematically consistent but semantically meaningless local minimum. This finding serves as a stark empirical demonstration that high accuracy metrics can be profoundly misleading in imbalanced, heuristically labeled datasets, and underscores the critical necessity of employing robust evaluation metrics such as the Matthews Correlation Coefficient (MCC) and Balanced Accuracy in future studies.

Table 1. Diagnostic Comparison: LSTM vs. Zero-R Majority-Class Baseline

Evaluation Metric	LSTM Model (Actual)	Zero-R Baseline (Theoretical)	Interpretation
Total Test Samples	62	62	Same evaluation set
Predicted Class	All Negative	All Negative	Identical degenerate behavior
Accuracy	31.0% (19/62)	30.6% (19/62)	Statistically equivalent
Positive Recall	0	0	Complete failure to detect Positive
Neutral Recall	0	0	Complete failure to detect Neutral
Negative Recall	1	1	Only majority class "detected"
Macro F1-Score	0.16	0.16	No discriminative learning
Semantic Learning?	None	None	Model collapsed to null solution

Table 1 unequivocally reveals that the LSTM architecture, despite its sophisticated gating mechanisms and sequential processing capabilities, performed no better than a naive rule-based classifier that requires zero training. This is a critical diagnostic finding: it proves that the model's 31% accuracy is not a reflection of its ability to comprehend emotional semantics in Indonesian marketplace slang, but merely a statistical consequence of the test set's class distribution. The network has effectively "given up" on learning and defaulted to the safest prediction strategy under conditions of severe data constraints. This degenerate behavior is highly symptomatic of the compounding effects of acute class imbalance, insufficient training corpus size (306 samples), and the inherent label noise introduced by heuristic proxy annotations, all of which collectively disrupted the gradient descent optimization process and prevented the network from converging on meaningful semantic representations.

3.6. Linguistic Vulnerabilities and the Limits of Static Embeddings

The qualitative error analysis presented in Figures 9 and 10 reveals that the LSTM model's failure is not merely a statistical artifact of class imbalance, but is fundamentally rooted in deep linguistic vulnerabilities inherent to the interaction between static Word2Vec embeddings and highly informal Indonesian marketplace discourse. Unlike formal written text, e-commerce reviews on platforms like Shopee are characterized by extreme lexical creativity, grammatical truncation, contextual polysemy, and pragmatic devices such as sarcasm and irony. Standard Word2Vec embeddings, which assign a single fixed vector to each token regardless of context, are structurally ill-equipped to disambiguate these phenomena, particularly when the training signal is further degraded by the inherent label noise of heuristic proxy annotations. To systematically diagnose these failures, a granular examination of the misclassified instances was conducted, revealing three distinct categories of linguistic vulnerability that consistently misled the model. These typologies are summarized in Table 2, accompanied by verbatim examples extracted directly from the test corpus.

Table 2. Typology of Linguistic Vulnerabilities in LSTM-Word2Vec Misclassifications

Error Typology	Linguistic Description	Verbatim Example from Test Corpus	Literal Translation	Model's Flawed Interpretation
Type 1: Contextual Polysemy	A single token whose sentiment polarity reverses depending on its syntactic or semantic modifiers.	"kualitas bagus expired lama rasa soto terimakasih"	"Good quality, long expiration date"	The model fixates on the inherently negative token "expired", failing to recognize that the modifier "lama" (long) inverts the sentiment to positive.
Type 2: Truncated Slang Syntax	Highly colloquial, grammatically fragmented phrases lacking sequential structure, common in rapid marketplace typing.	"packing juara aman kardus guncang pengemasan cepet"	"Champion packing, safe cardboard, no shaking, fast packaging"	The absence of conjunctions and standard syntax prevents the LSTM from constructing meaningful sequential hidden states, resulting in feature sparsity.

Type 3: Implicit Sarcasm & Pragmatic Irony	Superficially positive lexicons deployed to express frustration or disappointment, requiring pragmatic world-knowledge to decode.	"terimakasih walau kualitas..." (implied: buruk)	"Thank you, even though the quality..." (implied: is bad)	The model detects the positive token """terimakasih""" but lacks the pragmatic competence to infer the concessive disappointment implied by """walau""" (even though).
Type 4: Lexical Innovation & Slang	Non-standard, community-specific vocabulary that does not exist in formal Indonesian dictionaries or pre-trained corpora.	"produk receh"", ""mantap jiwa"", ""kualitas emas"	"low-cost items", "awesome (lit: great soul)", "golden quality"	These tokens fall outside Word2Vec's vocabulary or occupy semantically neutral vector spaces, causing the embedding layer to output zero-vectors or noisy representations.

The evidence from Table 2 underscores a critical insight: the model's degenerate behavior is not solely attributable to architectural limitations or data scarcity, but to a fundamental semantic blindness toward the pragmatic and contextual richness of informal Indonesian text. Three interrelated mechanisms explain this failure:

First, the static nature of Word2Vec embeddings assigns a single, context-independent vector to each token. In formal NLP tasks, this may suffice; however, in marketplace slang, the sentiment of a word like "expired" is entirely contingent on its collocations. Without contextualized representations (as provided by transformer-based models like IndoBERT), the embedding layer cannot dynamically adjust token meanings based on surrounding modifiers, leading to systematic misclassification of Type 1 errors.

Second, the LSTM's reliance on sequential structure becomes a liability when confronted with Type 2 errors. The gating mechanisms of LSTM are designed to capture long-term dependencies in grammatically coherent sentences. When the input text is reduced to fragmented keyword strings, common in rapid mobile typing on Shopee, the sequential signal degrades into noise, and the LSTM's hidden states fail to converge on meaningful emotional representations.

Third, the absence of pragmatic reasoning renders the model incapable of decoding Type 3 errors. Sarcasm and irony require not only lexical knowledge but also world-knowledge and discourse-level inference. The LSTM-Word2Vec architecture, operating purely on statistical co-occurrence patterns, has no mechanism to detect the concessive disappointment embedded in phrases like "terimakasih walau kualitas...". This is a well-documented limitation of non-transformer architectures in sentiment analysis literature.

Finally, the vocabulary mismatch inherent in Type 4 errors compounds the problem. Community-specific slang such as "produk receh" or "mantap jiwa" either does not exist in the Word2Vec vocabulary built from the limited 306-review corpus, or occupies semantically neutral vector spaces due to insufficient contextual exposure. This results in the embedding layer outputting zero-vectors or noisy representations, effectively starving the LSTM of discriminative features for these instances.

Collectively, these linguistic vulnerabilities demonstrate that the model's failure is multi-layered: it is simultaneously a problem of data (heuristic proxy labels introducing noise), architecture (static embeddings and sequential inductive biases), and linguistics (the extreme informality of marketplace discourse). This tripartite diagnosis reinforces the core thesis of this study: in low-resource, highly skewed e-commerce domains, architectural sophistication alone cannot compensate for the compounded challenges of noisy labels, limited data, and linguistically complex input. Future optimizations must therefore adopt a holistic approach, integrating contextualized pretrained embeddings (e.g., IndoBERT), rigorous expert annotation to disambiguate slang, and advanced class-balancing techniques to restore the model's semantic sensitivity to the full spectrum of consumer emotions.

3.7. Discussion

The empirical findings of this study reveal a critical vulnerability of the standard LSTM-Word2Vec architecture when deployed on real-world, highly imbalanced, and noisy e-commerce datasets. Contrary to studies that report idealized, high-accuracy results, this research demonstrates that architectural sophistication alone cannot compensate for fundamental data quality and distribution constraints. The catastrophic performance, characterized by a 31% accuracy and a complete "blanket prediction" bias toward the negative class, provides profound insights into the practical limitations of deep sequential models in niche marketplace environments.

The most striking finding is the model’s inability to discriminate between positive and neutral sentiments, resulting in an extreme prediction bias. This phenomenon, often referred to as majority-class collapse or blanket prediction, occurs when the model minimizes its loss function by simply outputting a single dominant class, entirely

disregarding the input features. While Ayu et al. [4] achieved a remarkable 97% accuracy using RNN-LSTM on Shopee reviews, their success was likely underpinned by a massive dataset (10,000 reviews) and a binary classification task, which inherently reduces model complexity. In contrast, our study's attempt at fine-grained, multi-class emotion analysis on a severely restricted dataset (306 samples) exposed the model's fragility. Furthermore, the inherent label noise introduced by heuristic proxy annotations during the initial training phase severely disrupted the gradient descent optimization process, preventing the network from converging on meaningful semantic representations. This aligns with the assertions of Gondhi et al. [6], who emphasized that in highly imbalanced distributions, standard accuracy is a misleading metric, and models are highly susceptible to bias unless rigorous data balancing and authentic ground-truth labeling are enforced. The high numerical error metrics (MSE = 1.66, RMSE = 1.28) further corroborate that the model's predictions deviated fundamentally from the actual emotional semantics.

The qualitative error analysis exposed the model's profound lack of semantic sensitivity to informal Indonesian marketplace slang. The model successfully identified explicit negative tokens (e.g., "expired") but catastrophically failed to recognize positive contextual cues such as "kualitas bagus" (good quality) or "expired lama" (long expiration date), erroneously classifying them as negative. This failure highlights a critical limitation of relying solely on standard Word2Vec embeddings for highly colloquial text. As demonstrated by Hariz et al. [8], deep learning models like LSTM and GRU often struggle with contextual ambiguity and irony in informal reviews, sometimes underperforming compared to traditional machine learning models like SVM that utilize robust TF-IDF n-gram features. To overcome this semantic blindness, future implementations must consider the feature fusion approach proposed by Akbar et al. [5], who proved that combining local contextual embeddings (Word2Vec) with global statistical semantics (GloVe) within a BiLSTM framework significantly enriches the model's ability to capture nuanced sentiment in Indonesian e-commerce reviews.

The stagnant validation loss and severe overfitting observed during the 10-epoch training phase underscore the heavy reliance of LSTM on massive data volumes. LSTM networks are inherently data-hungry architectures; their complex gating mechanisms require extensive examples to learn optimal weight distributions. The failure of the model to generalize is consistent with the findings of Azsyams et al. [9], whose LSTM-based forecasting model yielded a negligible R^2 value (0.09) when applied to a limited dataset of culinary products. They concluded that without sufficient data volume and external variables, LSTM cannot effectively capture temporal or sequential patterns. Additionally, the choice of the Adam optimizer in this study may have contributed to the unstable convergence. Nasution et al. [7] conducted a rigorous comparative study on the PREDECT-ID dataset and revealed that for Indonesian product reviews, the Stochastic Gradient Descent (SGD) optimizer, combined with a dropout rate of 0.5, yields vastly superior stability and generalization compared to Adam, which is prone to rapid overfitting on smaller datasets. The absence of such optimized regularization in our initial setup likely exacerbated the model's inability to learn discriminative emotional features.

Theoretically, this study establishes a crucial empirical baseline for e-commerce emotion mining. It challenges the prevailing assumption that deep learning models universally outperform traditional methods in all textual scenarios. Instead, it proves that in low-resource, highly imbalanced, and noisy domains, the "garbage in, garbage out" principle is amplified. Practically, the findings serve as a vital warning for e-commerce platform developers and sellers: deploying off-the-shelf LSTM models without addressing underlying data distribution and linguistic informality will lead to disastrous business insights, such as misclassifying satisfied customers as dissatisfied.

This study is not without limitations. The restricted dataset size (306 reviews), the reliance on heuristic proxy annotations that introduced inherent label noise, and the use of a single embedding technique (Word2Vec) constrained the model's learning capacity. To address these limitations, future research must adopt a multi-pronged optimization strategy. First, at the data level, advanced resampling techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or data augmentation must be applied to rectify the severe class imbalance, and meticulous expert-annotated ground truth must replace heuristic proxies to eliminate label noise and enhance semantic fidelity. Second, at the feature level, integrating GloVe embeddings or leveraging pre-trained transformer models like IndoBERT could vastly improve the model's semantic understanding of Indonesian slang. Finally, at the architectural level, future studies should explore the SGD optimizer with a 0.5 dropout rate [7], or compare the standard LSTM against more robust sequential architectures like Temporal Convolutional Networks (TCN) [10] or GRU [8], which have demonstrated superior resilience in handling noisy, low-resource Indonesian textual data.

4. CONCLUSION

This study aimed to develop and evaluate an emotion analysis system for e-commerce product reviews using a Long Short-Term Memory (LSTM) architecture integrated with Word2Vec embeddings. The empirical findings reveal that under real-world conditions characterized by severe class imbalance, limited dataset size (306 samples), and highly informal linguistic variations, the standard LSTM model failed to achieve optimal generalization. The model yielded a marginal overall accuracy of 31% and exhibited an extreme classification bias, resulting in a

"majority-class collapse" where all instances were indiscriminately predicted as the negative class. Despite the suboptimal predictive performance, this research provides a critical and realistic baseline for e-commerce emotion mining. It underscores a fundamental tenet in deep learning: architectural sophistication cannot compensate for fundamental data distribution constraints. The inability of the model to distinguish between positive and neutral sentiments highlights the profound vulnerability of sequential models to noisy, imbalanced, and heuristic proxy-labeled datasets. Consequently, this study serves as a vital empirical warning that deploying off-the-shelf LSTM models without addressing underlying data quality and class disparities will lead to erroneous business insights and misinterpretation of consumer satisfaction.

To address these constraints and advance the field, future research must adopt a multi-pronged optimization strategy. At the data level, the implementation of advanced resampling techniques such as SMOTE (Synthetic Minority Over-sampling Technique) and the transition from heuristic proxies to meticulous, expert-annotated ground-truth labels are imperative to rectify class imbalance and eliminate inherent label noise. At the feature and architectural levels, future studies should explore the fusion of global and local embeddings (e.g., combining Word2Vec with GloVe), leverage pre-trained transformer models tailored for Indonesian colloquialisms (e.g., IndoBERT), and experiment with alternative optimization strategies. Specifically, utilizing the SGD optimizer with rigorous dropout regularization, which has demonstrated superior stability in low-resource sequential modeling, is highly recommended to prevent overfitting and enhance the model's semantic sensitivity in dynamic marketplace environments

DECLARATIONS

Author Contributions

M.R. and T. conceived the study and designed the methodology. M.R. performed the data collection and preprocessing. M.R. and T. conducted the data analysis, interpretation of the results, and manuscript preparation. Both authors reviewed, edited, and approved the final version of the manuscript. Both authors have read and agreed to the published version of the manuscript.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Conflicts of Interest

The authors declare that they have no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

AI Use Statement

During the preparation of this work, the authors used Google Gemini for language paraphrasing and scientific figure visualization. After using this tool, the authors reviewed and edited the content and visualizations as needed and take full responsibility for the final content of the published article.

Additional Information

No additional information is available for this paper.

REFERENCES

[1] M. F. Septokasya, “The Influence of Consumer Reviews on Purchasing Decisions on Shopee E-Commerce, Indonesia,” *Int. J. Adm. Bus. Organ.*, vol. 5, no. 3, pp. 118–128, 2024, doi: 10.61242/ijabo.24.362 JEL.

[2] M. Andriyani and R. Ardianto, “Pengaruh Kualitas Layanan dan Kualitas Produk Terhadap Kepuasan Nasabah Bank,” *EKOMABIS J. Ekon. Manaj. Bisnis*, vol. 1, no. 02, pp. 133–140, 2020, doi: 10.37366/ekomabis.v1i02.73.

[3] R. Gunawan, M. B. Dimiliu, K. Valerine, and S. P. Tamba, “ANALISIS PREDIKSI PENJUALAN TOKO FURNITUR DENGAN METODE LONG SHORT-TERM MEMORY (LSTM),” *J. Tekinkom (Teknik Inf. dan Komputer)*, vol. 7, no. 2, pp. 716–725, 2024, doi: 10.37600/tekinkom.v7i2.1511

[4] S. Ayu, B. Irawan, and N. A. Ramdhan, “Sentiment Analysis of Shopee User Reviews Using Recurrent Neural Network with LSTM for Real-Time Web-Based Prediction,” *J. Comput. Networks, Archit. High Perform. Comput.*, vol. 08, no. 1, pp. 123–132, 2026, doi: 10.47709/cnahpc.v8i1.7824

[5] H. Akbar, D. Aryani, M. K. M. Al-shammari, and M. B. Ulum, “SENTIMENT ANALYSIS FOR E-COMMERCE PRODUCT REVIEWS BASED ON FEATURE FUSION AND BIDIRECTIONAL LONG

SHORT-TERM MEMORY," *J. Tek. Inform.*, vol. 5, no. 4, pp. 1385–1391, 2024, doi: 10.52436/1.jutif.2024.5.5.2675.

[6] N. K. Gondhi, E. Sharma, A. H. Alharbi, R. Verma, and M. A. Shah, "Efficient Long Short-Term Memory-Based Sentiment Analysis of E-Commerce Reviews," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/3464524

[7] K. Nasution, K. Saddami, R. Roslidar, and A. Akhyar, "Comparative Study of BiLSTM and GRU for Sentiment Analysis on Indonesian E-Commerce Product Reviews Using Deep Sequential Modeling," *J. Tek. Inform.*, vol. 6, no. 4, pp. 1881–1896, 2025, doi: 10.52436/1.jutif.2025.6.4.4878.

[8] A. Hariz, A. Azrir, P. Naveen, and S. Haw, "Sentiment Analysis using Machine Learning Models on Shopee Reviews," *J. Syst. Manag. Sci.*, vol. 14, no. 2, pp. 214–228, 2024, doi: 10.33168/JSMS.2024.0213.

[9] L. F. Azsyams, A. Pebrianggara, and I. K. Almanfaluti, "Analysis of the LSTM Model on the Demand Patterns of Indonesian Traditional Cookies in Online Marketplaces," *Int. J. Artif. Intell. Digit. Mark.*, vol. 2, no. 10, pp. 35–48, 2025, doi: 10.61796/ijaifd.v2i10.419.

[10] M. Bayu, B. Dharmasaguna, and A. Retnowardhani, "Comparison of LSTM and TCN Models for Customer Churn Prediction Based on Sentiment and Transaction Data," *Int. J. Eng. Sci. Inf. Technol.*, vol. 5, no. 4, pp. 64–74, 2025, doi: 10.52088/ijesty.v5i4.979

[11] D. B. W. Alfredo Gormantara, "KLASIFIKASI KATEGORI DAN PELABELAN BERITA BAHASA INDONESIA MENGGUNAKAN MUTUAL INFORMATION DAN K- NEAREST NEIGHBORS," *J. Temat.*, vol. 8, pp. 75–82, 2020, doi: 10.56963/tematika.vi.257

[12] O. Stephen, M. Sain, U. J. Maduh, and D. U. Jeong, "An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare," *J. Healthc. Eng.*, vol. 2019, 2019, doi: 10.1155/2019/4180949.

[13] G. N. A. H. Yar, M. Taha, Z. Afzal, F. Zafar, I. U. R. Shahid, and A. Noor-Ul-Hassan, "TexGAN: Textile Pattern Generation Using Deep Convolutional Generative Adversarial Network (DCGAN)," *Proc. - 2023 IEEE Int. Conf. Emerg. Trends Eng. Sci. Technol. ICES T 2023*, no. June, 2023, doi: 10.1109/ICEST56843.2023.10138848.

[14] Guruh Wijaya, Dudi Irawan, Zainul Arifin, Hardian Oktavianto, Miftahur Rahman, and Ginanjar Abdurrahman, "Studi Klasifikasi Topik Berita Dengan Algoritma Machine Learning," *J-Ensitec*, vol. 11, no. 01, pp. 10202–10206, 2024, doi: 10.31949/jensitec.v11i01.12037.

[15] P. Sayarizki and H. Nurrahmi, "Implementation of IndoBERT for Sentiment Analysis of Indonesian Presidential Candidates," *J. Comput.*, vol. 9, no. 2, pp. 61–72, 2024, doi: 10.34818/indojc.2024.9.2.934.

[16] A. Muhaddisi, "Sentiment Analysis With Sarcasm Detection on Politician's Instagram," *Ijccs (Indonesian J. Comput. Cybern. Syst.*, 2021, doi: 10.22146/ijccs.66375.

[17] N. Tabassum, A. Namoun, T. Alyas, A. Tufail, M. Taqi, and K. H. Kim, "Classification of Bugs in Cloud Computing Applications Using Machine Learning Techniques," *Appl. Sci.*, vol. 13, no. 5, 2023, doi: 10.3390/app13052880.

[18] F. D. K. Romario Onsu, Daniel Febrian Sengkey, "IMPLEMENTASI BI-LSTM DENGAN EKSTRAKSI FITUR WORD2VEC UNTUK PENGEMBANGAN ANALISIS SENTIMEN APLIKASI IDENTITAS KEPENDUDUKAN DIGITAL," vol. 10, no. 1, pp. 46–55, 2024, doi: 10.54914/jtt.v10i1.1225

[19] J. Yu, S. Park, S. H. Kwon, C. M. B. Ho, C. S. Pyo, and H. Lee, "AI-based stroke disease prediction system using real-time electromyography signals," *Appl. Sci.*, vol. 10, no. 19, 2020, doi: 10.3390/app10196791.

[20] A. M. Hasan, M. M. AL-Jawad, H. A. Jalab, H. Shaiba, R. W. Ibrahim, and A. R. AL-Shamasneh, "Classification of Covid-19 Coronavirus, Pneumonia and Healthy Lungs in CT Scans Using Q-Deformed Entropy and Deep Learning Features," *Entropy*, vol. 22, 2020, doi: 10.3390/e22050517

[21] B. Imran, H. Hambali, A. Subki, Z. Zaeniah, A. Yani, and M. R. Alfian, "Data Mining Using Random Forest, Naïve Bayes, and Adaboost Models for Prediction and Classification of Benign and Malignant Breast Cancer," *J. Pilar Nusa Mandiri*, vol. 18, no. 1, pp. 37–46, 2022, doi: 10.33480/pilar.v18i1.2912.