Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15607
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Sarcasm Detection in Indonesian YouTube Comments using Fine-Tuned IndoBERT with Class Imbalance Handling

Ahmad Muhlis Fanani[1)], Moh. Iwan Wahyuddin[2)*]
[1,2)]Magister Teknologi Informasi, Fakultas Teknologi Komunikasi dan Informatika, Universitas Nasional, Jakarta , Indonesia
[1)]ahmadmuhlisfanani.2024@civitas.unas.ac.id, [2)]iwan.wahyuddi@civitas.unas.ac.id

Submitted :Nov 21,2025 | **Accepted** : Nov 26, 2025 | **Published** : Jan 02, 2026

Abstract: Sarcasm detection in Indonesian social media faces challenges in natural language processing due to implicit meanings and limited labeled datasets. YouTube, with 143 million users in Indonesia, represents a largely unexplored source of sarcastic expressions. This study aims to develop an automatic sarcasm detection system for Indonesian YouTube comments using fine-tuned IndoBERT and evaluate the performance of two IndoBERT variants. A dataset of 5,291 YouTube comments was collected and automatically labeled using GPT-4o with structured prompts based on linguistic indicators of sarcasm. Two IndoBERT variants (IndoNLU and IndoLEM) were fine-tuned with three class imbalance mitigation strategies: imbalanced, under-sampling, and class weighting. Zero-shot evaluation was conducted as a baseline to measure fine-tuning effectiveness. Models were evaluated using accuracy, precision, recall, and F1-score metrics. Pre-trained models without fine-tuning showed very limited sarcasm detection capability with F1-scores of 0.1613 for IndoNLU and 0.3519 for IndoLEM. Fine-tuning with under-sampling dramatically improved F1-scores to 0.6499 for IndoNLU and 0.6568 for IndoLEM, showing improvements up to 303%. IndoBERT-IndoNLU provided more balanced performance with 0.6424 accuracy, while IndoLEM showed higher sarcasm recall of 0.7639. Fine-tuning IndoBERT is effective for detecting sarcasm in Indonesian YouTube comments. This study contributes by providing a new labeled dataset, demonstrating the effectiveness of automatic labeling using large language models, and providing empirical evidence of the significant value of fine-tuning for Indonesian sarcasm detection.

Keywords: BERT; class imbalance; GPT-40; IndoBERT; Indonesian language; natural language processing; sarcasm detection; YouTube comments

INTRODUCTION

Sarcasm is a form of indirect communication that conveys criticism or mockery through expressions that appear positive or neutral but carry opposite pragmatic meanings. In digital contexts, sarcasm has become a common phenomenon, particularly in social media comments. However, its implicit and ambiguous nature poses significant challenges for automatic systems, especially in sentiment analysis tasks, as it can unexpectedly reverse meaning polarity. Therefore, accurate sarcasm detection is important for improving the accuracy of natural language processing (NLP) systems and for content moderation and public opinion analysis purposes.

YouTube was selected as the research platform based on Indonesia's very large user base. According to We Are Social and Meltwater (2025), YouTube's advertising reach in Indonesia reaches 143 million users, or about 50% of the total national population, making Indonesia the country with the fourth largest YouTube users in the world. With a broad user base and open interaction characteristics, YouTube functions not only as entertainment media but also as a major digital public space where people express opinions, criticism, and sarcasm.

In NLP, sarcasm presents unique challenges because its meaning often contradicts literal sentence structure. Jia et al., (2024) explain that semantic ambiguity in sarcastic utterances can cause sentiment models to fail to recognize true polarity, especially without considering pragmatic and multimodal context. This is reinforced by Sharma et al. (2022) who show that the presence of sarcasm significantly reduces the accuracy of automatic sentiment analysis systems.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v10i1.15607 p-ISSN : 2541-044X

Early efforts in Indonesian sarcasm detection were conducted by Suhartono et al. (2024) through the development of the IdSarcasm dataset sourced from Twitter and Reddit platforms. Although significant, the dataset does not yet reflect the specific linguistic characteristics of YouTube comments. Previous studies on YouTube comments have focused on sentiment and emotion analysis but have not specifically addressed sarcasm as a distinct linguistic phenomenon.

Based on the literature review, three main gaps were identified in Indonesian sarcasm detection research. First, there is no labeled sarcasm dataset specifically for Indonesian YouTube comments, which have unique characteristics such as longer text length, mixed formal-informal language style, and topic diversity influenced by video content. Second, the effectiveness of fine-tuning pre-trained Indonesian language models (IndoBERT) for sarcasm detection on the YouTube platform has not been empirically tested. Third, there is no comparative study between different IndoBERT variants (IndoNLU vs IndoLEM) to understand the influence of pretraining corpus characteristics on sarcasm detection performance.

To address these gaps, this research aims to evaluate the effectiveness of fine-tuning IndoBERT models for sarcasm detection on Indonesian YouTube comments and compare the performance of two IndoBERT variants. IndoBERT was chosen because it is specifically trained on large Indonesian corpora and consistently outperforms multilingual models in various Indonesian language classification tasks (Wilie et al. 2020). This study also evaluates IndoBERT IndoLEM Koto et al. (2020) to observe the influence of different pretraining approaches on sarcasm detection performance.

This research not only aims to develop a classification model that is more sensitive to forms of irony in Indonesian but also contributes to enriching local NLP literature, particularly in the context of YouTube sarcasm detection. The research results are expected to become a foundation for developing automatic moderation systems and public opinion analysis that are more adaptive to the digital communication style of Indonesian society.

LITERATURE REVIEW

Sarcasm is a form of language style that is linguistically categorized as indirect speech, where the conveyed meaning contradicts its literal meaning. In pragmatic theory, sarcasm is often considered a form of violation of Grice (1975) maxim of quality, which suggests that speakers should convey something true. In sarcastic utterances, speakers instead convey statements that are literally inconsistent with their actual intent, so understanding sarcasm requires contextual reasoning and recognition of emotional contrasts in sentences.

According to Oxford Learner's Dictionaries, sarcasm is defined as a way of using words that are opposite to the actual intent to be unpleasant or mock someone. This definition emphasizes the role of meaning reversal as a key feature of sarcasm. Meanwhile, Kamus Besar Bahasa Indonesia (KBBI) defines sarcasm as the use of harsh words, mockery, or rough ridicule to hurt someone's feelings, showing a focus on emotional intensity and verbal intent to harm.

In the NLP context, sarcasm presents unique challenges because its meaning often contradicts literal sentence structure. Razali et al. (2021) show that sarcasm can affect sentiment analysis results and cause prediction errors because the true meaning is not explicitly conveyed. To address this, context-based approaches become important so models can understand hidden semantic and emotional information.

Several early studies attempted to address sarcasm detection challenges using manual feature-based approaches and classic classification algorithms such as SVM or Random Forest. However, such conventional approaches have limitations in capturing implicit meanings, especially without comprehensive contextual understanding (Ranti and Girsang 2020; Suhartono, Wongso, and Tri Handoyo 2024).

With the development of transformer architectures, models like BERT have shown better performance in understanding sentence structure and context. Sharma et al. (2022) show that hybrid deep learning approaches successfully improve sarcasm detection accuracy compared to conventional methods because they can capture emotional nuances and implicit meaning shifts in text. Qin et al. (2025) integrate transformers with gated graph neural networks to detect sarcasm in online content, demonstrating the importance of structural context understanding.

Early efforts in Indonesian sarcasm detection were conducted by Suhartono et al. (2024) through the development of the IdSarcasm dataset sourced from Twitter and Reddit platforms. The research identified that sarcastic comments in Indonesian often do not contain explicit cues such as hashtags or emojis, so detection depends on overall sentence context. Although significant, the dataset does not yet reflect the specific linguistic characteristics of YouTube comments which have text length, mixed language style, and diversity due to the video topics being responded to.

Several local studies have utilized YouTube comments for sentiment and emotion analysis tasks. Mandhasiya, Murfi, and Bustamam (2024) evaluated various hybrid deep learning models with BERT representations in the Indonesian political context, finding that BERT-CNN provides the highest accuracy in classifying positive and negative comments. Islam et al (2025) show that hybrid BERT-CNN approaches can improve classification performance for emotional and opinion comments.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15607
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Ma'aly et al. (2024) used a multi-label classification approach to identify emotions from YouTube comments using Bi-LSTM models achieving an AUC score of 0.91. The research also utilized GPT-3.5 for automatic emotion annotation. Nevertheless, both studies did not specifically address sarcasm as a distinct linguistic phenomenon.

The BERT (Bidirectional Encoder Representations from Transformers) model introduced by (Devlin et al. 2019) has revolutionized the NLP field with a bidirectional pretraining approach that allows models to understand left and right context simultaneously. The transformer architecture underlying BERT uses a self-attention mechanism to calculate the importance weight of each word in a sentence relative to other words, regardless of their distance in the sequence.

In the Indonesian language context, IndoBERT was developed by Wilie et al. (2020) as a transformer-based model specifically trained on large Indonesian corpora (Indo4B) covering formal and informal varieties from various domains. According to the IndoNLU benchmark, IndoBERT consistently surpasses the performance of multilingual models such as mBERT and XLM-R in various Indonesian language classification and information extraction tasks. Koto et al. (2020) developed the IndoBERT-IndoLEM variant with a different pretraining approach to evaluate the influence of corpus sources on NLP task performance.

The development of large language models (LLMs) such as GPT opens new opportunities in natural language processing, including data labeling tasks. Gole et al. (2024) tested the performance of various GPT models for sarcasm detection and found that the GPT-3 davinci model fine-tuned specifically on sarcasm data successfully outperformed the zero-shot GPT-4 approach in terms of accuracy and F1-score (0.81 vs 0.75). However, they also found that GPT-3.5 Turbo has limitations in recognizing sarcasm compared to fine-tuned models.

Qin et al. (2025) emphasize that LLMs often face difficulties detecting subtle sarcasm due to their tendency to rely on probabilistic patterns, thus risking missing irony nuances and pragmatic contrasts. Based on these findings, this research designs an automatic labeling strategy that relies on GPT-4o with emphasis on prompt design as the key to label quality.

The literature review reveals that sarcasm detection has evolved from classical approaches using algorithms such as SVM with manually defined features, to transformer-based models capable of learning contextual representations automatically with significantly improved performance. IndoBERT has proven effective for various Indonesian NLP tasks, but its application for sarcasm detection on YouTube comments remains unexplored. Previous Indonesian sarcasm research (IdSarcasm) focused on Twitter and Reddit platforms, which have different linguistic characteristics from YouTube. Additionally, while LLMs like GPT-4o show promise for automatic labeling, their effectiveness for Indonesian sarcasm annotation has not been validated. This study addresses these gaps by: (1) creating a YouTube-specific Indonesian sarcasm dataset using GPT-4o labeling, (2) empirically testing fine-tuned IndoBERT performance on this new domain, and (3) comparing IndoNLU and IndoLEM variants to understand pretraining corpus influence on sarcasm detection.

METHOD

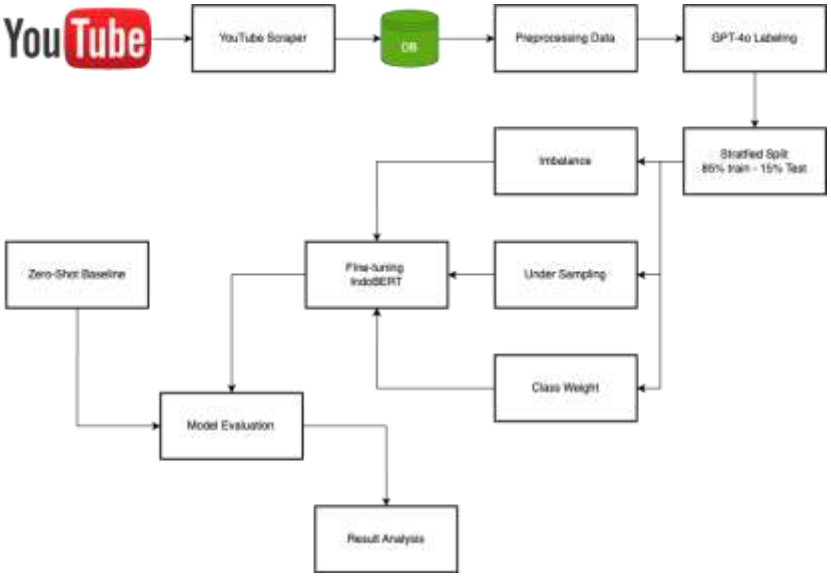

Fig.1. Proposed method YouTube Comments Detection Sarcasm using Fine-Tuned IndoBERT

Figure 1 shows that this research method follows an end-to-end flow as shown in the diagram. Data was collected through web scraping of YouTube comments using the Python library youtube-comment-downloader.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15607
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Videos were purposively selected based on topics with high potential for containing sarcastic comments, such as politics, public issues, military topics, and comedy content. A total of 5,291 comments were collected after filtering.

Preprocessing steps included removing duplicate comments, spam and gambling promotions, emoji characters, foreign characters and non-text symbols, URLs and mentions, timestamps and excessive spaces, and converting all text to lowercase. This preprocessing ensures data quality while maintaining the informal language characteristics typical of social media. All comments were automatically labeled using GPT-4o through API calls, following the approach demonstrated by Gole et al. (2024) for sarcasm detection using large language models. The labeling system was based on structured prompts containing operational definitions of sarcasm and explicit linguistic indicators such as irony, hyperbole, and veiled praise. The model classified comments into two classes: sarcasm (1) and non-sarcasm (0).

Sarcasm classification criteria included meaning contrast between literal expression and actual intent, irony or hyperbole, and veiled criticism through implicit expressions, analogies, or cynical assessments. Comments not showing these indicators were labeled as non-sarcasm. This approach enables efficient and consistent labeling across the dataset. The final dataset consisted of 1,555 sarcastic comments and 3,736 non-sarcastic comments, showing class imbalance requiring mitigation strategies during model training.

Two IndoBERT variants were fine-tuned: IndoBERT-IndoNLU (indobenchmark/indobert-base-p1) and IndoBERT-IndoLEM (indolem/indobert-base-uncased). Both models have similar architectures but differ in pretraining corpus and objectives (Koto et al. 2020; Wilie et al. 2020). A classification layer with sigmoid activation function was added on top of the [CLS] token output for binary classification, following the standard BERT fine-tuning approach (Devlin et al. 2019).

The dataset was split using stratified sampling with 85% for training and 15% for testing. Three class imbalance mitigation strategies were tested: imbalanced without adjustment, under-sampling the majority class, and class weighting during training. Training parameters were: 4 epochs, batch size 16 (training) and 32 (evaluation), learning rate 8e-5, 500 warmup steps, weight decay 0.01, maximum sequence length 128 tokens, using AdamW optimizer and Binary Cross-Entropy with Logits loss function.

All experiments were conducted locally on a MacBook Pro with Apple M4 Pro chip, 24 GB RAM, and Python 3.10. The fine-tuning process used the Huggingface Transformers library (v4.35.0) with PyTorch (v2.1.0) backend optimized for Apple Silicon through Metal Performance Shaders (MPS). Each fine-tuning run took approximately 5-10 minutes per epoch depending on the dataset size. For automatic labeling, GPT-4o was accessed through OpenAI API with the following system prompt:

"Anda adalah asisten yang bertugas mengklasifikasikan komentar YouTube ke dalam dua kategori: 'sarkasme/1' atau 'bukan sarkasme/0' HANYA berdasarkan teks yang diberikan. Definisi sarkasme untuk tugas ini (cukup memenuhi SALAH SATU kriteria): mengandung kontras antara makna literal dan maksud sebenarnya; berupa pujian palsu, hiperbola, atau sindiran halus; mengandung unsur ironi atau kontras makna. BUKAN sarkasme: kritik langsung dan lugas; hinaan atau makian biasa; humor tanpa sindiran; komentar netral atau positif."

To measure fine-tuning effectiveness, zero-shot evaluation was conducted as a baseline using pre-trained models without task-specific training. The non-fine-tuned IndoBERT-IndoNLU and IndoLEM models were evaluated directly on the same test set as the fine-tuned models. This evaluation aims to measure the inherent capability of pre-trained models in detecting sarcasm and quantify the improvement resulting from fine-tuning.

Model performance was evaluated using standard classification metrics calculated from the confusion matrix. Accuracy measures overall correctness across both classes:

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

Precision measures the proportion of correct sarcasm predictions among all sarcasm predictions:

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

Recall (sensitivity) measures the proportion of actual sarcastic comments correctly identified:

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

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15607
e-ISSN : 2541-2019
p-ISSN : 2541-044X

F1-score provides the harmonic mean of precision and recall, balancing both metrics:

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

where TP (True Positive) represents correctly identified sarcastic comments, TN (True Negative) represents correctly identified non-sarcastic comments, FP (False Positive) represents non-sarcastic comments incorrectly classified as sarcastic, and FN (False Negative) represents sarcastic comments incorrectly classified as non-sarcastic.

Special attention was given to F1-score for the sarcasm class as the main indicator of model sensitivity to sarcastic expressions. Confusion matrix was also analyzed to understand error distribution patterns. Improvement percentage was calculated by comparing fine-tuned model F1-scores against the zero-shot baseline.

The training process used Binary Cross-Entropy with Logits as the loss function:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[\, y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \quad (5)$$

where y_i is the true label (0 or 1), $\hat{y}_i$ is the predicted probability, and N is the number of samples in the batch.

RESULT

This study evaluated two IndoBERT variants (IndoNLU and IndoLEM) for sarcasm detection on Indonesian YouTube comments. The key findings are: (1) zero-shot models showed very limited capability with sarcasm F1-scores of only 0.1613 (IndoNLU) and 0.3519 (IndoLEM); (2) fine-tuning dramatically improved performance, with IndoNLU achieving 303% improvement (F1: 0.6499) and IndoLEM achieving 87% improvement (F1: 0.6568); (3) under-sampling proved most effective for handling class imbalance; and (4) IndoNLU provides more balanced predictions while IndoLEM shows higher sensitivity but more false positives. Table 1 presents a comprehensive comparison of zero-shot versus fine-tuned performance for both models.

Table 1. Zero-Shot vs Fine-Tuned Performance Comparison

Metric	**IndoNLU Zero-Shot**	**IndoNLU Fine-Tuned**	**IndoLEM Zero-Shot**	**IndoLEM Fine-Tuned**
Accuracy	0.6725	0.6424	0.5176	0.6017
F1-score (Macro)	0.4789	**0.6422**	0.4839	0.5912
Precision (Sarcasm)	0.3247	**0.6352**	0.2905	0.5761
Recall (Sarcasm)	0.1073	**0.6652**	0.4464	**0.7639**
F1-score (Sarcasm)	0.1613	**0.6499**	0.3519	**0.6568**
True Positive	25	**155**	104	**178**
False Positive	52	89	254	131
Improvement	—	**+303%**	—	**+87%**

Zero-shot evaluation revealed that pre-trained models have very limited sarcasm detection capability. IndoBERT IndoNLU only detected 25 out of 233 sarcastic comments (10.73%), while IndoLEM detected 104 (44.64%). IndoLEM's better zero-shot performance suggests its informal pretraining corpus provides initial

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15607
e-ISSN : 2541-2019
p-ISSN : 2541-044X

"sarcasm intuition." However, both models' performance was inadequate for practical applications without task-specific training.

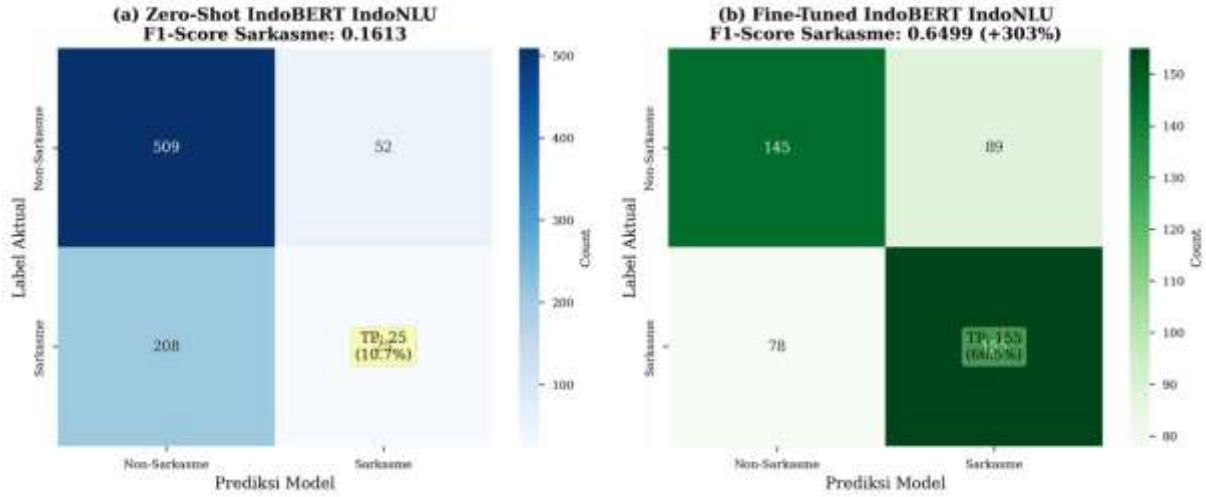

Fig.2. Confusion Matrix Comparison: Zero-Shot vs Fine-Tuned IndoBERT-IndoNLU

Figure 2 shows the confusion matrix comparison between zero-shot and fine-tuned models for IndoBERT-IndoNLU. A dramatic improvement is seen in True Positives increasing from 25 (10.73%) to 155 (66.52%), while False Negatives decreased from 208 to 78. This visualization confirms the effectiveness of task-specific training for sarcasm detection, where fine-tuning significantly improves the model's ability to recognize sarcastic comments.

Table 2 presents IndoBERT-IndoNLU evaluation results using three training strategies. The imbalanced strategy achieved the highest overall accuracy (0.7355) but showed low performance on the sarcasm class (recall: 0.4249, F1-score: 0.4853). The under-sampling strategy provided the best performance for sarcasm detection with recall of 0.6652 and F1-score of 0.6499, successfully detecting 155 out of 233 sarcastic comments (66.52%). Although overall accuracy decreased to 0.6424, this strategy showed the best balance between precision (0.6352) and recall. The class weighting strategy showed intermediate performance but did not surpass under-sampling for sarcasm detection.

Table 2. IndoBERT-IndoNLU Performance on Three Fine-Tuning Strategies

Metric	Imbalanced	Under-Sampling	Class Weighting
Test data (sarcasm/non-sarcasm)	233/561	233/234	233/561
Training data	4,497	2,643	4,497
Accuracy	0.7355	0.6424	0.7217
F1-score (Macro)	0.7232	0.6422	0.7090
Sarcasm Metrics			
Precision (Sarcasm)	0.5657	0.6352	0.5341
Recall (Sarcasm)	0.4249	**0.6652**	0.4034
F1-score (Sarcasm)	0.4853	**0.6499**	0.4597
Non-sarcasm Metrics			
F1-score (Non-sarcasm)	0.8220	0.6346	0.8126
Confusion Matrix			
True Positive	99	**155**	94
False Negative	134	**78**	139

Based on these results, under-sampling was selected as the optimal strategy for sarcasm detection, as it achieved the highest sensitivity to sarcastic comments despite lower overall accuracy.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15607
e-ISSN : 2541-2019
p-ISSN : 2541-044X

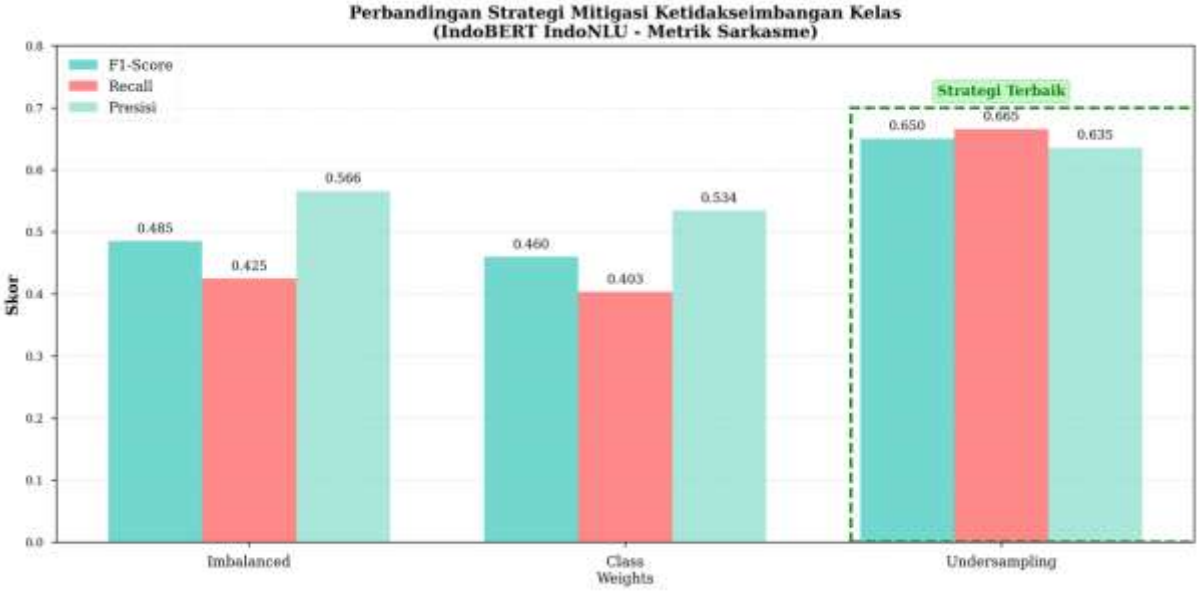

Fig. 3. Class Imbalance Mitigation Strategy Comparison

Figure 3 visualizes the comparison of three class imbalance mitigation strategies on IndoBERT-IndoNLU. Undersampling (marked with green box) provides the best performance for sarcasm detection with F1-score of 0.6499 and recall of 0.6652, surpassing imbalanced strategy (F1: 0.4853) and class weighting (F1: 0.4597). This result confirms that balanced data is more effective for training models to recognize sarcasm patterns compared to only adjusting loss function weights.

IndoBERT-IndoLEM was evaluated using the under-sampling strategy following IndoNLU results. Table 3 shows that IndoLEM achieved the highest sarcasm recall (0.7639), successfully detecting 178 out of 233 sarcastic comments (76.39%). This model shows very high sensitivity to sarcastic expressions with F1-score of 0.6568. However, this recall advantage comes with trade-offs of lower precision (0.5761) and increased false positives (131 cases), indicating a tendency to over-predict sarcasm. Performance on the non-sarcasm class decreased significantly with recall of only 0.4402 and F1-score of 0.5255.

Table 3. IndoBERT-IndoLEM Performance with Under-Sampling

Metric	Value
Accuracy	0.6017
Precision (macro)	0.6140
Recall (macro)	0.6021
F1-score (macro)	0.5912
Sarcasm Metrics	
Precision (Sarcasm)	0.5761
Recall (Sarcasm)	**0.7639**
F1-score (Sarcasm)	**0.6568**
Non-sarcasm Metrics	
Precision (Non-sarcasm)	0.6519
Recall (Non-sarcasm)	0.4402
F1-score (Non-sarcasm)	0.5255
Confusion Matrix	
True Positive	**178**
False Positive	131
False Negative	55
True Negative	103

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15607
e-ISSN : 2541-2019
p-ISSN : 2541-044X

IndoBERT IndoNLU shows more balanced performance on both classes with overall accuracy of 0.6424 and macro F1-score of 0.6422. IndoBERT IndoLEM shows higher sarcasm detection capability (178 vs 155 true positives) but with more false positives (131 vs 89), indicating aggressive sarcasm prediction behavior. IndoNLU is recommended for general applications requiring balanced predictions, while IndoLEM is suitable for content moderation scenarios where high sarcasm sensitivity is prioritized.

Figure 4 displays a radar chart comparing the performance of both IndoBERT variants across various metrics. IndoLEM (red) excels in sarcasm recall (0.7639) but with lower precision (0.5761), showing a tendency to over-predict sarcasm. IndoNLU (blue) provides more balanced performance with an F1-score gap between classes of only 1.53%, making it more stable for general applications.

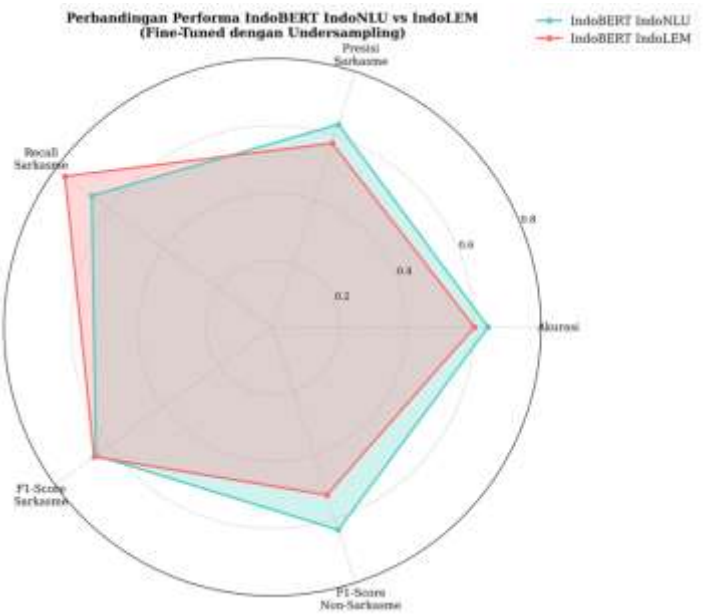

Fig.4. IndoBERT-IndoNLU vs IndoLEM Performance Comparison

DISCUSSIONS

The comparison between zero-shot baseline and fine-tuned models demonstrates the dramatic effectiveness of task-specific training. IndoBERT-IndoNLU showed a 303% increase in sarcasm F1-score (from 0.1613 to 0.6499) and a 520% increase in recall (from 0.1073 to 0.6652). IndoBERT-IndoLEM showed an 87% increase in F1-score (from 0.3519 to 0.6568) and a 71% increase in recall (from 0.4464 to 0.7639).

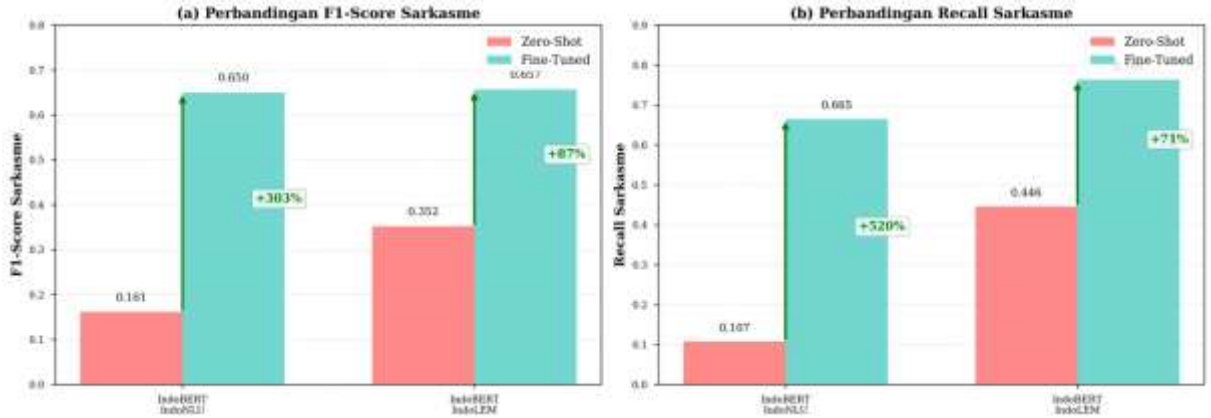

Fig.5. Zero-Shot vs Fine-Tuned Performance Comparison

Figure 5 visualizes the dramatic improvements resulting from fine-tuning. Panel (a) shows sarcasm F1-score increasing 303% for IndoNLU (0.1613 → 0.6499) and 87% for IndoLEM (0.3519 → 0.6568). Panel (b) displays even more dramatic recall improvements: 520% for IndoNLU and 71% for IndoLEM. Green arrows and improvement percentages clarify the magnitude of improvement, demonstrating that fine-tuning is not just beneficial but essential for effective sarcasm detection. Interestingly, although IndoLEM had better zero-shot

*name of corresponding author

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

performance, IndoNLU gained greater benefits from fine-tuning. This indicates that IndoLEM's more informal pretraining corpus provides initial "sarcasm intuition," but IndoNLU with more formal pretraining has greater capacity to learn sarcasm patterns through fine-tuning. This finding confirms that pre-trained models alone are insufficient for sarcasm detection—task-specific training is essential to achieve adequate performance.

It is worth noting that this study focuses on transformer-based models (IndoBERT) rather than classical machine learning approaches such as SVM or Random Forest. Previous research has shown that conventional methods with manually defined features have limitations in capturing implicit meanings inherent in sarcastic expressions (Ranti and Girsang 2020; Suhartono et al. 2024). The transformer architecture's ability to learn contextual representations provides significant advantages for understanding the semantic contrasts characteristic of sarcasm.

Experimental results demonstrate that class imbalance mitigation strategies significantly impact model performance in sarcasm detection. The under-sampling strategy achieved the highest F1-score for the sarcasm class, confirming that models better recognize sarcasm patterns when trained on balanced data, despite consequences of decreased performance on the non-sarcasm class. This finding can be explained through imbalanced data characteristics: models tend to be biased toward the majority class (non-sarcasm) when there is no distribution adjustment, causing sarcasm cases to be frequently misclassified. With under-sampling, the number of examples in each class becomes equal, forcing the model to pay more attention to characteristic features of sarcastic comments, such as irony, hyperbole, or praise implying sarcasm. Conversely, the class weighting strategy, although theoretically helpful, did not produce better F1-scores for the sarcasm class compared to under-sampling. This may be because limited feature representation in the minority class is not fully compensated by only adjusting loss function weights.

The comparison between IndoBERT-IndoNLU and IndoLEM reveals important insights about pretraining influence on sarcasm detection and the trade-off between recall and precision. IndoLEM shows very high sarcasm recall (0.7639 vs 0.6652), successfully detecting 23 more sarcastic comments (178 vs 155) and reducing false negatives by 29.5% (55 vs 78). However, this recall advantage comes with significant trade-offs: lower sarcasm precision (0.5761 vs 0.6352), 47% increase in false positives (131 vs 89), and drastically decreased non-sarcasm performance (recall 0.4402 vs 0.6197). IndoBERT-IndoNLU provides more balanced performance distribution across classes with an F1-score gap of only 1.53% (0.6346 vs 0.6499), compared to IndoLEM which has a gap of 13.13% (0.5255 vs 0.6568). This characteristic makes IndoNLU more stable and reliable in handling comment variations on platforms like YouTube. Balanced performance is crucial for practical applications where precision and recall are equally important.

IndoLEM is more suitable for applications prioritizing maximum sarcasm detection such as content moderation or early warning systems, where false positives can be tolerated or manually filtered. Conversely, IndoNLU is better suited for general classification systems requiring consistent and balanced performance on both classes without significant bias.

To contextualize this research's results, comparison was made with previous Indonesian sarcasm detection studies, particularly Suhartono et al. (2024) who developed the IdSarcasm dataset from Twitter and Reddit. Their research evaluated various pre-trained language models on a dataset consisting of 14,116 samples for Reddit and 2,684 samples for Twitter. For fair comparison, focus was given to their IndoBERT results which achieved macro F1-scores of 0.7273 for the Twitter subset and 0.6100-0.6184 for the Reddit subset. This study achieved macro F1-scores of 0.6422 for IndoNLU and 0.5912 for IndoLEM on YouTube comments, with sarcasm class F1-scores of 0.6499 and 0.6568 respectively.

This comparison reveals important insights about platform characteristics' influence on sarcasm detection complexity. Twitter, with its 280-character limit, forces users to express sarcasm more concisely and structurally, producing linguistic patterns that are easier for models to learn. Conversely, Reddit and YouTube, which allow longer and more informal text, present greater challenges because sarcasm can be spread across more complex narrative contexts. This study's macro F1-score (0.6422 for IndoNLU) is at a competitive level with IdSarcasm results on Reddit (0.6100-0.6274), showing that the fine-tuning strategy with under-sampling effectively handles YouTube comment complexity despite a smaller dataset.

Metric focus differences also need attention. The IdSarcasm study reported macro F1-scores giving equal weight to both classes, while this study gave special attention to sarcasm class performance because that class is the main detection target. This study's sarcasm class F1-scores (0.6499-0.6568) with high recall (0.6652-0.7639) show the model's ability to detect the majority of sarcastic comments, a critical aspect for practical applications like content moderation.

Zero-shot evaluation in this study is consistent with Suhartono et al. (2024) findings showing the limitations of pre-trained models without fine-tuning for sarcasm detection. This study quantifies improvement from fine-tuning with dramatic increases up to 303% for IndoNLU (from F1-score 0.1613 to 0.6499) and 87% for IndoLEM (from 0.3519 to 0.6568), providing additional empirical evidence about the importance of task-specific training that applies consistently across various Indonesian social media platforms.

*name of corresponding author

This comparison confirms that sarcasm detection is highly dependent on platform characteristics and data domain. Models trained on Twitter cannot be directly generalized to Reddit or YouTube without re-fine-tuning. This research's results provide specific baselines for Indonesian YouTube comments and show that achieved performance is at a competitive level with previous studies for platforms with similar complexity, while providing additional insights about fine-tuning effectiveness and model trade-offs not explored in previous research.

This research provides several important contributions to Indonesian NLP literature. The main contribution lies in developing a sarcasm detection system specifically for Indonesian YouTube comments, a platform with 143 million users representing a different communication context from Twitter and Reddit previously studied. This study demonstrates the effectiveness of automatic labeling using GPT-4o for sarcasm detection tasks, providing an efficient alternative to manual annotation and producing a new labeled dataset that can be used for further research. Systematic zero-shot baseline evaluation provides empirical evidence about the significant value of fine-tuning, showing performance improvements up to 303% for IndoNLU and 87% for IndoLEM. Comprehensive comparison between two IndoBERT variants reveals trade-offs between recall and precision important for model selection according to specific use cases, where informal corpus provides better initial intuition but formal corpus has greater learning capacity through fine-tuning.

This research has several limitations. The dataset is limited to text comments only without considering video content, metadata, or conversation threads. Class imbalance in the original dataset (29.4% sarcasm) may affect model generalization. Automatic labeling using GPT-4o, although efficient, may introduce systematic bias different from human annotation. The study only focuses on explicit linguistic sarcasm without multimodal elements.

CONCLUSION

This research successfully developed an automatic sarcasm detection system for Indonesian YouTube comments using fine-tuned IndoBERT models. Zero-shot evaluation showed that pre-trained models without fine-tuning have very limited sarcasm detection capability with F1-scores of 0.1613 for IndoNLU and 0.3519 for IndoLEM, confirming the importance of task-specific training. Fine-tuning with under-sampling strategy dramatically improved performance with F1-score increases up to 303% for IndoNLU and 87% for IndoLEM. There is a trade-off between recall and precision where IndoLEM achieves the highest recall but with lower precision, while IndoNLU provides optimal balance. IndoBERT-IndoNLU is more suitable for general applications requiring stability, while IndoLEM is better suited for content moderation prioritizing maximum detection. Automatic labeling using GPT-4o with structured prompts enables efficient dataset creation while maintaining consistency.

This research contributes to Indonesian NLP literature by providing a new labeled dataset, demonstrating the effectiveness of transformer-based models for sarcasm detection in informal social media contexts, and providing empirical evidence about the significant value of fine-tuning through zero-shot baseline comparison. Future research should explore increasing sarcasm class data to improve class distribution, testing models with additional context such as video titles or transcripts, comparison with other architectures such as IndoRoBERTa or multilingual models, and developing real-time web-based systems for YouTube comment moderation.

REFERENCES

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.”

Gole, Montgomery, Williams-Paul Nwadiugwu, and Andriy Miranskyy. 2024. “On Sarcasm Detection with OpenAI GPT-Based Models.” Pp. 1–6 in *2024 34th International Conference on Collaborative Advances in Software and COmputiNg (CASCON)*.

Grice, Herbert Paul. 1975. *Logic and Conversation*. New York: Academic Press.

Islam, Saiful, Mahmuda Ferdusi, and Tanjim Taharat Aurpa. 2025. “Words of War: A Hybrid BERT-CNN Approach for Topic-Wise Sentiment Analysis on The Russia-Ukraine War.” *Expert Systems with Applications* 284. doi:https://doi.org/10.1016/j.eswa.2025.127759.

Jia, Mengzhao, Can Xie, and Liqiang Jing. 2024. “Debiasing Multimodal Sarcasm Detection with Contrastive Learning.” *Proceedings of the AAAI Conference on Artificial Intelligence* 38(16):18354–62. doi:10.1609/aaai.v38i16.29795.

Koto, Fajri, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. “IndoLEM and IndoBERT: A Benchmark Dataset and Pre-Trained Language Model for Indonesian NLP.” Pp. 757–70 in *Proceedings of the 28th International Conference on Computational Linguistics*, edited by D. Scott, N. Bel, and C. Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Ma’aly, Ahmad Nahid, Dita Pramesti, Ariadani Dwi Fathurahman, and Hanif Fakhrurroja. 2024. “Exploring Sentiment Analysis for the Indonesian Presidential Election Through Online Reviews Using Multi-Label Classification with a Deep Learning Algorithm.” *Information* 15(11):705. doi:10.3390/info15110705.

*name of corresponding author

Mandhasiya, Dwi Guna, Hendri Murfi, and Alhadi Bustamam. 2024. "The Hybrid of BERT and Deep Learning Models for Indonesian Sentiment Analysis." *Indonesian Journal of Electrical Engineering and Computer Science* 33(1):591. doi:10.11591/ijeecs.v33.i1.pp591-602.

Qin, Zhenkai, Qining Luo, Zhidong Zang, and Hongpeng Fu. 2025. "Detecting Sarcasm in User-Generated Content Integrating Transformers and Gated Graph Neural Networks." *PeerJ Computer Science* 11:e2817. doi:10.7717/peerj-cs.2817.

Ranti, Kiefer Stefano, and Abba Suganda Girsang. 2020. "Indonesian Sarcasm Detection Using Convolutional Neural Network." *International Journal of Emerging Trends in Engineering Research* 8(9):4952–55. doi:10.30534/ijeter/2020/10892020.

Razali, Md Saifullah, Alfian Abdul Halin, Lei Ye, Shyamala Doraisamy, and Noris Mohd Norowi. 2021. "Sarcasm Detection Using Deep Learning With Contextual Features." *IEEE Access* 9:68609–18. doi:10.1109/ACCESS.2021.3076789.

Sharma, Dilip Kumar, Bhuvanesh Singh, Saurabh Agarwal, Hyunsung Kim, and Raj Sharma. 2022. "Sarcasm Detection over Social Media Platforms Using Hybrid Auto-Encoder-Based Model." *Electronics* 11(18):2844. doi:10.3390/electronics11182844.

Suhartono, Derwin, Wilson Wongso, and Alif Tri Handoyo. 2024. "IdSarcasm: Benchmarking and Evaluating Language Models for Indonesian Sarcasm Detection." *IEEE Access* 12:87323–32. doi:10.1109/ACCESS.2024.3416955.

We Are Social, and Meltwater. 2025. *Digital 2025 Global Overview Report*. *Research Report*. 2. London: We Are Social. https://wearesocial.com/wp-content/uploads/2025/02/GDR-2025-v2.pdf.

Wilie, Bryan, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding."

*name of corresponding author