Indonesian Journal of Data and Science

Volume 6 Issue 3
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v6i3.307

Research Article

Sarcasm and Irony Detection in Lazada App Reviews Using IndoBERT

Nabila Putri Yusal [1*]; Adhitia Erfina [2]; Cecep Warman [3]

[1]Universitas Nusa Putra, Sukabumi, 43152, Indonesia, nabila.yusal_si21@nusaputra.ac.id
[2]Universitas Nusa Putra, Sukabumi, 43152, Indonesia, adhitia.erfina@nusaputra.ac.id
[3]Universitas Nusa Putra, Sukabumi, 43152, Indonesia, cecep.warman@nusaputra.ac.id

Correspondence should be addressed to Nabila Putri Yusal; nabila.yusal_si21@nusaputra.ac.id

Received 17 October 2025; Accepted 20 December 2025; Published 31 December 2025

Abstract:
Digital technology has reshaped consumer behavior, particularly in e-commerce, where Google Play Store reviews provide rich feedback but often include sarcasm and irony that conventional sentiment models misread. This study proposes an Indonesian sarcasm–irony detection model using IndoBERT, a transformer pre-trained on Indonesian corpora. A dataset of 1,998 Lazada app reviews was collected via web scraping and preprocessed through text cleaning, tokenization, and stopword removal with the Sastrawi library. IndoBERT was fine-tuned to classify reviews into three classes: sarcasm, irony, and literal. Performance was assessed using accuracy, precision, recall, F1-score, and a confusion matrix. The model achieved 96.40% accuracy, with F1-scores of 0.9725 (sarcasm), 0.9675 (irony), and 0.9267 (literal). Word cloud visualizations revealed distinct lexical patterns across classes, supporting IndoBERT's ability to capture contextual cues behind implicit sentiment. The findings indicate IndoBERT is effective for advanced opinion mining in Indonesian e-commerce, with potential applications in customer feedback monitoring, surfacing hidden complaints, and improving recommendation systems beyond surface polarity. Limitations include reliance on a single platform (Google Play) and text-only input, without modeling non-textual signals such as emojis or punctuation intensity. Future work should test cross-platform generalization, incorporate non-textual cues, and apply data augmentation to reduce class imbalance, particularly for the less frequent literal class, to improve robustness for real-world deployment.

Keywords: Natural Language Processing, IndoBERT, Sarcasm Detection, Irony Detection, E-Commerce Reviews.

1. Introduction

The rapid advancement of digital technology has transformed consumption patterns and social interactions, particularly within the e-commerce sector. Platforms such as Lazada have become one of the primary choices for Indonesian consumers in online shopping, while also serving as an open space for users to express opinions through reviews [1], [9]. These reviews hold strategic value for evaluating services; however, they often contain implicit meanings such as sarcasm and irony that make it difficult for traditional sentiment analysis systems to interpret the actual intent [2], [3].

The main challenge in detecting sarcasm and irony lies in the complexity of semantic meaning and the frequent use of non-standard language, especially in the context of social media and online reviews [4], [5]. In the broader scope of digital education and communication, the ability to comprehend implicit meanings is also a critical issue that affects perception and the validity of information [10]. To address these challenges, Natural Language Processing (NLP) technologies have continued to evolve as a primary solution for understanding the structure and context of natural language sentences [6], [8]. One of the most significant innovations in NLP is the transformer architecture, such as Bidirectional Encoder Representations from Transformers (BERT), which is capable of modeling contextual

relationships between words in a bidirectional manner [7]. IndoBERT, an Indonesian adaptation of BERT, is trained on a large corpus encompassing news articles, Wikipedia, and Indonesian web content, making it well-suited to capture sentence patterns and syntactic structures in the Indonesian language [4].

Previous studies have demonstrated the effectiveness of IndoBERT in sentiment analysis across various domains, including applications like Tiket.com [2], electric vehicle discourse [12], and political discussions ahead of general elections [3]. In addition, prior research indicates that models such as Support Vector Machines (SVM) and other traditional classification methods still face limitations in capturing the context of complex utterances, including sarcastic comments [11].

Accordingly, this study aims to develop a sarcasm and irony detection model for user reviews of the Lazada application on Google Play Store using IndoBERT. The research questions addressed are: (1) How effective is IndoBERT in detecting sarcasm and irony in user reviews? and (2) What is its impact on overall classification performance? The scope of this study is limited to Indonesian-language reviews from Google Play Store and does not consider multimodal data. Furthermore, the focus is on implementing IndoBERT without direct comparison to other models such as RoBERTa or mBERT.

This research is expected to contribute to the development of more accurate Indonesian sarcasm and irony detection systems and strengthen the application of NLP in the e-commerce sector. The remainder of this article is organized as follows: Section II reviews related work on NLP and sarcasm detection; Section III explains the research methodology; Section IV presents experimental results and analysis; and Section V concludes the study and provides recommendations for future work.

2. Method

This study adopts a quantitative approach comprising five stages: data collection of Lazada reviews through scraping, preprocessing (text cleaning, tokenization, stopword removal), sarcasm and irony classification using IndoBERT, model evaluation (confusion matrix, accuracy, precision, recall, and F1-score), and result visualization using word clouds.

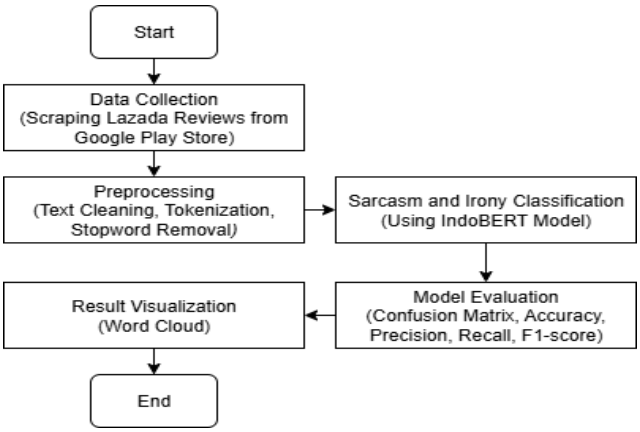

Figure 1. Research Flow

Data Selection

A total of 2,000 user reviews in Bahasa Indonesia were automatically collected from the Lazada application page on the Google Play Store using the google_play_scraper library in Python. The dataset included usernames, review content, and star ratings. The reviews were stored in structured CSV format for further processing.

Labeling Procedure

Labeling was carried out through a two-stage process. First, rule-based filtering was applied by identifying reviews where the sentiment of the text did not match the assigned rating for instance, negative content paired with

high ratings. These reviews were considered candidates for sarcasm or irony. Subsequently, two human annotators independently verified the classification into three categories: sarcasm, irony, and literal. Inter-annotator agreement was ensured through discussion and reconciliation of differences.

This study employed the following tools and libraries:

- Python 3.11.13
- google_play_scraper for data scraping [13]
- Sastrawi for text preprocessing (cleaning, tokenization, stopword removal) [15]–[17] transformers from Hugging Face for implementing IndoBERT [14]
- scikit-learn for model performance evaluation (confusion matrix, precision, recall, F1-score, and accuracy) [18], [19]
- matplotlib and wordcloud for result visualization [23]

Text Preprocessing

Text preprocessing involved several steps:

- Cleaning: Removal of punctuation, emojis, HTML tags, URLs, and user mentions.
- Tokenization: Splitting of sentences into word-level tokens.
- Stopword Removal: Elimination of non-informative words using the Sastrawi stopword list.

For example, the review "*Pelayanannya lambat banget* 😠 *tapi dapet cashback gede* 🤑" was transformed into "*pelayanan lambat cashback gede*" after cleaning and normalization.

Data Collection Process

Data collection was carried out through web scraping using the google_play_scraper library to retrieve user reviews from the Lazada application page on Google Play Store. The collected data included username, review text, and rating score, which were then stored in CSV format for further analysis [13].

Model Implementation

The IndoBERT model was fine-tuned using the labeled dataset. The model takes cleaned review texts as input and learns to classify them into one of the three categories. Each input was tokenized and converted into attention-masked tensors compatible with the transformer architecture. Model training utilized cross-entropy loss with the AdamW optimizer.

Data Analysis Methods

The review texts were processed through text cleaning, tokenization, and stopword removal to generate clean and analyzable data [15]–[17]. The primary model, IndoBERT, was employed to classify sarcasm and irony based on text input and rating scores [14].

The use of a transformer-based classification architecture is further supported by prior research demonstrating the effectiveness of neural network models for opinion classification tasks [24]. Model performance was evaluated using standard metrics, including confusion matrix, accuracy, precision, recall, and F1-score [18], [19]. In addition, word cloud visualization was applied to highlight dominant words in each category [23].

Model Evaluation Using Classification Metrics

Confusion Matrix

A confusion matrix is an evaluation table that represents the model's classification results against the actual labels. It consists of four key components [18]:

- True Positive (TP): Positive instances correctly classified as positive
- True Negative (TN): Negative instances correctly classified as negative

- False Positive (FP): Negative instances incorrectly classified as positive
- False Negative (FN): Positive instances incorrectly classified as negative

Precision

Measures how accurately the model predicts the positive class.

$$Precision = \left(\frac{TP}{TP + FP}\right) X\ 100\% \quad (1)$$

Recall

Measures the model's ability to identify all instances belonging to the positive class.

$$Recall = \left(\frac{TP}{TP + FN}\right) X\ 100\% \quad (2)$$

F1-Score

It is the harmonic mean of precision and recall, used to measure performance in a balanced manner.

$$F1 - Score\ = \frac{2 \times (Precision \times Recall)}{Precision + Recall} X\ 100\% \quad (3)$$

Accuracy

Measures how many predictions are correct out of the total data.

$$Accuracy\ = \frac{TP + TN}{TP + TN + FP + FN} X\ 100\% \quad (4)$$

3. Result and Discussion

Result

The IndoBERT model was fine-tuned to classify 1,998 Lazada app reviews into three expression categories: sarcasm, irony, and literal. The predicted class distribution is 995 sarcasm, 718 irony, and 285 literal.

Confusion Matrix

Figure 2 presents the confusion matrix for the three-class classification task. Most instances fall along the diagonal, indicating strong agreement between the predicted and true labels. Misclassifications occur primarily between sarcasm and irony—a reasonable outcome given their semantic proximity—while literal reviews show comparatively fewer cross-class errors. This pattern suggests that the model effectively captures explicit expressions and is also robust in distinguishing non-literal usage, though some overlap remains where sarcastic and ironic cues intersect.

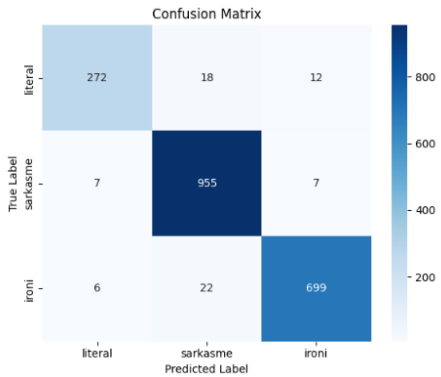

Figure 2. Confusion Matrix

Class-wise evaluation metrics derived from the confusion matrix are summarized in **Table 1**. All classes achieve high scores: irony reports the highest precision 0.9735, while sarcasm attains the highest recall 0.9856. F1-scores exceed 0.92 across all labels, and the overall accuracy reaches 0.964, indicating stable and well-balanced performance. These results confirm that IndoBERT can represent contextual nuances in Indonesian text sufficiently to separate literal from non-literal expressions at scale.

Table 1. Evaluation Model

Class	Precision	Recall	F1-Score
Literal	0.9544	0.9007	0.9267
Sarkasme	0.9598	0.9856	0.9725
Ironi	0.9735	0.9615	0.9675
Accuracy	-	-	**0.964**

Word Cloud Visualization

To explore lexical tendencies within each class, Figure 3 shows word clouds generated from the review texts grouped by predicted label. The sarcasm cloud is dominated by words such as "*udah*" and "*kecewa*," reflecting complaint-driven or mocking tone. The irony cloud surfaces ostensibly positive terms "*bagus*," "*cepat*" often used in contrastive or non-literal contexts. In the literal class, descriptive and product-related tokens such as "*barang*" and "*pengiriman*" predominate, aligning with direct experience reporting. These lexical patterns qualitatively support the quantitative classification results

Figure 3. Word Cloud Sarkasme, Ironi, and Literal

Discussion

IndoBERT achieved strong performance across all three expression labels, with F1-scores above 0.95. Class-wise analysis shows the highest precision for irony (0.9735) and the highest recall for sarcasm (0.9856), indicating the model is both selective and highly sensitive to indirect expressions—able to go beyond literal meaning and capture contextual cues associated with sarcastic or ironic intent.

These findings are consistent with prior transformer-based research demonstrating that bidirectional contextual representation improves modeling of implicit meaning and figurative language in text classification tasks. [7], [10].

Further support for the effectiveness of IndoBERT in Indonesian language understanding comes from an Automated Essay Scoring study in which a fine-tuned IndoBERT outperformed traditional feature-based baselines (e.g., TF-IDF similarity) when predicting essay scores, underscoring its strength in capturing semantic similarity and nuanced textual signals in Indonesian. [25]

In our study, the model reached an overall accuracy of 96.40%, reflecting stable behavior despite class imbalance (literal < sarcasm, irony). This level of reliability suggests IndoBERT is a strong candidate for large-scale opinion mining pipelines where non-literal user expressions must be interpreted correctly—an important need in e-commerce environments.

Practical implication: An automated sarcasm/irony detector can enhance downstream applications such as customer feedback triage, escalation of hidden complaints, and refinement of recommendation or reputation systems that currently rely on surface polarity signals.

Limitations: The dataset was drawn from a single platform (Lazada reviews on Google Play), so domain transfer to other marketplaces or social platforms remains untested. Future work should include cross-platform validation, multimodal extensions incorporating non-textual cues (emoji, punctuation intensity, star ratings), and class-balancing or augmentation strategies to improve performance on lower-frequency literal reviews.

4. Conclusion

This study applied IndoBERT to detect sarcasm, irony, and literal expressions in 1,998 Lazada app reviews from Google Play Store. The model achieved an overall accuracy of 96.40% with F1-scores above 0.95 for all classes, indicating strong capability in capturing explicit and implicit meanings within user reviews. These findings confirm IndoBERT's effectiveness for sarcasm and irony detection in Indonesian text and its ability to maintain robust multi-class performance.

The research contributes to Indonesian NLP by demonstrating how a transformer-based model can overcome the limitations of traditional sentiment analysis in handling non-literal expressions. This approach offers practical value for automated opinion mining in e-commerce, supporting applications such as customer feedback analysis and recommendation systems.

Future studies should extend this work by validating the model across multiple platforms and domains, incorporating multimodal features such as emojis and punctuation patterns, and applying augmentation or class-balancing strategies to improve robustness for underrepresented categories.

Reference

[1] S. Rajwal, "LiHiSTO: a comprehensive list of Hindi stopwords," *Multimed. Tools Appl.*, vol. 83, no. 17, pp. 50047–50059, 2024, doi: 10.1007/s11042-023-17205-9.

[2] D. Nuryadi *et al.*, "Fine Tuning Indobert Untuk Analisis Sentimen Pada Ulasan Pengguna Aplikasi Tiket.Com Di Google Play Store," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 2, pp. 3577–3583, 2025, doi: 10.36040/jati.v9i2.13204.

[3] L. Geni, E. Yulianti, and D. I. Sensuse, "Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using IndoBERT Language Models," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 9, no. 3, pp. 746–757, 2023, doi: 10.26555/jiteki.v9i3.26490.

[4] H. H. Friedman, "The education irony: when college degrees lead to unemployment, mindless thinking, debt, and despair," *Acad. Ment. Heal. Well-Being*, vol. 2, no. 2, pp. 1–10, 2025, doi: 10.20935/mhealthwellb7661.

[5] 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.

[6] S. Arora *et al.*, "Simple linear attention language models balance the recall-throughput tradeoff," *Proc. Mach. Learn. Res.*, vol. 235, pp. 1763–1840, 2024.

[7] M. D. Hilmawan, "Deteksi Sarkasme Pada Judul Berita Berbahasa Inggris Menggunakan Algoritme Bidirectional LSTM," *J. Dinda Data Sci. Inf. Technol. Data Anal.*, vol. 2, no. 1, pp. 46–51, 2022, doi: 10.20895/dinda.v2i1.331.

[8] A. J. Putri, A. S. Syafira, M. E. Purbaya, and D. Purnomo, "Analisis Sentimen E-Commerce Lazada pada Jejaring Sosial Twitter Menggunakan Algoritma Support Vector Machine," *J. TRINISTIK J. Tek. Ind. Bisnis Digit. dan Tek. Logistik*, vol. 1, no. 1, pp. 16–21, 2022, doi: 10.20895/trinistik.v1i1.447.

[9] K. A. Pradani and L. H. Suadaa, "Automated Essay Scoring Menggunakan Semantic Textual Similarity Berbasis Transformer Untuk Penilaian Ujian Esai," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 6, pp. 1177–

1184, 2023, doi: 10.25126/jtiik.2023107338.

[10] G. Z. Nabiilah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, "Indonesian multilabel classification using IndoBERT embedding and MBERT classification," *Int. J. Electr. Comput. Eng.*, vol. 14, no. 1, pp. 1071–1078, 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.

[11] B. V. Kartika, M. J. Alfredo, and G. P. Kusuma, "Fine-Tuned IndoBERT based model and data augmentation for indonesian language paraphrase identification," *Rev. d'Intelligence Artif.*, vol. 37, no. 3, pp. 733–743, 2023, doi: 10.18280/ria.370322.

[12] S. C. M. D. S. Sirisuriya, "Importance of Web Scraping as a Data Source for Machine Learning Algorithms - Review," *2023 IEEE 17th Int. Conf. Ind. Inf. Syst. ICIIS 2023 - Proc.*, pp. 134–139, 2023, doi: 10.1109/ICIIS58898.2023.10253502.

[13] V. Çetin and O. Yıldız, "A comprehensive review on data preprocessing techniques in data analysis," *Pamukkale Univ. J. Eng. Sci.*, vol. 28, no. 2, pp. 299–312, 2022, doi: 10.5505/pajes.2021.62687.

[14] E. Y. Daraghmi, S. Qadan, Y. A. Daraghmi, R. Yousuf, O. Cheikhrouhou, and M. Baz, "From Text to Insight: An Integrated CNN-BiLSTM-GRU Model for Arabic Cyberbullying Detection," *IEEE Access*, vol. 12, no. August, pp. 103504–103519, 2024, doi: 10.1109/ACCESS.2024.3431939.

[15] E. Dotan, G. Jaschek, T. Pupko, and Y. Belinkov, "Effect of tokenization on transformers for biological sequences," *Bioinformatics*, vol. 40, no. 4, pp. 1–15, 2024, doi: 10.1093/bioinformatics/btae196.

[16] E. Helmud, E. Helmud, F. Fitriyani, and P. Romadiana, "Classification Comparison Performance of Supervised Machine Learning Random Forest and Decision Tree Algorithms Using Confusion Matrix," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 13, no. 1, pp. 92–97, 2024, doi: 10.32736/sisfokom.v13i1.1985.

[17] Z. Jannah, R. Kurniawan, and S. Anwar, "Studi Algoritma Neural Network Dalam Klasifikasi Sentimen Pengguna Shopee: Peningkatan Akurasi Model," *J. Inform. dan Tek. Elektro Terap.*, vol. 13, no. 2, 2025, doi: 10.23960/jitet.v13i2.6113.

[18] A. Upadhyay *et al.*, "Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture," *Artif. Intell. Rev.*, vol. 58, no. 3, 2025, doi: 10.1007/s10462-024-11100-x.

[19] Z. Niu *et al.*, "Piscis: a novel loss estimator of the F1 score enables accurate spot detection in fluorescence microscopy images via deep learning," *bioRxiv*, pp. 1–21, 2024, [Online]. Available: https://doi.org/10.1101/2024.01.31.578123

[20] C. Ma *et al.*, "Multi-objective topology optimization for cooling element of precision gear grinding machine tool," *Int. Commun. Heat Mass Transf.*, vol. 160, no. November 2024, p. 108356, 2025, doi: 10.1016/j.icheatmasstransfer.2024.108356.

[21] M. Furqan, S. Sriani, and M. N. Shidqi, "Chatbot Telegram Menggunakan Natural Language Processing," *Walisongo J. Inf. Technol.*, vol. 5, no. 1, pp. 15–26, 2023, doi: 10.21580/wjit.2023.5.1.14793.

[22] S. Sharma and P. Chaudhary, "Machine learning and deep learning," *Quantum Comput. Artif. Intell. Train. Mach. Deep Learn. Algorithms Quantum Comput.*, pp. 71–84, 2023, doi: 10.1515/9783110791402-004.

[23] M. Munir and D. Darmawan, "The Role of Trust, Ease of Use and Security on Shopping Interests at Lazada,"

Eng. Technol. Int. J., vol. 4, no. 3, pp. 81–90, 2022.

[24] N. M. Gardazi, A. Daud, M. K. Malik, A. Bukhari, T. Alsahfi, and B. Alshemaimri, "BERT applications in natural language processing: a review," *Artif. Intell. Rev.*, vol. 58, no. 6, 2025, doi: 10.1007/s10462-025-11162-5.

[25] A. Pramudita, A. F. Nugroho, and T. B. Adji, "Aspect-based sentiment analysis for Indonesian hotel reviews using multilingual BERT," Indonesian Journal of Electrical Engineering and Computer Science, vol. 23, no. 1, pp. 456–464, Jan. 2026, doi: 10.11591/ijeecs.v23.i1.pp456-464.