Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16165
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Comparative Sentiment Analysis of GrabFood Reviews Using BiLSTM and BiGRU

Jakasurya Siswoyo[1*)], Andreas Leonardo Sumendap[2)], Lorna Yertas Baisa[3)]
[1,2,3)] Universitas Papua
jakakuliah16@gmail.com, al.sumendap@unipa.ac.id, l.baisa@unipa.ac.id

Submitted :May 1, 2026 | **Accepted** : June 2, 2026 | **Published** : July 5, 2026

Abstract: The exponential growth of user-generated reviews on digital platforms has made manual sentiment interpretation of Online Food Delivery (OFD) services increasingly impractical. GrabFood, operating within the Grab ecosystem, has accumulated over 16.1 million reviews on the Google Play Store, necessitating an automated and scalable approach to sentiment monitoring. Conventional labeling approaches, including star-rating proxies and lexicon-based annotation, are inadequate for capturing contextual nuance, negation, and informal linguistic patterns prevalent in Indonesian-language OFD reviews. Furthermore, limited research has systematically compared BiLSTM and BiGRU architectures within a transformer-assisted labeling framework for Indonesian OFD sentiment analysis. This study aims to implement RoBERTa-based automatic sentiment labeling and to comparatively evaluate BiLSTM and BiGRU models for three-class sentiment classification of GrabFood reviews. A corpus of 265,500 raw reviews was collected via web scraping, filtered to 17,709 reviews through rigorous preprocessing, and annotated using the *w11wo/indonesian-roberta-base-sentiment-classifier*. Random Oversampling was applied to address class imbalance. BiLSTM and BiGRU models were trained and benchmarked against Support Vector Machine (SVM) and Naïve Bayes baselines. BiLSTM achieved 86% accuracy while BiGRU attained 85%, both substantially outperforming SVM (82%) and Naïve Bayes (77%). However, BiGRU demonstrated superior convergence speed and more stable per-class performance, particularly on the neutral category (F1: 51% vs. 50%). Transformer-assisted automatic labeling combined with bidirectional recurrent architectures constitutes an effective and scalable pipeline for Indonesian OFD sentiment classification, with neutral sentiment remaining the primary classification challenge.

Keywords: Automatic Labeling, BiGRU, BiLSTM, GrabFood, RoBERTa, Sentiment Analysis

INTRODUCTION

The rapid proliferation of digital technology has fundamentally restructured consumer markets, accelerating the transition from conventional commerce toward fully digitalized service ecosystems. Among the most prominent manifestations of this shift is the emergence of Online Food Delivery (OFD) services, exemplified by GrabFood under the Grab platform and GoFood under the Gojek ecosystem (Rangaswamy, Nawaz, & Changzhuang, 2022; Safira & Chikaraishi, 2022). Several factors influence the choice of OFD services, including individual factors, technological factors, and external factors (Sabilaturrizqi & Subriadi, 2024). GrabFood is a service feature within the Grab app that facilitates food ordering and delivery. Using the GrabFood function, users can place orders at restaurants affiliated with PT Grab (Hasanah & Hargyatni, 2022). As of February 24, 2026, statistics from the Google Play Store show that the Grab app has surpassed 100 million users, has a rating of 4.8, and has 16.1 million reviews. Reviews are customer comments about a product or service posted on a third party. Existing reviews show remarkable growth. The large number of reviews makes it difficult to understand user attitudes and satisfaction. Therefore, sentiment analysis is used to automatically understand reviewer opinions and sentiments (Sabilaturrizqi & Subriadi, 2024).

The goal of sentiment analysis, a subfield of Natural Language Processing (NLP) is to extract subjective feelings and views about a product or service from written content (Bhadane, Dalal, & Doshi, 2015; Pang & Lee, 2008). The deep learning technique is used to do sentiment analysis. This strategy involves training artificial neural

e-ISSN : 2541-2019
p-ISSN : 2541-044X

networks to analyze data and make use of error values. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are two deep learning approaches to be used (Chung, Gulcehre, Cho, & Bengio, 2014; Hochreiter & Schmidhuber, 1997).

Nevertheless, one of the most persistent challenges in sentiment analysis research pertains to the reliable assignment of sentiment labels particularly when distinguishing among positive, negative, and neutral categories in domain-specific corpora. When working with massive datasets, manual labeling becomes a tedious and inefficient process. Transformer-based automated labeling offers a scalable and context-aware alternative to manual annotation. Among such approaches, the Robustly Optimized BERT Pretraining Approach (RoBERTa) has demonstrated superior labeling fidelity owing to its enhanced pre-training strategy and dynamic masking mechanism (Robustly Optimized BERT Pre-training Approach). RoBERTa, for instance, is another BERT model that produces better results with the modified pretraining procedure (Liu et al., 2019).

A critical examination of extant literature reveals several substantive gaps. Prior studies on GrabFood and analogous OFD platforms predominantly rely on Twitter-derived datasets, star-rating-based annotation, or lexicon-driven labeling tools approaches that are inherently limited in their capacity to encode contextual nuance, sarcasm, or culturally specific expressions (Arifin & Purnama, 2023; Muhammad, Kusumaningrum, & Wibowo, 2021). Furthermore, the majority of prior works treat the Grab application as a monolithic entity, failing to differentiate sentiment directed specifically at the GrabFood service component. The application of transformer-based automatic labeling pipelines to large-scale Indonesian-language OFD review corpora has, to the best of the authors' knowledge, not been previously investigated. This study proposes an improved hybrid sentiment analysis pipeline for Indonesian OFD reviews that can overcome these limitations. To address the aforementioned gaps, this study proposes a two-phase sentiment analysis pipeline for Indonesian GrabFood reviews. The first phase employs RoBERTa-based automatic labeling utilizing the w11wo/indonesian-roberta-base-sentiment-classifier to generate high-fidelity sentiment annotations that surpass conventional rating-based or lexicon-based methods in contextual accuracy. The second phase conducts a systematic comparative evaluation of BiLSTM and BiGRU architectures for three-class sentiment classification, supplemented by Support Vector Machine (SVM) and Naïve Bayes (NB) as conventional baselines. Based on the preceding discussion, this study seeks to address the following research questions: (1) How effective is RoBERTa-based automatic labeling in generating reliable sentiment annotations for large-scale Indonesian OFD review data? (2) Which architecture BiLSTM or BiGRU yields superior performance in three-class sentiment classification of GrabFood reviews? (3) To what extent do deep learning models outperform conventional machine learning baselines in this domain?

Accordingly, the objectives of this study are: (1) to implement and validate a transformer-based automatic labeling approach using RoBERTa for Indonesian-language GrabFood reviews; (2) to comparatively evaluate the classification performance of BiLSTM and BiGRU models under equivalent experimental conditions; and (3) to benchmark deep learning architectures against SVM and Naïve Bayes baselines to quantify the performance advantage of sequential deep learning in this context. The principal contributions of this study are threefold: (1) the development of a scalable, context-aware automatic labeling pipeline for Indonesian OFD reviews using RoBERTa, mitigating the subjectivity and scalability constraints inherent in manual annotation; (2) a rigorous empirical comparison of BiLSTM and BiGRU architectures in a three-class Indonesian sentiment classification task an underexplored configuration in prior literature; and (3) an empirical benchmark establishing the performance advantage of deep learning approaches over conventional classifiers within the GrabFood review domain.

LITERATURE REVIEW

Comparative Analysis of Bidirectional Deep Learning Architectures

Recurrent neural networks have undergone considerable architectural refinement in response to the limitations of early sequential models. Conventional RNNs were fundamentally constrained by the vanishing gradient problem, which degraded their capacity to retain information across extended text sequences. Long Short-Term Memory addressed this by introducing a structured gating mechanism, input, forget, and output gates, enabling selective memory retention over longer contextual spans (Hochreiter & Schmidhuber, 1997). A subsequent advancement, Bidirectional LSTM, extended this capability by processing input sequences simultaneously in forward and backward directions, substantially enriching the contextual embeddings available to downstream classification layers (Abduljabbar, Dia, & Tsai, 2021). Nevertheless, BiLSTM's architectural richness comes at a cost: its greater number of trainable parameters increases susceptibility to overfitting when applied to domain-specific datasets of moderate size (Nasution et al., 2025).

As a more computationally efficient alternative, the Bidirectional Gated Recurrent Unit compresses the gating mechanism into two components, a reset gate and an update gate, reducing model complexity without severely compromising representational capacity (Chung et al., 2014; Hochreiter & Schmidhuber, 1997). This architectural economy makes BiGRU particularly well-suited for deployment in resource-constrained environments. However, its reduced parametric expressiveness may limit performance on semantically ambiguous

e-ISSN : 2541-2019
p-ISSN : 2541-044X

categories such as neutral sentiment, where fine-grained contextual discrimination is most critical (Redjeki et al., 2025). The trade-off between these two architectures, expressive power versus computational efficiency, remains an empirically unresolved question in the context of Indonesian-language online food delivery (OFD) review classification.

Transformer-Based Labeling versus Conventional Annotation

The quality of training labels is a foundational determinant of supervised learning performance. Lexicon-based annotation approaches are operationally simple yet structurally constrained: they rely on fixed word inventories that cannot accommodate contextual polarity reversals, negation constructions, or culturally embedded expressions prevalent in informal Indonesian text (Mutinda, Mwangi, & Okeyo, 2023). Star-rating-based annotation introduces a different class of error by assuming a direct mapping between numerical scores and underlying sentiment polarity, an assumption that breaks down in the presence of mixed-emotion reviews or culturally specific communication norms. RoBERTa, developed through dynamic masking, extended training batches, and elimination of the Next Sentence Prediction objective, offers a principled alternative that substantially outperforms earlier BERT-based approaches across diverse NLP benchmarks (Liu et al., 2019). Domain-adapted variants, including the w11wo/indonesian-roberta-base-sentiment-classifier, extend this capability to Indonesian informal registers characterized by code-switching and non-standard orthography.

Identified Research Gaps

Prior Indonesian sentiment analysis studies employing LSTM with Word2Vec have achieved binary classification accuracy in the 85–93% range (Arifin & Purnama, 2023; Muhammad et al., 2021), but three-class performance degrades notably, dropping to 74–82% on Twitter corpora (Ahmed, Yuosif, Ahmed, & Mohammed, 2025) as low as 47% on hospitality review datasets (Husein, Livando, Andika, Chandra, & Phan, 2023) GrabFood-specific studies have relied predominantly on TextBlob or star-rating proxies (Arifin & Purnama, 2023; Muhammad et al., 2021), which inadequately capture the imbalanced distributions and linguistic heterogeneity of Indonesian OFD corpora (Singgalen, 2024). Although BiLSTM has demonstrated 83.4% accuracy under manual annotation conditions (Redjeki et al., 2025), scalable labeling pipelines for large-scale Google Play Store datasets remain absent from existing literature. Three gaps emerge: the absence of scalable transformer-based labeling for Indonesian OFD corpora, the lack of a controlled BiLSTM–BiGRU comparative evaluation in this domain, and the underrepresentation of app-store review data relative to social media sources. The present study is designed to address each of these gaps within a unified experimental framework (Nasution et al., 2025).

METHODS

The methodology adopted in this study is structured into five sequential phases: data collection, data preprocessing, data preparation, model development, and model evaluation. The overall workflow is illustrated in Figure 1. A distinguishing characteristic of the proposed pipeline lies in its integration of transformer-based automatic labeling specifically RoBERTa as an upstream annotation instrument, followed by comparative classification using BiLSTM and BiGRU architectures. This two-phase design addresses the dual limitations of conventional labeling scalability and recurrent model comparability within the Indonesian OFD sentiment analysis domain.

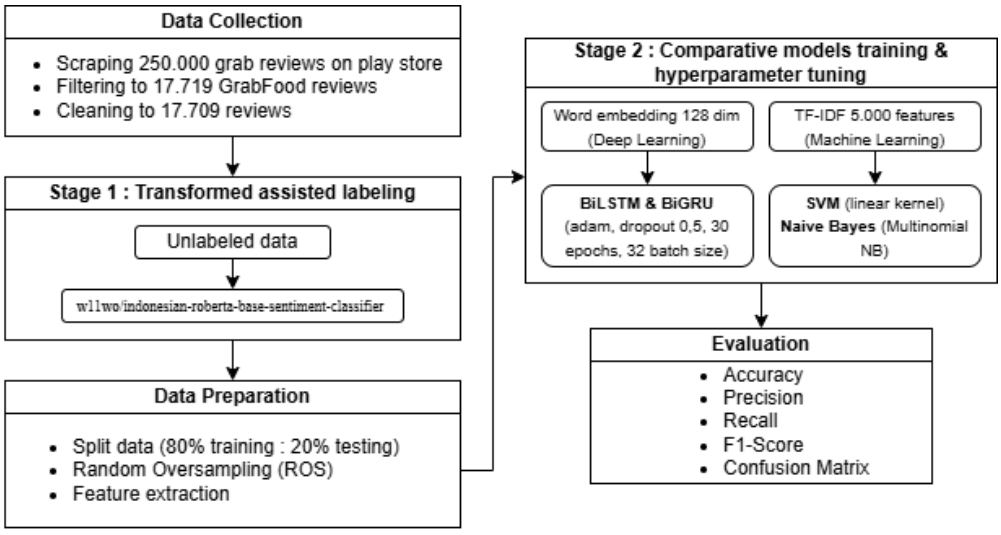
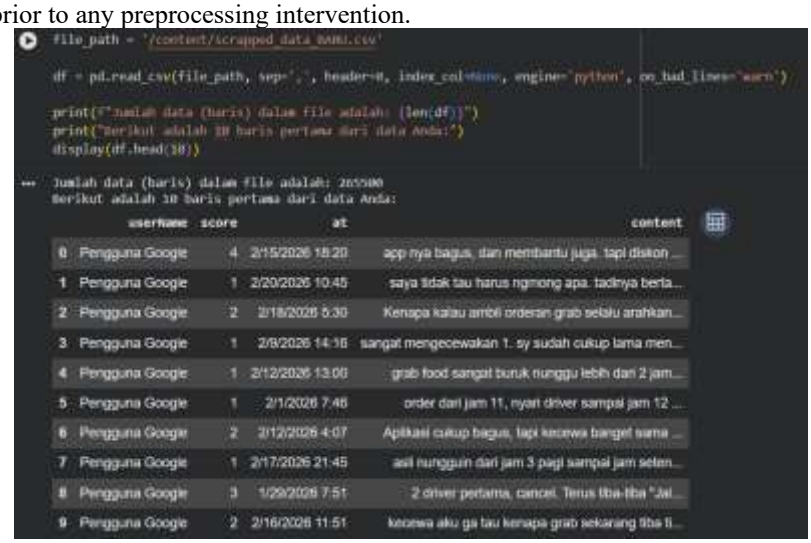


Figure 1 Research processes

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16165
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Data collection

Reviews were collected using the web scraping method carried out on the Google Colaboratory (Google Colab) platform with the assistance of the Python language using the Google-play-scraper library. A total of 265,500 user reviews were obtained within the time range of December 9, 2018 to February 24, 2026. To enhance transparency and reproducibility of the dataset, Table 1 presents a representative sample of raw reviews obtained from the scraping process prior to any preprocessing intervention.



```
file_path = '/content/scrapped_data_mami.csv'

df = pd.read_csv(file_path, sep=',', header=0, index_col=None, engine='python', on_bad_lines='warn')

print(f"Jumlah data (baris) dalam file adalah: {len(df)}")
print("Berikut adalah 10 baris pertama dari data Anda:")
display(df.head(10))
```

```
Jumlah data (baris) dalam file adalah: 265500
Berikut adalah 10 baris pertama dari data Anda:
```

	userName	score	at	content
0	Pengguna Google	4	2/15/2026 16:20	app nya bagus, dan membantu juga. tapi diskon ...
1	Pengguna Google	1	2/20/2026 10:45	saya tidak tau harus ngmong apa. tadinya berta...
2	Pengguna Google	2	2/18/2026 5:30	Kenapa kalau ambil orderan grab selalu arahkan...
3	Pengguna Google	1	2/8/2026 14:16	sangat mengecewakan 1. sy sudah cukup lama men...
4	Pengguna Google	1	2/12/2026 13:00	grab food sangat buruk nunggu lebih dari 2 jam...
5	Pengguna Google	1	2/1/2026 7:46	order dari jam 11, nyari driver sampai jam 12 ...
6	Pengguna Google	2	2/12/2026 4:07	Aplikasi cukup bagus, tapi kecewa banget sama ...
7	Pengguna Google	1	2/17/2026 21:45	asli nunggunin dari jam 3 pagi sampai jam seten...
8	Pengguna Google	3	1/29/2026 7:51	2 driver pertama, cancel. Terus tiba-tiba "Jal...
9	Pengguna Google	2	2/16/2026 11:51	kecewa aku ga tau kenapa grab sekarang tiba ti...

Figure 2 Sample of Raw GrabFood Review Data from Scraping

Data preprocessing

Raw reviews were preprocessed through three sequential stages within Google Colab using Python: keyword-based filtering, data cleaning, and case folding. Keyword filtering isolated GrabFood-specific reviews using domain-relevant terms, yielding 17,719 reviews, subsequently refined to 17,709 after duplicate removal, punctuation stripping, emoji elimination, and hyperlink deletion. Case folding standardized all text to lowercase. Notably, stopword removal and stemming were deliberately omitted, as RoBERTa relies on complete contextual representations where functional words carry sentiment-bearing capacity, and stemming risks distorting Indonesian informal morphological structures (Khairani, Mutiawani, & Ahmadian, 2024). Removing such linguistic features would compromise the semantic fidelity available to the transformer-based labeling model.

Table 1 Example of Preprocessing Transformation (Before vs. After)

Before	After
app nya bagus, dan membantu juga. tapi diskon gede itu kalo akun baru. biasanya kalo udah rada lamaan diskon nya ga sebanyak waktu masih baru", kecuali resto prioritas, biasanya itu yang diskon gede. cuma kalo resto biasa kecil promo nya. soalnya saya beli di resto biasa cuma dapat diskon ongkir 3rb saja, dan ongkir nya dari 2k jdi 5k, padahal waktu awal awal pake ongkir hanya 2k dan dapat diskon 30%. tapi yang penting kalo pesen makanan gak delayed.	app nya bagus dan membantu juga tapi diskon gede itu kalo akun baru biasanya kalo udah rada lamaan diskon nya ga sebanyak waktu masih baru kecuali resto prioritas biasanya itu yang diskon gede cuma kalo resto biasa kecil promo nya soalnya saya beli di resto biasa cuma dapat diskon ongkir rb saja dan ongkir nya dari k jdi k padahal waktu awal awal pake ongkir hanya k dan dapat diskon tapi yang penting kalo pesen makanan gak delayed
Aplikasi cukup bagus, tapi kecewa banget sama grab,promo grabfood cuma diawal awal aja, semakin sering pesen lewat grab palah semakin dikit promonya dan jadi lebih mahal banget, mau ambil yang berlangganan juga jadi mikir dua kali karna grab hanya memprioritaskan untuk memancing pelanggan bukan untuk buat pelanggan nyaman dan terus menerus pakai grab, saran untuk seterusnya perhatikan juga untuk loyalitas pelanggan lama bukan cuma pengguna baru aja.	aplikasi cukup bagus tapi kecewa banget sama grabpromo grabfood cuma diawal awal aja semakin sering pesen lewat grab palah semakin dikit promonya dan jadi lebih mahal banget mau ambil yang berlangganan juga jadi mikir dua kali karna grab hanya memprioritaskan untuk memancing pelanggan bukan untuk buat pelanggan nyaman dan terus menerus pakai grab saran untuk seterusnya perhatikan juga untuk loyalitas pelanggan lama bukan cuma pengguna baru aja

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16165
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Terimakasih sudah menyediakan pembayaran lewat Qris. Sangat berguna apalagi untuk orang-orang yang jarang pakai cash. Fitur pembayaran yang sangat berguna ðŸ˜‡ Pindah dari G*Jek ke Grab karena diskon dan promo nya asli, pembayaran gampang dan sejauh ini pesan makanan, berjalan dengan baik. Terimakasih Grab ðŸ˜‡	terimakasih sudah menyediakan pembayaran lewat qris sangat berguna apalagi untuk orangorang yang jarang pakai cash fitur pembayaran yang sangat berguna pindah dari gjek ke grab karena diskon dan promo nya asli pembayaran gampang dan sejauh ini pesan makanan berjalan dengan baik terimakasih grab

Data preparation

Preprocessed data underwent four sequential preparation stages: automatic labeling, data splitting, class balancing, and word embedding. All 17,709 reviews were annotated into three sentiment categories positive, negative, and neutral using the w11wo/indonesian-roberta-base-sentiment-classifier, an Indonesian-adapted RoBERTa variant that extends BERT through dynamic masking, larger batch training, and removal of the Next Sentence Prediction objective, collectively enhancing fine-grained contextual inference (Liu et al., 2019). Uniquely, RoBERTa was deployed as an upstream labeling instrument rather than a terminal classifier, enabling scalable high-fidelity annotation. Labeling reliability was validated through manual annotation of 200 stratified reviews by two independent evaluators, with inter-rater agreement assessed via Cohen's Kappa and RoBERTa–human concordance subsequently computed to confirm pipeline adequacy.

Table 2 Distribution of Grabfood review sentiment

Sentiment	Number of reviews	Percentage
Positive	5181	29.26%
Negative	11309	63.86%
Neutral	1219	6.88%

Split Data: There are two sections to the GrabFood review dataset: training and testing. We trained the BiLSTM and BiGRU models using 80% of the data and tested their performance with 20%. We used 14,167 reviews for training and 3,542 reviews for testing out of a total of 17,709 review data sets. An analysis of the training and testing data sets reveals the distribution of each emotion in Figure 3.

Figure 3 Sentiment distribution on testing and training data

The labeled dataset exhibited severe class imbalance, with negative reviews dominating at 63.86% and neutral reviews comprising only 6.88%, a disparity that risks systematically biasing model predictions toward majority classes and suppressing minority-class recall (Wibowo & Fatichah, 2021). Random Oversampling (ROS) was selected as the balancing strategy, stochastically replicating minority instances to restore distributional equilibrium without generating synthetic data points (Ependi, Rochim, & Wibowo, 2023). Alternative techniques were deliberately excluded: SMOTE is unsuitable for high-dimensional discrete token sequences, focal loss introduces architectural complexity and hyperparameter sensitivity, and cost-sensitive learning requires objectively undefined misclassification cost matrices. ROS, being computationally lightweight and architecture-agnostic, proved most operationally compatible with the sequential embedding pipeline employed in this study (Ependi et al., 2023). The class distribution following ROS application is illustrated in Figure 6.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16165
e-ISSN : 2541-2019
p-ISSN : 2541-044X

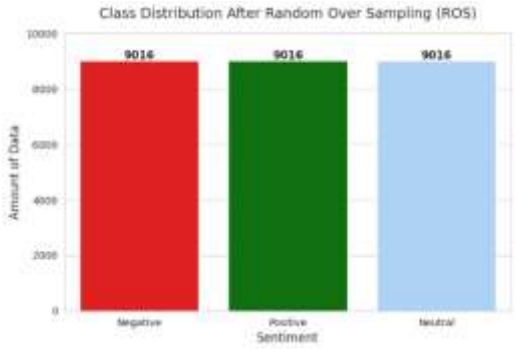


Figure 4 Class distribution after ROS

Tokenization was performed using the TensorFlow Keras tokenizer, generating 20,781 unique tokens, of which the 5,000 most frequent were retained; out-of-vocabulary tokens were mapped to an <OOV> placeholder. Sequences were padded to a uniform length of 100 tokens. Each token was subsequently mapped to a 128-dimensional trainable dense vector, with weights updated adaptively during training, enabling semantically proximate terms to converge within the embedding space and capturing domain-specific linguistic patterns characteristic of informal GrabFood reviews (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013).

BiLSTM and BiGRU Modeling

Both BiLSTM and BiGRU models were constructed following an identical sequential architecture to ensure comparability. Each model begins with an Embedding Layer (input dimension: 5,000; output dimension: 128), followed by a Bidirectional Layer comprising 64 units, a Dropout Layer (rate: 0.5) to mitigate overfitting, and a final Dense Layer with Softmax activation for three-class probability output. Baseline training parameters are specified in Table 3, and hyperparameter tuning was subsequently applied to identify optimal configurations.

Table 3 BiLSTM and BiGRU parameter baselines

Parameter	Value
Units	64
Embedding Dimension	128
Dropout Rate	0.5
Learning rate	0.001
Optimizer	Adam
Output Activation	Softmax
Batch Size	32
Epoch	30
Early Stopping	5 Patience

Baseline Comparison Models

To contextualize the performance of the proposed deep learning architectures, two conventional machine learning classifiers were incorporated as empirical baselines: Support Vector Machine (SVM) with a linear kernel and Multinomial Naïve Bayes (NB). Both models were trained on TF-IDF feature representations with a maximum vocabulary of 5,000 features, consistent with the vocabulary scope used in deep learning models, and evaluated under the same 80:20 train-test partitioning protocol (Arifin & Purnama, 2023; Muhammad et al., 2021). Configuration details are summarized in Table 4.

Table 4 Configuration of Baseline Comparison Models

Model	Feature Extraction	Configuration
Support Vector Machine (SVM)	TF-IDF (max 5,000 features)	Kernel: Linear, C: 1.0
Naïve Bayes (NB)	TF-IDF (max 5,000 features)	Variant: Multinomial NB, α: 1.0
BiLSTM	Word Embedding (128-dim)	See Table 3
BiGRU	Word Embedding (128-dim)	See Table 3

Evaluation

The effectiveness of the BiGRU and BiLSTM models was tested. A classification report and a confusion matrix were used to evaluate the models in the study. Accuracy, recall, precision, and f1-score were the main metrics used in the classification report.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16165
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Accuracy is used to measure the number of correct predictions for the entire data set.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision is used to measure the accuracy of a model in predicting a particular class.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall is a metric that indicates the model's ability to determine all data that truly belong to a particular class.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

The F1 score is a combination of precision and recall and is used to measure model balance.

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

TP (True Positive) = a situation in which the model accurately predicts data as positive, neutral, or negative and the prediction outcomes correspond to the real label.
TN (True Negative) = a situation in which the model accurately predicts that data does not belong to a particular negative, neutral, or positive class and the prediction results agree with the actual label.
FP (False Positive) = a situation in which the data does not truly fall into the negative, neutral, or positive class that the model predicts.
FN (False Negative) = a situation in which, despite the fact that the data truly falls into that class, the model is unable to predict it as negative, neutral, or positive

RESULT

This section presents research findings from sentiment analysis of GrabFood reviews using the BiLSTM and BiGRU methods. The results include data visualization, baseline modeling results, parameter tuning results, and model performance evaluation.

Data visualization

In this data visualization stage, the data is presented in the form of a bigram word cloud. A word cloud is a data visualization technique used to display the occurrence of words with the highest frequency. The word clouds for each sentiment are displayed in Figure 5 to Figure 7.

Figure 5 Positive sentiment word cloud

Figure 6 Negative sentiment word cloud

Figure 7 Neutral sentiment word cloud

Bigram word cloud analysis revealed distinct lexical patterns across all three sentiment categories. Positive sentiment was characterized by expressions of gratitude and satisfaction, including phrases such as *"sangat membantu," "terima kasih," "banyak promo,"* and *"saya suka,"* reflecting users who perceived GrabFood as genuinely beneficial. Negative sentiment was dominated by frustration-laden bigrams including *"tidak bisa," "double order," "lama banget,"* and *"di cancel,"* collectively indicating recurring service failures encompassing order cancellation failures, duplicate transactions, and delivery delays. Neutral sentiment exhibited informational

rather than evaluative language, with bigrams such as *"tolong di," "untuk grabfood,"* and *"di update"* suggesting constructive feedback or feature requests devoid of explicit emotional polarity. Notably, several bigrams bridged sentiment boundaries *"tidak bisa"* and *"terima kasih"* appeared within neutral reviews, underscoring the inherent lexical ambiguity that makes neutral classification particularly challenging. Regarding class distribution, negative sentiment constituted the overwhelming majority at 63.86% (11,309 reviews), followed by positive at 29.26% (5,181 reviews), with neutral representing a marginal 6.88% (1,219 reviews). This pronounced imbalance introduces systematic bias toward majority-class predictions, artificially inflating overall accuracy while suppressing per-class precision, recall, and F1-score for underrepresented categories. Although Random Oversampling was applied to partially redress this disparity, the intrinsic semantic vagueness of neutral sentiment where conflicting polarity signals coexist within single reviews constitutes a structural classification challenge that resampling techniques alone cannot resolve.

Conventional Machine Learning Baseline Results

Conventional machine learning baselines were trained and evaluated before neural network models, end providing an empirical reference prior to deep learning model evaluation. SVM with linear kernel and Multinomial NB both were used on TF-IDF feature extraction method (with maximum 5,000 features) to the same preprocessed and labeled dataset with a train-test split of 80:20. The classification results are shown in
Table *5*.

Table 5 Classification Report of Conventional Machine Learning Baselines

Model	Sentiment Class	Precision	Recall	F1-score	Accuracy
SVM	Negative	0.87	0.91	0.89	0.82
	Neutral	0.52	0.28	0.36	
	Positive	0.74	0.74	0.74	
Naïve Bayes	Negative	0.81	0.93	0.87	0.77
	Neutral	0.37	0.19	0.25	
	Positive	0.69	0.55	0.61	

Like the results of the baseline models, both conventional models did well on the negative class, which is naturally shows majority on the dataset. Despite this, both models performance dropped heavily on the neutral class with SVM getting an F1-score at 0.36 and Naïve Bayes scoring only 0.25 for this particular class. This performance difference for the neutral class is over four times higher than that in the deep learning models, indicating that a bag-of-words representation like TF-IDF cannot capture the contextual subtleties needed to classify sentiment into exactly positive/negative and neutral expressions. Results for SVM (82%) and Naïve Bayes (77%), on the other hand, are lower than both BiLSTM (86%) and BiGRU (85%), proving that recurrent architectures with word embedding representations are more suitable models for this task. This is because BiLSTM and BiGRU can learn contextual characteristics of words bearing sentiment from both right and left sides, while static TF-IDF representations cannot encode properties leveraging sequential dependencies (Arifin & Purnama, 2023; Muhammad et al., 2021).

BiLSTM and BiGRU Baseline Modeling Results

The baseline modeling process was conducted to determine the initial performance of the model before the hyperparameter tuning process. The modeling results from the BiLSTM and BiGRU baseline models can be seen in Figure 8 and Figure 9. Furthermore, the evaluation results are presented in the form of a classification report in Table 6.

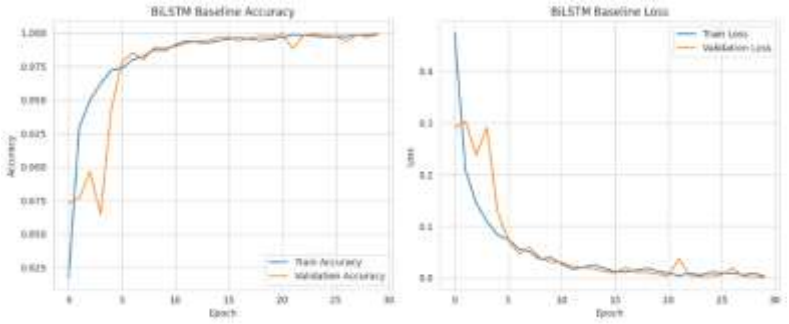


Figure 8 Learning curves of the baseline BiLSTM model

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16165
e-ISSN : 2541-2019
p-ISSN : 2541-044X

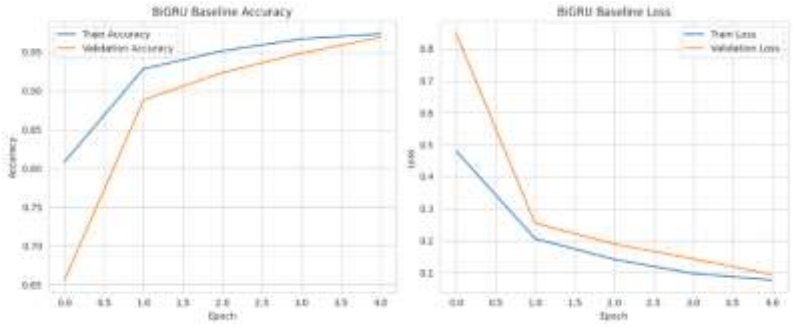


Figure 9 Learning curves of the baseline BiGRU model

Table 6 Classification report from BiLSTM and BiGRU baselines

Model	Sentiment Class	Precision	Recall	F1-score	Accuracy
BiLSTM	Negative	0.90	0.88	0.89	0.83
	Neutral	0.45	0.38	0.41	
	Positive	0.75	0.82	0.78	
BiGRU	Negative	0.84	0.95	0.89	0.83
	Neutral	0.46	0.60	0.52	
	Positive	0.95	0.60	0.73	

The BiGRU and BiLSTM models both showed significant performance gains during training, as seen in Figure 8 and Figure 9, Improved accuracy and lower loss values in validation and training data show this. Also, the BiGRU model was terminated after 6 epochs of Early Stopping since the validation loss value had not increased over the previous several epochs. According to Table 6 the BiLSTM and BiGRU models did well in the positive and negative categories, but they were not as strong in the neutral category. There has to be an attempt to enhance the performance of the models since they both reached 83% accuracy. We ran a hyperparameter tuning procedure to get the optimal values for the parameters that would make the model work better.

Hyperparameter Tuning

The tuning process is performed using the grid search method. This method involves searching for predetermined parameter combinations to obtain the most optimal model performance based on model evaluation values. The parameters used in the model training process are shown in Table 7.

Table 7 Hyperparameter Combination Used

Parameter	Value
Embedding	128
Hidden Units (LSTM/GRU)	64, 128
Dropout Rate	0.4, 0.5, 0.6
Learning Rate (Adam)	0.001, 0.0005, 0.0003
Batch Size	16, 32

Through the hyperparameter tuning process, the results of the best parameters are obtained, which are shown in Figure 10

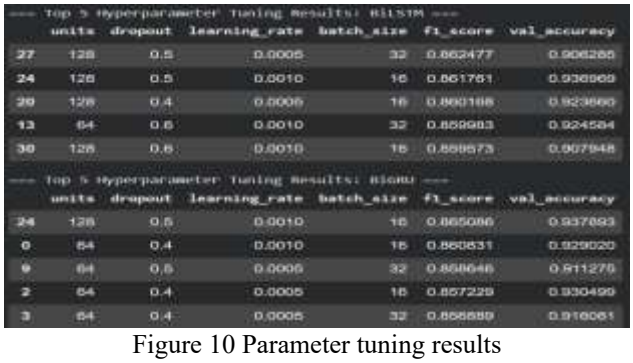


Figure 10 Parameter tuning results

The tuning results revealed that several parameter combinations produced different performance results for the two models. These differences indicate that parameter selection impacts model performance. Based on these

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16165
e-ISSN : 2541-2019
p-ISSN : 2541-044X

results, three optimal parameter combinations were selected for model evaluation. These three parameter combinations were chosen to ensure the model can provide a thorough evaluation and achieve optimal performance.

Evaluation

Classification report and confusion matrix provide the assessment findings, which include the f1-score values for each emotion as well as accuracy, precision, and recall. The models were trained for up to 30 epochs with early stopping using a patience of 10 epochs. Given the class imbalance described in the data visualization section, particular attention is directed toward per-class F1-score values, especially for the neutral class, as overall accuracy alone is insufficient to capture the true classification capability of each model across all sentiment categories. Figure 11 displays the results of the assessment for each model's top three parameters as determined by the confusion matrix.

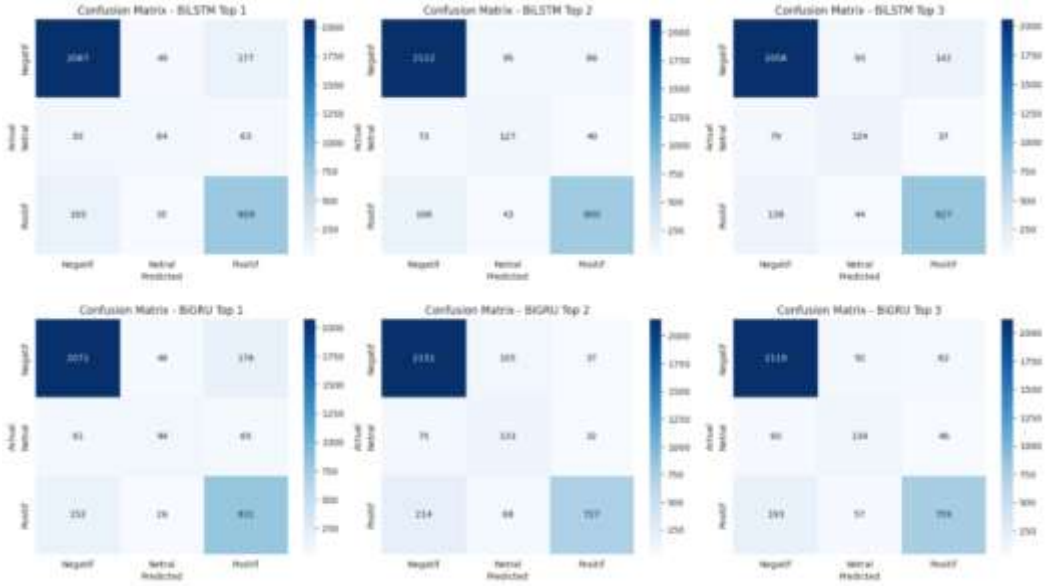


Figure 11 Confusion matrix of top 3 parameters for each model

The confusion matrices in Figure 11, show a distinct pattern of misclassification for the neutral class. A substantial fraction of neutral reviews was mistakenly predicted as negative. It happens because neutral reviews in the GrabFood dataset usually has complaint like word (e.g. mohon diperbaiki, tolong update) that visually recognized as negative class by model since mostly are from negative dataset. This indicates that semantic overlap between neutral and negative classes is still a significant problem, even with ROS. Classification report to evaluate model performance used it to visualize precision, recall and f1-score values for each sentiment. The classification report appears tabular as seen in Table 8.

Table 8 Classification report of top 3 parameters for each model

Models & Parameters	Sentiment Class	Precision	Recall	F1-score	Accuracy
BiLSTM Top 1	Negative	0.89	0.90	0.90	0.84
	Neutral	0.50	0.35	0.41	
	Positive	0.77	0.80	0.79	
BiLSTM Top 2	Negative	0.90	0.92	0.91	**0.86**
	Neutral	0.48	0.53	0.50	
	Positive	0.86	0.79	0.83	
BiLSTM Top 3	Negative	0.90	0.90	0.90	0.85
	Neutral	0.48	0.52	0.50	
	Positive	0.82	0.82	0.82	
BiGRU Top 1	Negative	0.90	0.90	0.90	0.85
	Neutral	0.57	0.39	0.46	
	Positive	0.78	0.82	0.80	
BiGRU Top 2	Negative	0.88	0.94	0.91	0.85
	Neutral	0.43	0.55	0.49	
	Positive	0.91	0.72	0.81	

BiGRU Top 3	Negative	0.89	0.92	0.91	**0.85**
	Neutral	0.47	0.56	0.51	
	Positive	0.86	0.75	0.80	

Both BiLSTM and BiGRU demonstrated strong overall performance, with BiLSTM achieving 83–86% accuracy and BiGRU maintaining a stable 85% across optimal hyperparameter configurations. Both architectures yielded consistently high F1-scores for negative sentiment (0.90–0.91), reflecting the model's capacity to learn discriminative features from the majority class (63.86%). Conversely, neutral sentiment produced the lowest F1-scores (0.41–0.51), attributable to two interconnected factors: feature scarcity arising from severe underrepresentation (6.88%), and intrinsic semantic ambiguity wherein neutral reviews frequently contain conflicting polarity signals. These findings confirm that class imbalance and linguistic ambiguity are the primary drivers of per-class performance disparity, necessitating more sophisticated embedding strategies in future work.

DISCUSSIONS

In this work, investigated BiLSTM and BiGRU model performance for three-class sentiment classification of GrabFood service reviews from Google Play Store, with RoBERTa as the automatic labeling mechanism. The model performance, data characteristics considered, comparisons with previous research and implications of these findings are discussed in the next section.

Model Performance and Architecture Comparison

BiLSTM achieved the highest overall accuracy at 86%, with BiGRU closely following at 85%. However, this marginal one-percent difference in aggregate accuracy is insufficient as a sole basis for model selection; a more nuanced evaluation across per-class F1-scores, training convergence behavior, and computational efficiency is warranted. The architectural distinction between the two models has direct implications for their respective performance profiles. BiLSTM's three-gate memory mechanism affords finer control over long-range sequential dependencies, which is advantageous in sentiment tasks where polarity cues are distributed across extended contextual spans. Conversely, BiGRU's consolidated two-gate structure yields a computationally leaner model that is less susceptible to overfitting on moderately sized corpora and converges more rapidly during training — in this study, BiGRU's early stopping was triggered earlier than BiLSTM's, confirming its superior convergence efficiency (Redjeki et al., 2025).

Notably, BiGRU demonstrated more consistent per-class performance, recording a neutral class F1-score of 51% compared to BiLSTM's 50%. This suggests that BiGRU's reduced architectural complexity may confer a generalization advantage on the semantically ambiguous neutral category. Both deep learning models substantially outperformed conventional baselines: SVM achieved 82% accuracy with a neutral F1-score of 0.36, while Naïve Bayes recorded 77% accuracy with a neutral F1-score of 0.25. The performance gap reflects the fundamental limitation of TF-IDF bag-of-words representations, which treat words as independent features and fail to capture the sequential contextual dependencies essential for fine-grained sentiment discrimination (Yadav & Vishwakarma, 2020).

The Challenge of Neutral Sentiment Classification

Across all model configurations, neutral sentiment classification yielded the lowest F1-scores, ranging from 0.41 to 0.52, compared to 0.73–0.91 for positive and negative classes. Two interconnected factors underlie this pattern. First, severe class imbalance with neutral reviews constituting only 6.88% of the dataset limits the model's exposure to discriminative neutral features (Ahmed et al., 2025). While Random Oversampling was applied to address this disparity, replicating existing instances does not introduce additional linguistic diversity, constraining the model's capacity to learn robust neutral representations (Ependi et al., 2023). Second, neutral sentiment is inherently linguistically ambiguous; neutral reviews frequently contain both positive and negative lexical markers within a single text, generating semantic overlap that confounds classification boundaries. Expressions such as "tidak bisa" and "terima kasih" both prevalent in the neutral word cloud exemplify this cross-class lexical ambiguity. This finding is consistent with prior three-class sentiment studies, which uniformly report neutral as the most challenging category regardless of model architecture (Husein et al., 2023).

Comparison with Prior Studies

The accuracy results obtained in this study 86% for BiLSTM and 85% for BiGRU are competitive with and in several instances surpass those reported in comparable Indonesian-language sentiment analysis research. (Arifin & Purnama, 2023; Muhammad et al., 2021) achieved equivalent 86% accuracy on GrabFood-related Twitter data; however, their pipeline relied exclusively on TextBlob lexicon-based annotation, which is fundamentally constrained in its capacity to resolve contextual polarity, negation structures, and informal linguistic expressions

e-ISSN : 2541-2019
p-ISSN : 2541-044X

characteristic of user-generated reviews. A lexicon-based model, for instance, would systematically misclassify the Indonesian phrase "tidak terlalu jelek" (not too bad) as negative due to the presence of the word "jelek" (ugly), whereas RoBERTa correctly resolves the negation through full bidirectional contextual attention (Liu et al., 2019). Similarly, rating-based annotation which assumes direct correspondence between star scores and sentiment polarity is susceptible to labeling noise introduced by users who assign high ratings while expressing complaints, or vice versa. RoBERTa circumvents both limitations by deriving sentiment labels exclusively from textual content, without dependence on external lexical resources or user-provided ratings (Mutinda et al., 2023). (Redjeki et al., 2025) reported 83.4% accuracy using BiGRU on hospitality review corpora, while Singgalen (2024) attained 91% accuracy using LSTM and GRU on a markedly smaller dataset of 326 hotel reviews. The elevated accuracy observed in smaller-dataset studies is attributable to overfitting toward domain-specific linguistic patterns rather than acquisition of generalizable representations. The present study's corpus of 17,709 reviews introduces substantially greater lexical diversity and classification complexity, rendering the accuracy achieved proportionally more meaningful in terms of real-world applicability. This finding aligns with evidence from (Onan, 2021) and (Yadav & Vishwakarma, 2020), who demonstrated that transformer-assisted labeling and deep sequential modeling consistently outperform conventional approaches in multi-class text classification settings, particularly where informal language and class imbalance coexist.

Practical Implications for GrabFood

Besides the contribution to academics and theory, this study's results have straight implications for GrabFood service management. The 63.86% share of negative sentiment in the dataset indicates pervasive issues with service quality, which go far beyond individual complaints from users. From the above word cloud for negative sentiments, multiple bigrams like "tidak bisa" (does not work), "double order", "lama banget" (very slow), and "di cancel" (order was cancelled) show that failure to process an order, duplicate orders and delayed deliveries are among key pain-points faced by users of GrabFood. These issues lend themselves to being actionable: outputs from this sentiment classification pipeline can be provided to the GrabFood service teams and platform developers as a real-time monitoring tool to better prioritize which of the most popular complaints regarding service failures should have resolution efforts allocated to them. The challenge is also in that it captures neutral sentiment (which I call challenging) users who submit informational or improvement oriented feedback like "tolong di update", which essentially becomes a rich source of constructive input rather than being lost in the binary sentiments. Utilising the proposed pipeline at scale would allow GrabFood to transition from aggregate star-rating analysis, towards user experience measured on a text-level basis thereby driving targeted and evidence based improvements to service efficiency.

Limitations

This study acknowledges several limitations. That said, the start with automatic labeling using RoBERTa is at least more scalable than manual annotation, promoting a source for noise in labeling that could never be estimated due to an absence of gold-standard human-annotated labels. Although this dataset is on the large side, it is mostly imited by just one platform (Google Play Store) and one service (GrabFood), meaning fg to be generalisazle to all other OFDS or review channels. Moreover, both models still misclassify the neutral class which indicates that standard oversampling techniques are not adequate and future work should investigate advanced methods like cost-sensitive learning or transformer-based end-to-end classification.

CONCLUSION

The present experimental study shows that transformer-based automatic labeling using RoBERTa with bidirectional recurrent neural networks is an effective avenue for sentiment analysis of mobile application reviews. The experimental results indicates that the BiLSTM model attained highest overall accuracy (86%) and BiGRU obtained similar performance of 85% but with better stability across sentiment classes (i.e., a 51% F1-score for neutral sentiment) and convergence speed. A major conclusion of this research is that neutral sentiment remains very difficult to classify, with between 50-85% lower F1-scores than the positive and negative classes. This is partly due to class imbalance, and also because neutral expressions are often ambiguous with sentiment signals overlapping between classes. Such results reaffirm that performing sentiment analysis on mobile application reviews is a different problem versus traditional datasets like those based on social media or product reviews. To this end, this research delivers an empirical assessment of automatic labeling using RoBERTa in the Indonesian-language GrabFood reviews and comparing BiLSTM, BiGRU against traditional machine learning models. The results show that although deep learning methods are superior to traditional classification algorithms, data characteristics still limit performance and this is even greater in the case of multi-class classifiers. Future work in this area should aim at overcoming the limitations identified above, for example through hybrid human–machine annotation to improve labeling reliability, through advanced imbalance handling techniques such as cost-sensitive learning or even end-to-end transformer-based architectures that can leverage cross-sentence context.

REFERENCES

Abduljabbar, R. L., Dia, H., & Tsai, P. W. (2021). Development and evaluation of bidirectional LSTM freeway traffic forecasting models using simulation data. *Scientific Reports*, *11*(1), 23899-. https://doi.org/10.1038/s41598-021-03282-z

Ahmed, R. Y., Yuosif, N. F., Ahmed, S. A., & Mohammed, A.-B. A. (2025). Comparison of RNN and LSTM Classifiers for Sentiment Analysis of Airline Tweets. *Journal of Information Systems and Informatics*, *7*(2), 1893–1913. https://doi.org/10.51519/journalisi.v7i2.1140

Arifin, R., & Purnama, D. A. (2023). Identifying customer preferences on two competitive startupproducts: An analysis of sentiment expressions and textmining from Twitter data. *Jurnal Infotel*, *15*(1), 66–74. https://doi.org/10.20895/infotel.v15i1.906

Bhadane, C., Dalal, H., & Doshi, H. (2015). Sentiment Analysis: Measuring Opinions. *Procedia Computer Science*, *45*(C), 808–814. https://doi.org/10.1016/j.procs.2015.03.159

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *ArXiv Preprint*, 1–9. https://doi.org/10.48550/arXiv.1412.3555

Ependi, U., Rochim, A. F., & Wibowo, A. (2023). A Hybrid Sampling Approach for Improving the Classification of Imbalanced Data Using ROS and NCL Methods. *International Journal of Intelligent Engineering and Systems*, *16*(3), 345–361. https://doi.org/10.22266/ijies2023.0630.28

Hasanah, Z. M., & Hargyatni, T. (2022). Analisis pengaruh kualitas pelayanan, harga dan promosi terhadap keputusan pembelian GrabFood di kota Boyolali. *MANAJEMEN*, *2*(2), 115–124. https://doi.org/10.51903/MANAJEMEN.V2I2.173

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/NECO.1997.9.8.1735

Husein, A. M., Livando, N., Andika, A., Chandra, W., & Phan, G. (2023). Sentiment Analysis of Hotel Reviews With LSTM And ELECTRA. *Sinkron : Jurnal Dan Penelitian Teknik Informatika*, *7*(2), 733–740. https://doi.org/10.33395/sinkron.v8i2.12234

Khairani, U., Mutiawani, V., & Ahmadian, H. (2024). Pengaruh Tahapan Preprocessing Terhadap Model Indobert Dan Indobertweet Untuk Mendeteksi Emosi Pada Komentar Akun Berita Instagram. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *11*(4), 887–894. https://doi.org/10.25126/jtiik.1148315

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv Preprint*. https://doi.org/10.48550/arXiv.1907.11692

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, *26*.

Muhammad, P. F., Kusumaningrum, R., & Wibowo, A. (2021). Sentiment Analysis Using Word2vec And Long Short-Term Memory (LSTM) For Indonesian Hotel Reviews. *Procedia Computer Science*, *179*, 728–735. https://doi.org/10.1016/j.procs.2021.01.061

Mutinda, J., Mwangi, W., & Okeyo, G. (2023). Sentiment Analysis of Text Reviews Using Lexicon-Enhanced Bert Embedding (LeBERT) Model with Convolutional Neural Network. *Applied Sciences (Switzerland)*, *13*(3). https://doi.org/10.3390/app13031445

Nasution, K., Saddami, K., Roslidar, R., Akhyar, A., Fathurrahman, F., & Aulia, N. (2025). Comparative Study of BiLSTM and GRU for Sentiment Analysis on Indonesian E-Commerce Product Reviews Using Deep Sequential Modeling. *Jurnal Teknik Informatika (Jutif)*, *6*(4), 1881–1896. https://doi.org/10.52436/1.jutif.2025.6.4.4878

Onan, A. (2021). Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurrency and Computation: Practice and Experience*, *33*(23), e5909.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Rangaswamy, E., Nawaz, N., & Changzhuang, Z. (2022). The impact of digital technology on changing consumer behaviours with special reference to the home furnishing sector in Singapore. *Humanities and Social Sciences Communications*, *9*(1). https://doi.org/10.1057/s41599-022-01102-x

Redjeki, S., Joshi, B., Situmorang, A., Guntara, M., Nursari, S. R. C., & Kusumawati, D. (2025). Bidirectional GRU for Aspect-Based Sentiment Classification in Multi-Dimensional Review Analysis. *ZERO: Jurnal Sains, Matematika Dan Terapan*, *9*(2), 344–354. https://doi.org/10.30829/zero.v9i2.25754

Sabilaturrizqi, M., & Subriadi, A. P. (2024). Online food delivery adoption: In Search For Dominantly Influencing Factors. *Procedia Computer Science*, *234*(0), 1519–1528. https://doi.org/10.1016/j.procs.2024.03.153

Safira, M., & Chikaraishi, M. (2022). The impact of online food delivery service on eating-out behavior: a case of Multi-Service Transport Platforms (MSTPs) in Indonesia. *Transportation 2022 50:6*, *50*(6), 2253–2271. https://doi.org/10.1007/s11116-022-10307-7

Singgalen, Y. A. (2024). Sentiment Analysis and Trend Mapping of Hotel Reviews Using LSTM and GRU. *Journal of Information Systems and Informatics*, *6*(4), 2814–2836.

https://doi.org/10.51519/journalisi.v6i4.926

Wibowo, P., & Fatichah, C. (2021). An in-depth performance analysis of the oversampling techniques for high-class imbalanced dataset. *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, *7*(1), 63–71. https://doi.org/10.26594/register.v7i1.2206

Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, *53*(6), 4335–4385. https://doi.org/10.1007/s10462-019-09794-5