VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8222

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

SARCASM DETECTION IN PALEMBANG LANGUAGE USING HYBRID CLASSICAL AND DEEP LEARNING APPROACHES

Johannes Petrus[1*]; Antonius Wahyu Sudrajat[1]; Muhammad Rachmadi[2]

Master of Information System[1]
Information System[2]
Universitas Multi Data Palembang, Palembang, Indonesia[1,2]
www.mdp.ac.id[1,2]
johannes@mdp.ac.id*, wahyu.sudrajat@mdp.ac.id, rachmadi@mdp.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *Research on sarcasm detection in sentences generally focuses on English and the national languages of certain countries, so that regional languages such as Palembang are still underrepresented in NLP research. This study aims to classify sarcasm in Palembang sentences by applying 12 models and evaluating their performance using a dataset containing 1,952 manually annotated samples. The experiment was also conducted using a 5-fold cross-validation scheme, and the statistical significance of the test results was analyzed using the Friedman t-test (χ_F^2= 41.96, FF=12.90). Contrary to common expectations, the Ensemble classifier achieved the best overall performance, attaining a mean F1-score of 0.877 and outperforming larger pre-trained and dialect-specific models. Even though it is still in the same language family, MelayuBERT's performance is very far behind (F1:0.691). Furthermore, the relatively lower performance of IndoBERT-1.5G (0.731) suggests that model scale alone does not guarantee effectiveness in low-resource language settings. These findings highlight the robustness of feature-engineered classical models compared to large-scale or dialect-specific pre-trained models for sarcasm detection in low-resource regional languages. Our research results with datasets sourced from varied domains have exceeded several national benchmarks. This study can be a baseline for further research.*

Keywords: *Cross-Validation, Low-Resource NLP, Palembang, Sarcasm Detection, T-Friedman Test*

Intisari— *Penelitian deteksi sarkasme pada kalimat umumnya berfokus pada bahasa Inggris dan bahasa nasional suatu negara, sehingga bahasa daerah seperti Palembang masih kurang terwakili dalam penelitian NLP. Penelitian ini bertujuan mengelompokkan sarkasme pada kalimat berbahasa Palembang dengan menerapkan 12 model, serta mengevaluasi performanya menggunakan dataset berisi 1.952 sampel yang dianotasi secara manual. Eksperimen juga dilakukan dengan skema validasi silang 5-fold, dan signifikansi statistik hasil pengujian dianalisis menggunakan uji t-Friedman (χ_F^2= 41.96, FF=12.90). Diluar ekspektasi, pengklasifikasi Ensemble menunjukkan kinerja terbaik dengan memperoleh skor F1 sebesar 0,877, serta melampaui model pra-latih dan model spesifik dialek yang berukuran lebih besar. Disisi lain, walaupun MelayuBERT masih dalam satu rumpun bahasa yang sama dengan Palembang, kinerjanya masih jauh tertinggal (F1:0.691). Selain itu, capaian kinerja IndoBERT-1.5G (0,731) mengisyaratkan bahwa besarnya skala model tidak secara otomatis menjamin efektivitas pada konteks bahasa dengan sumber daya terbatas. Temuan ini menegaskan kemampuan model klasik berbasis rekayasa fitur dibandingkan model pra-latih berskala besar atau spesifik dialek dalam tugas deteksi sarkasme pada bahasa daerah dengan sumber daya terbatas. Hasil penelitian kami dengan dataset bersumber dari domain yang bervariasi telah melampaui beberapa tolok ukur nasional. Studi ini dapat menjadi dasar untuk penelitian lebih lanjut.*

Kata Kunci: *Deteksi Sarkasme, Palembang, Pemrosesan Bahasa Alami Sumber Daya Rendah, Uji t-Friedman, Validasi Silang.*

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8222

INTRODUCTION

Indonesia is the largest archipelagic country in the world with 17,380 islands, has extraordinary linguistic richness, reflected in 718 regional languages which makes it the country with the largest number of regional languages in the world [1]. Although there are many regional languages, Indonesian serves as the national language that functions as a means of communication between citizens from different regions. One of these linguistic treasures is the Palembang language, which is widely used in South Sumatra Province, especially in its capital city, Palembang. This regional language that rich in philosophy and has a long history dating back to the Sriwijaya Kingdom, is used in everyday conversation and is widely understood, even when used sarcastically. The Palembang language, also known as Palembang Malay (*Baso Pelembang*), is a variant of Malay [2] spoken in the Greater Palembang area in South Sumatra Province. This language has two main linguistic levels, namely Bebaso or Palembang alus language and Baso Palembang sari-sari. Palembang alus is a more formal and refined Palembang language, while Baso Palembang sari-sari is a more informal form [3] but is the most widely used in everyday conversation. On the other hand, there is also the Malay language used by Malaysian citizens.

The Palembang Malay language and the Malaysian Malay language originate from the same language family. Sarcasm is defined as a form of linguistic expression that conveys a meaning opposite to what is literally written, spoken, or implied [4],[5],[6]. This expression is often accompanied by emotional tones such as anger, disappointment, ridicule, sarcasm, or even humor [7]. In today's digital age, sarcasm has become a common phenomenon in text-based communication on various social media platforms such as X (formerly Twitter), Instagram, and YouTube comment sections. Sarcastic statements are often difficult to understand because their literal meaning is far from their original intent, and they are often not explicitly expressed in the sentence. Detecting sarcasm in text presents significant challenges [8] compared to spoken communication. The absence of nonverbal cues such as body language, facial expressions, or tone of voice makes it difficult for text-based systems to identify the true meaning of a sarcastic statement. An inability to understand sarcasm can lead to miscommunication and even potential conflict. Conversely, the ability to grasp the implied meaning of sarcasm allows for a more thoughtful and sensitive response to a situation.

Sarcasm introduces noise into textual data [9], and the inability of automated systems to detect it can lead to serious misinterpretations, ultimately impairing the functionality of text-based applications. This underscores the importance of developing systems capable of detecting sarcasm. The main purpose of sarcasm detection is to determine whether a sentence contains sarcasm or not. However, this task is very challenging due to the subjective nature of sarcasm which varies depending on the topic, geographical region, user mentality, and other factors [10]. Therefore, sarcasm detection has become a rapidly growing field of research in NLP, especially for English or other languages with abundant linguistic resources such as by [11], [12], [13] and [14]. Research on detecting sarcasm in languages other than English or known as low-resource languages, has also been conducted, such as in Arabic, Hindi [15], Malaysian [10], Nigerian [16], Persian also Tamil and Malaysian language [17], including Indonesian [18],[19] and [20] which is still very limited.

Most research still focuses on formal Indonesian, ignoring the pragmatic complexity of Indonesia's more than 700 regional languages. Consequently, general pre-trained models are less sensitive to cultural metaphors and local intensifiers, such as those in the Palembang dialect, which degrades performance. To date, there has been no comprehensive study of sarcasm detection in Indonesian regional dialects. This research fills this gap by developing a sarcasm detection framework specifically for the Palembang dialect. Sarcasm detection in low-resource languages (including Palembang) faces two major challenges: the scarcity of labeled corpora and the high dependence on cultural context. This paper evaluates various algorithms, from lexical feature-based methods to Transformer models with deep contextual representations. Although Indonesia has 718 regional languages, NLP efforts are still largely concentrated on Indonesian. Regional languages such as Javanese, Sundanese, Batak, or Minangkabau, despite having millions of speakers, are significantly underrepresented in the NLP literature compared to Indonesian. The main challenge is the limited availability of linguistic resources, including the scarcity of annotated corpora [21].

Although there have been attempts to conduct research on regional languages in Indonesia for classification purposes, the use of Palembang language datasets is still lacking, especially in the context of sarcasm detection and to the best of author's knowledge, there has been no research that specifically discusses sarcasm

detection in Palembang language. This research serves as an initial benchmarking for this dialect, which is an essential step in language research with limited resources. Sarcasm detection in resource-limited languages (such as the Palembang dialect) is an important research challenge, because conventional NLP methods—which generally work well in English—are often ineffective in such contexts. The Palembang language offers a unique and untouched linguistic landscape in the context of sarcasm detection. This research gap is the main justification and fundamental novelty of this study. Indonesia, with more than 700 regional languages, faces significant challenges in NLP research due to limited data sources and annotated data. Therefore, this study not only fills a scientific gap in the literature on sarcasm detection, but also has great significance in expanding the horizons of NLP to less resourceful regional languages.

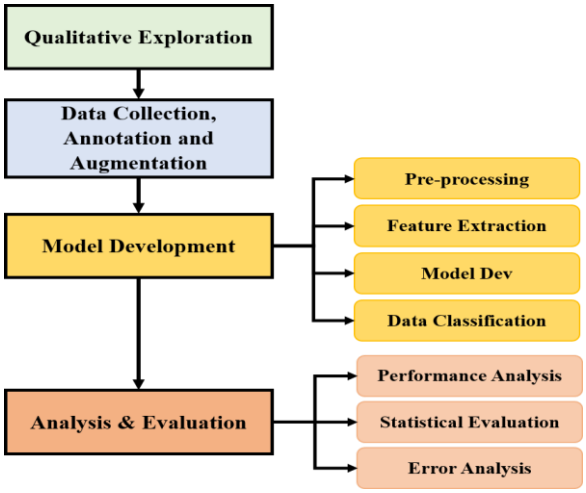
This study presents a qualitative error analysis to bridge quantitative metrics and linguistic nuances. By highlighting the model's difficulties in recognizing Palembang's signature sarcasm—such as implicit metaphors and pragmatic shifts—this study reveals the unique challenges of regional language NLP beyond statistical performance. This study does not include an analysis of multimodal sarcasm. This study uses both classical and modern machine learning methods, including models specifically designed for Indonesian and Malay, as well as multilingual models. These models are grouped into three approaches. The first approach includes classical machine learning models, namely LR, SVM, RF, and NB, trained using TF-IDF representations. The second approach applies deep learning methods, including Bi-LSTM and Ensemble architectures, as well as six pre-trained transformer models for Indonesian and Malay: IndoBERT and its variants, mBERT, IndoRoBERTa, and MelayuBERT. All models are equipped with linguistic features.

The dataset was trained using this twelve model groups and evaluated using standard metrics, namely Accuracy, Precision, Recall, and F1-score. The results of single-split training show that the Ensemble is superior with the highest F1-score of 0.877. Training a model with a single-split pattern can introduce bias in testing, or some useful samples may not be included in the training set. Therefore, cross-validation is applied. Cross-validation provides statistically more reliable performance estimates than a single training/test set split. In this study, model validation used 5-fold Cross-Validation to assess model performance, and Ensemble emerged as the best model with an average F1 score of 0.896.

This study aims to classify sentences in the Palembang language into two groups: sarcastic and non-sarcastic. Theoretically, this study contributes to the development of a very rare, if not non-existent, approach to classifying sarcastic sentences in the Palembang language. Practically, the results of this study will be useful in designing a monitoring system and translating sarcastic sentences in the Palembang language that are harsh or hurtful into more humane ones. By presenting an annotated dataset and evaluating the baseline method, this research builds an important benchmark as a reference for the development of language technology research in various dialects in Nusantara (Indonesia).

MATERIALS AND METHODS

The proposed workflow consists of four main stages, namely Qualitative Exploration, Data Collection and Data Annotation, Model Development, and finally Analysis and Evaluation, as shown in Figure 1.



Source: (Research Results, 2026)

Fig. 1 Propose Workflow Chart

1. Qualitative Exploration

Conducting preliminary corpus analysis of authentic sources, both directly from native community conversations and from social media (X/Twitter, Instagram, Facebook, etc.) that use the Palembang language, to identify the characteristics of sarcastic sentences.

2. Data Collection, Annotation And Augmentation

Data collection is the first step before data processing activities can be carried out [22]. Compiling a sarcasm dataset is a challenging

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8222

process because sarcastic language is highly context-dependent, making sarcastic statements difficult to recognize and annotate without considering the context in which they are used [10]. Unlike general sarcasm research, which relies primarily on Twitter for its data, this study explores cross-domain sarcasm in the Palembang dialect to ensure linguistic representativeness and mitigate domain bias. Data sources include transcriptions of everyday spoken conversations, social media with hashtags related to Palembang, such as #BahasoPlembang and #BahasaPalembang, and comments on Instagram accounts such as plg_yedak and palembangbedebis.id. It also includes vocabulary collected from wikikamus (https://id.wiktionary.org/wiki/Kategori:Kata_bahasa_Palembang). This multi-domain approach allows the model to learn sarcasm patterns across various levels of formality and writing styles. However, we acknowledge some challenges, such as balancing the proportions of these data sources, which can affect the model's sensitivity to sarcasm cues across domains. Sarcasm in one source (e.g., on a wiki) can be more subtle or less frequent than in face-to-face conversations. This can make it challenging for the model to recognize more formal sarcasm.

The first data collection consisted of 1,220 Palembang language sentences, which were manually annotated using binary labels (1 = sarcastic and 0 = not sarcastic). The dataset size of 1,220 was deemed insufficient, therefore, to address this limitations while maintaining sarcastic intent, we implemented a context-aware augmentation technique that increased our dataset by 60% (from 1,220 to 1,952 samples). Our approach focuses on three transformations for sarcastic utterances: (1) punctuation strengthening (e.g., '!' → '!!!'), (2) insertion of dialect-specific intensifiers (*'nian', 'banget'*), and (3) contextual synonym replacement of sarcastic keywords. Most importantly, we avoided machine translation-based methods that risk distorting pragmatic markers crucial for sarcasm detection in Palembang language. The statistics of the dataset are shown in Table 1.

Table 1. Data Collection Statistic

Data Class	# of Sentences	Percentage
Sarcasm	1,039	53.23
Non-sarcasm	913	46.77
	1.952	100.00

Source: (Research Results, 2026)

Source: (Research Results, 2026)

The dataset annotation was conducted through a rigorous manual process involving three native Palembang speakers to ensure pragmatic accuracy. Each annotator independently labeled samples as "Sarcastic" or "Non-Sarcastic" based on guidelines that included linguistic cues such as tonal emphasis, regional intensifiers (examples: "*nian*," "*cak*," or "*awak ni*"), and contextual contradictions. If labeling differences arise, a consensus-based approach is used through in-depth discussions until a final agreement is reached. This mechanism is prioritized over automated metrics to capture the subtle cultural nuances of the Palembang dialect that are often overlooked by quantitative evaluations.

Several challenges emerged during the annotation process, such as: first, pragmatic ambiguity, as many expressions in the Palembang dialect rely heavily on unwritten social context. For example, the particle "*la*" can neutrally denote a completed action or a sarcastic, cynical rejection, requiring annotators to interpret intent without the aid of visual or audio context. Second, orthographic inconsistencies, given the highly informal and non-standard nature of social media data, require manual normalization and interpretation of spelling variations in dialect terms before labeling. Third, implicit sentiment, where forms of "*soft sarcasm*" are difficult to detect because they do not contain offensive words and are often hidden within positive metaphorical expressions typical of Palembang culture.

3. Model Development

This study developed twelve models consisting of 3 groups, namely classical machine learning, deep-learning and transformers models. All transformer models were fine-tuned for five epochs using the AdamW optimizer, adaptive batch size (2–16), mixed precision (fp16), and gradient accumulation to improve memory efficiency. In the deep-learning group, an ensemble classifier based on the average probability was also developed, with the decision threshold determined by optimizing the F1 score on the validation set.

As a pioneering study in Palembang dialect sarcasm detection, testing through 12 models aims to map the capabilities of NLP in low-resource languages through a cross-paradigm evaluation—from lexical methods to transformers—in capturing regional nuances. With this approach, this study not only establishes a comprehensive empirical foundation but also reveals the extent to which the accuracy of regional dialect sarcasm classification is influenced by the linguistic proximity of the pre-training corpus.

Several steps in this phase consist of:

a. Preprocessing Data:

This stage is treated as a crucial data cleaning process, adjusting for the orthographic inconsistencies typical of the Palembang dialect. Unlike formal Indonesian, some texts sourced from regional social media tend to be very noisy and non-standard. At this stage, text normalization, case folding, and text tokenization are performed. Normalization aims to change non-standard linguistic variations in the Palembang language used in everyday life [23], by applying two schemes: (i) lexical normalization to change non-standard local words into standard forms (e.g., *awak* → *kamu* (you), *idak* → *tidak* (no)), and (ii) expressive normalization to standardize emotive patterns such as laughter (e.g., wkwk, hahaha → *tertawa* (laughing)). This two schemes are crucial steps to suppress Out-of-Vocabulary (OOV) levels and reduce model noise due to spelling variations. By transforming unstructured input into more standardized input, the model can more consistently capture the emotional and pragmatic cues needed to detect sarcasm in low-resource settings. This process is designed to reduce interference, ensure consistency in tokenization, and retain the pragmatic cues that are essential for detecting sarcasm. Next, the entire text is converted to lowercase, and tokenization is performed by dividing sentences into tokens consisting of single words [22].

b. Feature Extraction

This stage was designed as a hybrid representation that combines statistical and knowledge-based approaches to address data limitations. Statistical representation using TF-IDF (a vocabulary size of 5,000, n-grams 1–2) was used to capture dominant word and phrase patterns in Palembang dialect. Simultaneously, knowledge-based features were integrated through seven manual linguistic features designed to represent sarcasm characteristics, such as contextual irony phrase indicators, irony intensity measures (frequency of sarcastic phrases and punctuation marks), proportion of capital letters as markers of emphasis, and sentence length. This approach provide explicit domain knowledge to offset the limitations of model learning mechanisms on small datasets. All featured were combined as input to all models to capture pragmatic nuances not visible in raw text.

c. Model Development

Involving twelve classifiers grouped into three approaches: (i) classical machine learning (LR, SVM, RF, NB), (ii) sequential deep learning based on Bi-LSTM and Ensemble (combination of models), and (iii) pre-trained transformer models for Indonesian and Malay, including variants of IndoBERT, IndoRoBERTa, mBERT, and MelayuBERT. Experiments were conducted using Python 3.10 with scikit-learn 1.4.0, Transformers 4.44.2, and PyTorch 2.1.0.

d. Data Classification

Assigning a specific label (sarcastic or non-sarcastic) to a sentence based on the patterns the model has learned from previous data.

4. Analysis and Evaluation:

a. Performance Analysis

The performance evaluation of classification models is based on a confusion matrix [24] that describes the relationship between actual labels and model predictions, with four components: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Based on these components, the key performance metrics are calculated as follows: accuracy, precision, recall (sensitivity), and F1-score, each of which is formulated as a function of TP, FP, TN, and FN to measure prediction accuracy, the ability to detect sarcasm, and the balance between precision and recall.

b. Statistical Evaluation:

Model evaluation was performed using the McNemar test and the t-Friedman test. The McNemar test is used to compare two classifiers and assess performance differences on the same instances [25][26], while the t-Friedman test is used to compare the performance of twelve classification models based on the F1-score values generated from 5-fold cross-validation [27].

The McNemar test is essentially a form of the paired chi-square test, which requires the calculation of the χ^2 value with continuity correction. The t-Friedman test is essentially a modified version of the Friedman test [27] which is applied to assess whether there is a significant difference in performance, such as F1-score, between the models as a whole. The t-Friedman test ranks models based on cross-validation results. The k-fold cross-validation technique divides the data into k subsets (folds), and the model is tested k times with one subset as test data and the rest as training data. The t-Friedman test was used to detect whether there were significant differences in performance among all tested models. To identify which pairs of models differed significantly, a post-hoc analysis was performed using the paired Wilcoxon signed-rank test, accompanied by the

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8222

Holm correction.

RESULTS AND DISCUSSION

The dataset in this study consists of 1,952 annotated data samples, with a two-class distribution: sarcasm and non-sarcasm. Evaluation was performed using standard metrics such as accuracy, F1-score, precision, and recall, calculated on the test dataset.

1. Experimental Setup

Experiments were conducted with general parameter data adjusted to each model. All transformer models were trained using the AdamW optimizer (lr=2e-5, weight decay=0.01) with gradient accumulation 1-8 to achieve an effective batch size of 2-16 for 5 epochs and random seed 42. Early stopping was applied based on the F1-score performance on the validation set. Each model uses its own tokenizer and threshold optimum such as Indobenchmark/indobert-base-p2 for IndoBERT-p2 (0.58), Google-bert/bert-base-multilingual-cased for mBERT (0.63), Cahya/bert-base-indonesian-522M for IndoBERT-522M (0.55), Cahya/Roberta-base-indonesia-522M for IndoRoBERTa (0.52), Cahya/bert-base-indonesian-1.5G for IndoBERT-1.5G (0.58) and stevenLimcorn/MelayuBERT for MelayuBERT (0.62).

The dataset divided into 70% or 1,366 sentences for training data, 15% or 293 sentences for test data, and 15% or 293 sentences for validation data, as shown in Table 2.

Table 2 Dataset composition

Sentences	Actual		Data Train (70%)	Test (15%)	Valid (15%)
Sarcastic	1,039	53%	727	156	156
Non-Sarcastic	913	47%	639	137	137
	1,952	100%	1,366	293	293

Source: (Research Results, 2026)

Source: (Research Results, 2026)

The available dataset was then trained using 12 different classification models. This training aimed to evaluate each model's ability to recognize sarcasm patterns, thus identifying the model or combination of models with the best performance. After the model was trained, the classification of sarcastic sentences in the Palembang language was carried out through a standard inference flow, which combined probabilistic output, threshold optimization, and ensemble consensus. In the probabilistic inference stage, each model processes the input text and generates probability scores for each class: sarcastic and non-sarcastic. These probabilities scores indicate the model's confidence level in predicting sarcastic sentences. For classical machine learning models such as LR, SVM, NB and RF, probability scores estimates are obtained directly through built-in probability prediction functions (predict_proba). In contrast, neural network-based models, including Bi-LSTM and Transformer architectures, generate probabilities scores derived from softmax over logits.

The next step involves threshold optimization. We did not use the default threshold of 0.5, but implemented a data-driven approach to determine the optimal decision threshold. This threshold was selected by maximizing the F1 score on the validation dataset. A sentence will be classified as sarcastic only if its predicted probability exceeds the optimal threshold. Next, ensemble classification is performed to improve prediction reliability. The final classification decision is generated based on a probability-based voting mechanism, where a sentence is classified as sarcastic if the ensemble probability exceeds an optimized global threshold. The models were evaluated using statistical methods. To ensure balanced statistical comparisons between models, evaluations were also conducted using a 5-fold stratified cross-validation scheme.

2. Model Performance

As mentioned, this study evaluated twelve models. Each model was trained and validated using the same dataset. The evaluation was carried out using standard metrics. Experiments were conducted to assess how well the model performs in detecting sarcastic sentences, with the results as listed in Table 3.

Table 3. Training outcome performance

Model	Acc	F1	Prec	Rec	TP	FP	TN	FN
Ensemble	0.860	0.877	0.834	0.924	146	29	106	12
Random Forest	0.857	0.874	0.830	0.924	146	30	105	12
IndoBERT-522M	0.846	0.860	0.847	0.873	138	25	110	20
Bi-LSTM	0.833	0.851	0.823	0.880	139	30	105	19
mBERT	0.829	0.838	0.860	0.817	129	21	114	29
IndoBERT-p2	0.826	0.837	0.845	0.829	131	24	111	27
IndoRoBERTa	0.816	0.833	0.813	0.854	135	31	104	23
LR	0.795	0.827	0.761	0.905	143	45	90	15
SVM	0.812	0.822	0.841	0.804	127	24	111	31
Naive Bayes	0.782	0.818	0.742	0.911	144	50	85	14
IndoBERT-1.5G	0.659	0.731	0.636	0.861	136	78	57	22
MelayuBERT	0.628	0.691	0.626	0.772	122	73	63	36

Source: (Research Results, 2026)

3. Best Performing Model: Ensemble

The Ensemble and Random Forest models dominated the performance with F1 scores of 0.877 and 0.874, respectively, demonstrating the robustness of the hybrid and classical approaches on the limited Palembang dialect sarcasm dataset. The Ensemble model emerged as the best architecture with an F1 score of 0.877 and an accuracy of 0.860. These results indicate that the strategy of combining predictions from multiple models optimally balances precision (0.834) and recall (0.924), making it stable in minimizing prediction errors in both sarcasm and non-sarcasm classes. This is one of the main advantages of Ensemble, making it the best candidate for the task of sarcasm detection in the Palembang language context.

In second place, Random Forest demonstrated highly competitive performance with an F1-score of 0.874 and the highest recall (0.924). This high recall indicates that the classical TF-IDF feature-based model is highly effective in capturing explicit lexical signals of sarcasm in the Palembang regional language. The Transformer-based model group demonstrated fairly solid performance, although it wasn't yet the best. IndoBERT-522M achieved the best result in this group with an F1-score of 0.860, followed by IndoBERT-p2 (0.837) and IndoRoBERTa (0.833). Meanwhile, mBERT (0.838) and MelayuBERT (0.691) trailed behind, indicating that the multilingual or Malay-based models are still less than optimal in handling the nuances of Palembang's sarcasm. However, IndoBERT-1.5G exhibited an anomaly with an F1-score of 0.731 and very low precision (0.636). This model suffered from bias, classifying almost all data as sarcasm (0.861 recall, 78 FP), confirming that large-scale models are susceptible to overfitting on dialect-specific datasets.

Classical models demonstrate consistent performance with an emphasis on high recall. Naive Bayes (F1: 0.818) and Logistic Regression (LR) (F1: 0.827) are able to detect almost all sarcastic sentences (recall > 0.90), although their precision tends to be lower. Meanwhile, SVM (F1: 0.822) shows a better balance between the two. This finding indicates that word frequency-based features (TF-IDF) are quite reliable in capturing keywords of sarcasm in dialects, such as local intensifiers (*"nian", "cak"*). Bi-LSTM recorded an F1-score of 0.851 with a good performance balance, indicating that sequential architectures utilizing manual linguistic features remain competitive. Classical models tend to produce high False Positive rates, often mistakenly classifying neutral sentences as sarcasm. Naive Bayes (FP=50, FN=14) recorded the highest number of FPs, which indicates that this model is too sensitive to intensifier words in dialect, so it often considers informative sentences as a form of sarcasm.

Transformer models tend to produce high FN, indicating that they often fail to detect subtle or implicit sarcasm. For example, MelayuBERT (FP=73, FN=36) indicates an inability to capture the nuances of the Palembang dialect, while mBERT (FP=21, FN=29) indicates that the multilingual model is less sensitive to local cultural context. MelayuBERT performed the worst with an F1 score of 0.691, FP of 73, and FN of 36, indicating that the multilingual model struggled to capture the sarcasm of the Palembang dialect, confirming the existence of a 'Linguistic Proximity Gap' despite the Palembang dialect being related to Malay. With very high FP, the model's precision drops to its lowest level, showing the most aggressive tendency to misclassify neutral sentences as sarcasm. This model also tends to over-generalize across cultural contexts (Palembang), reflected in FN=36. A recall of only 0.772 indicates its inability to detect satire without explicit profanity.

The best models exhibited an optimal performance balance. Ensemble (FP=29, FN=12) recorded the lowest total errors (41), with an FP:FN ratio of approximately 2.4:1, reflecting excellent balance. This approach was able to suppress the FP of the classical model while reducing the FN of the Transformer model through a majority voting mechanism. Random Forest (FP=30, FN=12) had an FP:FN ratio of 2.5:1 and proved effective in detecting explicit sarcasm. Meanwhile, IndoBERT-522M (FP=25, FN=20) showed an FP:FN ratio of 1.25:1, making it a local Transformer model with an ideal FP and FN balance, indicating that the Indonesian corpus-based model is more attuned to the Palembang dialect than the multilingual model.

In a global context, sarcasm detection shows that its performance varies greatly depending on language features and dataset resource availability. In Indonesian studies, research by [18] recorded the highest F1-score of 0.769 on Twitter data, while other studies [20] showed lower figures, around 0.540-0.610. Our research findings on the Palembang dialect (F1-score of 0.877) exceed the national benchmark, indicating that the use of local linguistic features and augmentation techniques is highly effective in capturing the more specific nuances of regional dialects compared to standard formal language models.

Compared to other languages, the performance of our hybrid model is comparable to the results on English ([11], [12]) which are in the range of 0.72–0.89. An interesting phenomenon has also been observed in research on other resource-limited

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

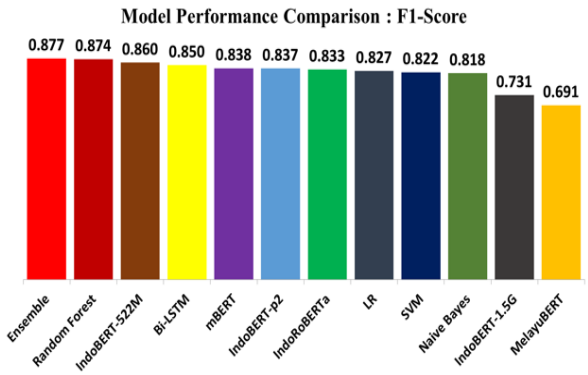
VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8222

languages; similar to the findings in Nigerian Pidgin [16] and Hindi [15], the application of ensemble models and manual feature incorporation has been shown to provide higher stability. Specifically, the success of our Random Forest model aligns with the results in Pidgin, which also demonstrate the robustness of classical models in data-limited scenarios. This confirms the contribution of this research in offering a robust methodological solution for the NLP community focused on resource-limited regional languages.

Increasing the data size to 1,952 samples improved model stability. This is evident in the performance of Random Forest (F1: 0.874), which outperformed several other Transformer models, such as IndoRoBERTa and MelayuBERT. This finding suggests that after augmentation, keyword features in the Palembang language became more consistent, allowing even conventional algorithms to perform reasonably well.

4. Performance Evaluation

In evaluating model performance, we chose the F1-score as the primary metric for discussion, given that the F1-score provides a more objective or balanced assessment of model performance [28]. As a harmonic mean between precision and recall, the F1-score takes into account both False Positives and False Negatives, thereby better reflecting the model's ability to predict sarcasm classes accurately and completely without being affected by class imbalance. Comparison of model performance based on F1-score values as shown in Figure 2.



Source: (Research Results, 2026)

Fig. 1. Model Performance - F1 score

Based on the model performance data, the Deep Learning group achieved the highest average F1-score of 0.864 based on the two models tested, indicating superior performance compared to other approaches. The classic models group came next with an average F1-score of 0.835 from four models, reflecting relatively stable and balanced performance. Meanwhile, the Transformer group obtained an average F1-score of 0.798 from the six models evaluated.

a. McNemar Test

The McNemar test is used to test whether there is a significant difference between two binary classification models on the same dataset [25]. McNemar focuses more on classification errors (whether one model classifies as positive and another model classifies as negative, and vice versa). In this study, the McNemar test was conducted between the IndoBERT-p2 and other models using formula 1. With a significance level of 0.05 and a degree of freedom (df)=1, any $p < 0.05$ indicates a significant difference between the two models. The test results are presented in Table 4.

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} \quad (1)$$

Where b indicates the number of samples misclassified by model #1 but classified correctly by model #2, and c indicates the number of samples misclassified by model #2 but not by model #1. Table 4 indicates that of the 11 models compared, there are three models, namely LR (χ^2=4.17, p=0.0412). , IndoBERT-1.5G (χ^2=25.3, p<0.0001). and MelayuBERT (χ^2=35, p=<0.0001), showed a significant difference compared to IndoBERT-p2.

Table 4. McNemar test results between indoBERT-p2 and other models

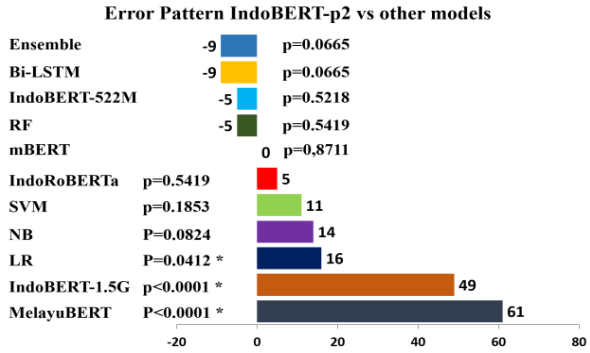
Model	a	b	c	d	|b-c|	χ2	p-value	Sig?
LR	207	35	19	32	16	4.17	0.0412	Yes
SVM	208	34	23	28	11	1.75	0.1853	No
RF	223	19	24	27	5	0.37	0.5419	No
NB	207	35	21	30	14	3.02	0.0824	No
Bi-LSTM	237	5	14	37	9	3.37	0.0665	No
mBERT	223	19	19	32	0	0.03	0.8711	No
IndoBERT-522M	225	17	22	29	5	0.41	0.5218	No
IndoRoBERTa	218	24	19	32	5	0.37	0.5419	No
IndoBERT-1.5G	172	70	21	30	49	25.3	<0.0001	Yes
MelayuBERT	160	82	21	30	61	35	<0.0001	Yes
Ensemble	237	5	14	37	9	3.37	0.0665	No

Source: (Research Results, 2026)

MelayuBERT ($\chi^2 = 35.0$; $p < 0.0001$) showed the largest error deviation (|b-c| = 61), confirming that its classification pattern was significantly worse at capturing the nuances of the Palembang dialect. IndoBERT-1.5G ($\chi^2 = 25.3$; $p < 0.0001$) as the largest

model actually shows significant inconsistency, with column b = 70 vs c = 21, indicating that the reference model is more effective in correcting the errors of the 1.5G variant. Logistic Regression / LR (χ^2 = 4.17; p = 0.0412) shows a different error pattern although it is slight, indicating the limitations of the linear model compared to understanding the Transformer context. The majority of other models, including Ensemble, RF, and IndoBERT-522M, have p-values > 0.05, indicating their error patterns are statistically equivalent to IndoBERT-p2. mBERT has the lowest χ^2 (0.03) and a p-value of 0.8711, indicating its predictions are almost identical to IndoBERT-p2 on this dataset. Ensemble and Bi-LSTM have very low b-values (5), indicating that both models are rarely wrong on data that the IndoBERT-p2 correctly answers.

The McNemar test confirms that the local Transformer model (IndoBERT-p2) is more stable than MelayuBERT or the large model (IndoBERT-1.5G). The statistical similarity of RF/NB to Transformer shows that on limited regional language datasets, the simple model remains statistically competitive. To determine which model is more often correct or incorrect compared to IndoBERT-p2, we can look at the difference between the b and c values. This can be seen in Figure 3. The bar moving to the right indicates that the IndoBERT-p2 model makes more errors than other models. Conversely, the bar moving to the left indicates that the IndoBERT-p2 model is more accurate than other models. The * sign next to the p value indicates a significant difference (p < 0.05).



Source: (Research Results, 2026)

Fig. 2. Comparison of Error Pattern of IndoBERT-p2 vs other Models

b. T-Friedman test

Statistical validation is crucial in machine learning, as the reported model performance may not necessarily be statistically significant. Using a single metric is often insufficient because it ignores variation in the data, therefore non-parametric statistical methods are the main approach in comparing models. The t-Friedman test was applied to evaluate whether there were significant differences in performance (in this case using the F1-score value) among all models as a whole, by utilizing the measurement data from 5-fold cross-validation as shown in Table 5. The t-Friedman's statistic was calculated using formula 2.

$$X_F^2 = \frac{12N}{k(k+1)} \sum_{j=1}^{k} \left(\bar{R}_j - \frac{k+1}{2} \right)^2 \quad (2)$$

with distribution F:

$$\mathrm{FF} = \frac{(N-1)X_F^2}{N(k-1)-X_F^2} \quad (3)$$

A post-hoc analysis was performed using the paired Wilcoxon signed-rank test, accompanied by the Holm correction. The number of model pairs (m) is calculated using the formula K(K-1)/2, and the significance threshold (t) is determined using formula α/m.

Table 5. The 5-Fold Cross Validation Of The Performance Of The Entire Model

MODEL	Fold1	Fold2	Fold3	Fold4	Fold5
LR	0.826	0.809	0.810	0.799	0.830
SVM	0.675	0.734	0.675	0.694	0.697
RF	0.901	0.867	0.847	0.884	0.923
NB	0.799	0.803	0.796	0.812	0.805
Bi-LSTM	0.797	0.823	0.772	0.813	0.815
IndoBERT-p2	0.878	0.848	0.868	0.862	0.876
mBERT	0.861	0.813	0.813	0.890	0.652
IndoBERT-522M	0.868	0.836	0.846	0.885	0.873
IndoRoBERTa	0.850	0.879	0.843	0.860	0.918
IndoBERT-1.5G	0.865	0.861	0.836	0.861	0.860
MelayuBERT	0.844	0.868	0.834	0.870	0.698
Ensemble	0.895	0.886	0.856	0.916	0.925

Source: (Research Results, 2026)

The results of the 5-fold cross-validation test show high stability in most models, with Ensemble and Random Forest consistently dominating each fold. The Ensemble model is the most superior and stable, with an F1-score of 0.856–0.925; the highest achievements on Fold 4 (0.916) and Fold 5 (0.925) indicate that the combination of architectures is able to overcome the weaknesses of a single model. Random Forest (RF) nearly matched the highest performance, reaching 0.923 on Fold 5, demonstrating the high efficiency of traditional lexical features on low-resource regional languages.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8222

The Indonesian-based Transformer model is more stable than the multilingual model, with IndoBERT-p2 (0.848–0.878) and IndoBERT-522M (0.836–0.885) showing minimal fluctuation, and IndoRoBERTa reaching 0.918 on Fold 5, indicating good generalization to the Palembang dialect. mBERT and MelayuBERT show a sharp drop at Fold 5 (mBERT 0.652; MelayuBERT 0.698) despite being high at the time, indicating the sensitivity of multilingual models to variations in dialect distribution. IndoBERT-1.5G is more stable (~0.86) but does not surpass the smaller variance, demonstrating that parameter improvements do not always improve accuracy on this dataset. NB and Bi-LSTM show moderate but stable performance (~0.80) with low variance between folds, while SVM is consistently the lowest, never reaching 0.75. The next step is to transform the data into a ranking for each fold, and the results can be seen in Table 6.

Table 6. Model Performance Based on F1-Score (5F-CV), Standard Deviation and Ranking

Model	F1	F2	F3	F4	F5	Mean	Std	Mean Rank	Mean Rank Sq
Ensemble	2	1	2	1	1	0.896	0.024	1.40	1.96
RF	1	4	3	4	2	0.884	0.026	2.80	7.84
IndoBERT-p2	3	6	1	6	4	0.866	0.011	4.00	16.00
IndoBERT-522M	4	7	4	3	5	0.862	0.018	4.60	21.16
IndoRoBERTa	7	2	5	8	3	0.870	0.027	5.00	25.00
IndoBERT-1.5G	5	5	6	7	6	0.857	0.010	5.80	33.64
MelayuBERT	8	3	7	5	10	0.823	0.064	6.60	43.56
mBERT	6	9	8	2	12	0.806	0.082	7.40	54.76
LR	9	10	9	11	7	0.815	0.012	9.20	84.64
Bi-LSTM	11	8	11	9	8	0.804	0.018	9.40	88.36
NB	10	11	10	10	9	0.803	0.005	10.00	100.00
SVM	12	12	12	12	11	0.695	0.022	11.80	139.24
								78.00	616.16

Source: (Research Results, 2026)

Based on Table 6, Ensemble established itself as the best architecture (Mean F1 0.896; Mean Rank 1.40) with consistent rank 1–2 across folds, while Random Forest (Mean Rank 2.80) stood out as the single model, proving the power of traditional lexical features on low-resource dialects. The IndoBERT model family demonstrated exceptionally high stability. IndoBERT-p2 and IndoBERT-1.5G recorded the lowest Standard Deviations (Std) at 0.011 and 0.010, respectively. These models demonstrated their ability to generalize Palembang sarcasm patterns. IndoRoBERTa had a slightly higher average F1 (0.870) but with slightly larger variance. mBERT and MelayuBERT exhibit high instability; mBERT has the highest Std (0.082) with rank jumps of 2→12 between folds, confirming that multilingual models are less consistent in capturing local dialect nuances than Indonesian corpus-based models.

NB and LR remain relevant with high stability (Std 0.005–0.012) and F1 ~0.80, making them strong baselines, while SVM is consistently the weakest (Mean Rank 11.80) due to its kernel function being less effective in separating sarcasm classes. The low standard deviation across folds indicates that the expanded dataset successfully stabilized the training process, providing a solid basis for comparison between models. With parameter data N=5, k=12 and α=0.05, the t-Friedman statistic (χ_F^2) value obtained was 41.985 and the FF value was 12.903. With degrees of freedom df=(k-1)(n-1)=44 and α=0.05, the critical value (two-tailed) $t_{0.025,44}$ is 2.01. Since the FF value (12.903) is greater than the critical value (2.01), this indicates that there is at least one pair of models that has a significant difference in F1-score performance.

To identify which model pairs were significantly different, a post-hoc analysis was performed using the paired Wilcoxon signed-rank test with Holm correction. This approach was chosen because the available data was ordinal (rank) and the number of folds was limited (n=5). The p-value was determined based on the two-sided Wilcoxon signed-rank table. Overall, the number of model pairs (m) is 66 pairs, and a Holm correction is required at α=0.05, thus obtaining an adjusted significance threshold (t) of 0.00076. If the p-value > t, then there is no significant difference between the two models. This report only presents a comparison of all models against the transformer model IndoBERT-p2 (ranked 4.0), the results shown in Table 7. After Holm correction for pairwise comparisons, it was found that no model pairs showed a significant difference in F1-score at α = 0.05. Although the Friedman t-test showed that, of all models, at least one pair of models had a significant difference, this significance did not hold up in the pairwise test after correction. This is likely due to the small number of folds (N = 5) or because the differences were spread across many model pairs.

Table 7. Paired Model P-Value

Model	# of rank		p-value	Different?
	Pos	Neg		
Ensemble	1	14	0,8680556	No
RF	2	13	1,3020833	No
IndoBERT-522M	10	5	3,0381944	No
IndoRoBERTa	6	9	3,90625	No
IndoBERT-1.5G	12	3	1,7361111	No
MelayuBERT	12	3	1,7361111	No
mBERT	13	2	1,3020833	No
LR	15	0	0,4340278	No
Bi-LSTM	15	0	0,4340278	No
NB	15	0	0,4340278	No
SVM	15	0	0,4340278	No

Source: (Research Results, 2026)

These experimental results enhance our methodological understanding of model behavior in low-resource domains. The finding that classical models like Random Forest remain competitive with large-scale Transformers. The evaluation also indicates that overly complex models are prone to overfitting on small datasets, recommending a hybrid feature approach as a more robust strategy for NLP in minority languages.

5. Error Analysis

Error analysis was used to investigate the causes of incorrect predictions. For example, the sentence "*galak apo kamu kalo lokak saro?*" in English "*do you want to be if it's going to be difficult?*" (non-sarcastic) was predicted as sarcastic by LR, RF, NB, IndoBERT-p2, and MelayuBERT (Table 8), likely because the phrase "*lokak saro*" *(facing difficulties / serious matters)* frequently appears in sarcastic contexts in the training data. The models tended to overgeneralize, associating negative words with sarcasm without considering the question structure. Despite the differences in individual predictions, the Ensemble model produced the correct final decision (Label 0), demonstrating the effectiveness of combining models in reducing false positives.

Table 8. The First-Sentence Error Analysis Results

Sample	Incorrect Prediction (FP)	Correct Prediction (TN)	Main features
galak apo kamu kalo lokak saro?	LR, RF, NB, IndoBERT-p2, MelayuBERT	SVM, Bi-LSTM, mBERT, IndoBERT-522M, IndoRoBERTa, IndoBERT-1_5G, MelayuBERT, Ensemble	Informative question sentences that contain negative vocabulary.

Source: (Research Results, 2026)

The sentence "*kamu beduo itu memang pantes cocok la*" ("*You two really are a perfect match*") was labeled sarcastic. Most models (10/12) predicted it incorrectly, except for NB and mBERT (Table 9).

Table 9. The Second-Sentence Error Analysis Results

Sample	Incorrect Prediction (FP)	Correct Prediction (TN)	Main features
kamu beduo itu memang pantes cocok la	LR, RF, SVM, Bi-LSTM, IndoBERT-p2, IndoBERT-522M, IndoRoBERTa, IndoBERT-1_5G, MelayuBERT, Ensemble	NB and mBERT	Compliment as an insult and no explicit negative words.

Source: (Research Results, 2026)

This sentence looks like a compliment but is actually a sarcasm; the positive words "*pantes*" *(proper)* and "*cocok*" *(suitable)* misled the models. IndoBERT varians failed to distinguish between sincere praise and subtle sarcasm, while the particle "*la*" indicated a sarcastic intonation not captured in the text. The success of NB and mBERT was not significant enough to influence Ensemble's decision.

The sentence "*dek ini ndak lamo lagi jadi peman sinetron*" *("this young sibling will soon be a soap opera actor")* is labeled sarcastic, but all models predict non-sarcastic (Table 10). The sarcasm in this sentence is implicit and situational; its meaning—that someone is "*acting*" or being dishonest—can only be understood with cultural/pragmatic knowledge, for example, that "*pemaen sinetron*" ("*opera actor*") is often used metaphorically. Without this context, the Transformer model reads the sentence as ordinary information about someone's career.

Table 10. The third sentence error analysis table

Sample	Incorrect Prediction (FP)	Correct Prediction (TN)	Main features
dek ini ndak lamo lagi jadi pemaen sinetron	LR, RF, SVM, NB, Bi-LSTM, mBERT, IndoBERT-p2, IndoBERT-522M, IndoRoBERTa, IndoBERT-1_5G, MelayuBERT, Ensemble	-	Implicit sarcasm

Source: (Research Results, 2026)

6. Performance Metrics Summary

For a more comprehensive evaluation, the performance of the 12 models was summarized in a systematic comparison to identify models that not only outperformed the single-split model but were also stable in cross-validation. Table 11 shows convergence among the leading models, such as Ensemble, Random Forest and IndoBERT-p2, which have very low variance. This evaluation serves as an important baseline for research on sarcasm detection in regional languages, given the limited digital resources.

The stability classification thresholds are determined using the Standard Deviation (Std) value, where Very Stable (Std < 0.020), Fairly Stable (0.020 ≤ Std < 0.040), Less Stable (0.040 ≤ Std < 0.100), and Unstable (Std ≥ 0.100)

Table 11. Summary of Model Performance (single-split vs 5-fold CV)

Model	Single Split		5-Fold CV			
	F1 score	Rank	Mean F1	Rank	Std	Stability Status
Ensemble	0.877	**1**	0.896	**1.40**	0.024	Fairly St
RF	0.874	**2**	0.884	**2.80**	0.026	Fairly St

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8222

Model	Single Split		5-Fold CV			
	F1 score	Rank	Mean F1	Rank	Std	Stability Status
IndoBERT-p2	0.837	**6**	0.866	**4.00**	0.011	Very St
IndoBERT-522M	0.859	**3**	0.862	**4.60**	0.018	Very St
IndoRoBERTa	0.833	7	0.870	5.00	0.027	Fairly St
IndoBERT-1.5G	0.731	11	0.857	5.80	0.010	Very St
MelayuBERT	0.691	12	0.823	6.60	0.064	Less St
mBERT	0.838	5	0.806	7.40	0.082	Less St
LR	0.827	8	0.815	9.20	0.012	Very St
Bi-LSTM	0.851	4	0.804	9.40	0.018	Very St
NB	0.818	10	0.803	10.00	0.005	Very St
SVM	0.822	9	0.695	11.80	0.022	Fairly St

Source: (Research Results, 2026)

A summary of the single-split and 5-fold cross-validation results shows Ensemble as the most consistent model in CV (mean F1: 0.896; mean rank 1.40; Std 0.024), while also outperforming in single-split (F1: 0.877).

7. Limitations

Although this study encompasses diverse linguistic sources—social media, Wikipedia, and spoken conversation—there are several limitations that warrant consideration. First, the predominance of informal styles in social media may still impact the model's ability to capture sarcasm in formal texts.

Second, lexical marker-based augmentation (e.g., local intensifiers) has the potential to introduce bias, making the model more sensitive to explicit keywords but less effective for implicit sarcasm. Furthermore, the limited dataset size indicates that this study is still an initial baseline and requires further validation on a larger corpus of regional dialects.

CONCLUSION

This study evaluated 12 cross-paradigm models in detecting sarcasm in the Palembang dialect. The Ensemble model excels with an average F1 of 0.896, peaking at 0.925 at Fold 5, demonstrating the power of combining multiple architectures. Random Forest ranks second as a single model (Mean 0.884; Fold 5 = 0.923), confirming that for the Palembang dialect, lexical features in classical models remain highly competitive compared to complex deep learning. Although Ensemble has the highest average score, the highest consistency is found in the local Transformer, with IndoBERT-1.5G being the most stable (Std 0.010), followed closely by IndoBERT-p. mBERT and MelayuBERT exhibit significant instability. mBERT (Mean 0.806; Std 0.082) fluctuates sharply from 0.890 (Fold 4) to 0.652 (Fold 5), indicating the multilingual model's susceptibility to dialect variation. MelayuBERT (Mean 0.823; Std 0.064) has a higher average, but a large Std indicates inconsistent performance across folds. Classical models show consistency: Naive Bayes (Mean 0.803; Std 0.005) is the most stable, Logistic Regression (Mean 0.815; Std 0.012) is stable around 0.80, and Bi-LSTM (Mean 0.804; Std 0.018) offers a balance with intermediate stability. MelayuBERT alone is inadequate for handling regional dialects, emphasizing the need for dialect-specific data and careful feature engineering to achieve reliable sarcasm detection in low-resource languages. Our findings suggest that NLP research on regional languages should prioritize a balance between model architecture and the incorporation of local linguistic knowledge, rather than relying solely on large parameter counts, as illustrated by the underperformance of the 1.5G variant. The baseline from this study is expected to serve as a reference for addressing data limitations and encouraging innovation toward digital inclusivity for other Indonesian languages.

REFERENCE

[1] E. Andina, "Implementasi dan Tantangan Revitalisasi Bahasa Daerah di Provinsi Lampung," *Aspir. J. Masal. Sos.*, vol. 14, no. 1, Jun. 2023, doi: 10.46807/aspirasi.v14i1.3859.

[2] Izzati, W. A. Rais, and H. Yustanto, "The Impact of Cultural Contact Between Javanese and Palembang Malay On the Language System In the City of Palembang," *Ichss*, vol. 2, pp. 1–11, 2022.

[3] H. L. H. S. Warnars, J. Aurellia, and K. Saputra, "Translation Learning Tool for Local Language to Bahasa Indonesia using Knuth-Morris-Pratt Algorithm," *TEM J.*, vol. 10, no. 1, pp. 55–62, 2021, doi: 10.18421/TEM101-07.

[4] N. H. Jeremy, "The Impact of Text Preprocessing in Sarcasm Detection on Indonesian Social Media Contents," *Eng. Math. Comput. Sci. J.*, vol. 7, no. 2, pp. 183–189, 2025, doi: 10.21512/emacsjournal.v7i2.13503.

[5] Z. Wen *et al.*, "Sememe knowledge and auxiliary information enhanced approach for sarcasm detection," *Inf. Process. Manag.*, vol. 59, no. 3, p. 102883, 2022, doi: 10.1016/j.ipm.2022.102883.

[6] A. Alqahtani, A. Alsheddi, and L. Alhenaki, "Text-based Sarcasm Detection on Social

Networks: A Systematic Review," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 3, pp. 313–328, 2023, doi: 10.14569/IJACSA.2023.0140336.

[7] F. B. Kader, N. H. Nujat, T. B. Sogir, M. Kabir, H. Mahmud, and K. Hasan, "Computational Sarcasm Analysis on Social Media: A Systematic Review," Sep. 2022, [Online]. Available: http://arxiv.org/abs/2209.06170

[8] Y. Y. Tan, C. O. Chow, J. Kanesan, J. H. Chuah, and Y. L. Lim, "Sentiment Analysis and Sarcasm Detection using Deep Multi-Task Learning," *Wirel. Pers. Commun.*, vol. 129, no. 3, pp. 2213–2237, 2023, doi: 10.1007/s11277-023-10235-4.

[9] A. C. Băroiu and Ştefan Trăuşan-Matu, "Automatic Sarcasm Detection: Systematic Literature Review," *Inf.*, vol. 13, no. 8, pp. 1–17, 2022, doi: 10.3390/info13080399.

[10] S. H. Suhaimi, N. A. A. Bakar, and N. F. Nurulhuda, "Constructing and Analysing the MalaySarc Dataset: A Resource for Detecting and Understanding Sarcasm in Malay Language," *Proc. World Congr. Electr. Eng. Comput. Syst. Sci.*, pp. 1–10, 2023, doi: 10.11159/cist23.126.

[11] R. Misra and P. Arora, "Sarcasm detection using news headlines dataset," *AI Open*, vol. 4, no. October 2022, pp. 13–18, 2023, doi: 10.1016/j.aiopen.2023.01.001.

[12] M. A. Rosid, M. A. K. Sambada, S. Busono, and F. Muharram, "Ensemble Machine Learning to Detect Sarcasm in English on Twitter Social Media," *J. Informatics Inf. Syst. Softw. Eng. Appl.*, vol. 6, no. 1, pp. 11–20, 2023, doi: 10.20895/inista.v6i1.1073.

[13] D. Šandor and M. Bagić Babac, "Sarcasm detection in online comments using machine learning," *Inf. Discov. Deliv.*, vol. 52, no. 2, pp. 213–226, 2024, doi: 10.1108/IDD-01-2023-0002.

[14] D. K. Sharma, B. Singh, S. Agarwal, N. Pachauri, A. A. Alhussan, and H. A. Abdallah, "Sarcasm Detection over Social Media Platforms Using Hybrid Ensemble Model with Fuzzy Logic," *Electronics*, vol. 12, no. 4, p. 937, Feb. 2023, doi: 10.3390/electronics12040937.

[15] I. A. Ahmad, P. Gatla, and R. K. Mundotiya, "Sarcasm Identification and Classification in Hindi Newspaper Headlines," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 24, no. 4, pp. 1–21, Apr. 2025, doi: 10.1145/3714469.

[16] K. T. Ladoja and R. T. Afape, "Sarcasm Detection in Pidgin Tweets Using Machine Learning Techniques," *Asian J. Res. Comput. Sci.*, vol. 17, no. 5, pp. 212–221, 2024, doi: 10.9734/ajrcos/2024/v17i5450.

[17] B. R. Chakravarthi, "Sarcasm Detection in Tamil and Malayalam YouTube Comments," *Soc. Netw. Anal. Min.*, vol. 15, no. 1, pp. 1–18, 2025, doi: 10.1007/s13278-025-01486-z.

[18] D. Suhartono, W. Wongso, and A. Tri Handoyo, "IdSarcasm: Benchmarking and Evaluating Language Models for Indonesian Sarcasm Detection," *IEEE Access*, vol. 12, no. May, pp. 87323–87332, 2024, doi: 10.1109/ACCESS.2024.3416955.

[19] J. Amalia, D. F. Matondang, G. E. M. Hutajulu, and A. Hasibuan, "Impact Of Sarcasm Detection on Sentiment Analysis Using Bi-LSTM and FastText," *J. Sist. Inf. Bisnis*, vol. 14, no. 4, pp. 353–362, 2024, doi: 10.21456/vol14iss4pp353-362.

[20] R. Kusumastuti, E. Utami, and A. Yaqin, "Detection of Sarcasm Sentences in Indonesian Tweets using SentiStrength," *Proceeding - 6th Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. Appl. Data Sci. Artif. Intell. Technol. Environ. Sustain. ICITISEE 2022*, pp. 93–98, 2022, doi: 10.1109/ICITISEE57756.2022.10057904.

[21] A. F. Aji *et al.*, "One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 1, pp. 7226–7249, 2022, doi: 10.18653/v1/2022.acl-long.500.

[22] J. Petrus, Ermatita, Sukemi, and Erwin, "An adaptable sentence segmentation based on Indonesian rules," *IAES Int. J. Artif. Intell.*, vol. 12, no. 3, pp. 1491–1499, 2023, doi: 10.11591/ijai.v12.i3.pp1491-1499.

[23] A. A. Aliero, B. S. Adebayo, H. O. Aliyu, A. G. Tafida, B. U. Kangiwa, and N. M. Dankolo, "Systematic Review on Text Normalization Techniques and its Approach to Non-Standard Words," *Int. J. Comput. Appl.*, vol. 185, no. 33, pp. 44–55, 2023, doi: 10.5120/ijca2023923106.

[24] G. Zeng, "Invariance Properties and Evaluation Metrics Derived from the Confusion Matrix in Multiclass Classification," *Mathematics*, vol. 13, no. 16, 2025, doi: 10.3390/math13162609.

[25] A. R. Aditama and A. F. Wicaksono, "Classification of customer complaints on social media for e-commerce in Indonesia," *Int. J. Electr. Comput. Eng.*, vol. 15, no. 3, p. 2977, 2025, doi: 10.11591/ijece.v15i3.pp2977-2985.

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8222

[26] M. Martinović, K. Dokic, and D. Pudić, "Comparative Analysis of Machine Learning Models for Predicting Innovation Outcomes: An Applied AI Approach," *Appl. Sci.*, vol. 15, no. 7, pp. 1–44, 2025, doi: 10.3390/app15073636.

[27] J. Liu and Y. Xu, "T-Friedman Test: A New Statistical Test for Multiple Comparison with an Adjustable Conservativeness Measure," *Int. J. Comput. Intell. Syst.*, vol. 15, no. 1, pp. 1–19, 2022, doi: 10.1007/s44196-022-00083-8.

[28] Wella, R. I. Desanti, and Suryasari, "A Comparative Study of Machine Learning Approaches to Megathrust Earthquake Prediction in Subduction Zones," *J. Appl. Data Sci.*, vol. 6, no. 4, pp. 2421–2435, 2025, doi: 10.47738/jads.v6i4.904.