JURNAL JTIK
(JURNAL TEKNOLOGI INFORMASI DAN KOMUNIKASI)
KITA Komunitas Informasi Teknologi Aceh
OPEN ACCESS
E-ISSN: 2580-1643
Volume 11, Nomor 1, Januari 2027 | Pages: 172-183 | DOI: https://doi.org/10.35870/jtik.v11i1.7711

Directional Robustness Asymmetry in Indonesian Sarcasm Detection: A Cross-Platform Evaluation of IndoRoBERTa on X and Reddit

Marcellinus Brendan Hananta [1*], Wiwin Sulistyo [2]

[1*,2] Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Kristen Satya Wacana, Kota Salatiga, Jawa Tengah, Indonesia.

Corresponding Email: 672022301@student.uksw.edu [1*]

Article History: Received: 25 June 2026, Revision: 8 July 2026, Accepted: 14 July 2026, Available Online: 1 January 2027.

Abstract

Sarcasm can reverse sentence polarity, making it one of the hardest problems in natural language processing and a persistent obstacle to reliable sentiment analysis on social media, where models are often deployed on platforms they were not trained on. Research on Indonesian sarcasm has largely stayed within a single domain or tested transfer in only one direction, leaving the cross-domain robustness of the IndoRoBERTa family on the IdSarcasm benchmark unclear. This study measures and compares the cross-domain robustness of IndoRoBERTa-small and IndoRoBERTa-base, with IndoBERT as a baseline, across X (Twitter) and Reddit, using the F1 gap (ΔF1) after cleaning anonymization artifacts, adding a data-size control, and running Stratified 5-Fold cross-validation. The results reveal a one-sided asymmetry: X → Reddit transfer is far more fragile (ΔF1 0.345–0.400) than Reddit → X (0.135–0.240), failing by missing sarcasm (false negatives) rather than over-flagging it (false positives). Practically, training on the context-rich domain (Reddit) and choosing IndoRoBERTa-base give the most reliable cross-platform sarcasm detection, with direction-specific mitigation recommended for robust deployment.

Keyword: Cross-Domain Robustness; IdSarcasm; IndoRoBERTa; Natural Language Processing; Sarcasm Detection.

Abstrak

Sarkasme dapat membalik polaritas kalimat sehingga menjadi salah satu persoalan tersulit dalam pemrosesan bahasa alami sekaligus penghambat analisis sentimen yang andal di media sosial, tempat model kerap diterapkan pada platform yang berbeda dari data pelatihannya. Penelitian sarkasme bahasa Indonesia umumnya masih terbatas pada satu domain atau menguji transfer hanya satu arah, sehingga ketangguhan lintas-domain keluarga IndoRoBERTa pada benchmark IdSarcasm belum jelas. Penelitian ini mengukur dan membandingkan ketangguhan lintas-domain IndoRoBERTa-small dan IndoRoBERTa-base, dengan IndoBERT sebagai baseline, antara X (Twitter) dan Reddit, menggunakan selisih F1 (ΔF1) setelah artefak anonimisasi dibersihkan, kontrol ukuran data ditambahkan, dan validasi silang Stratified 5-Fold dijalankan. Hasilnya menunjukkan asimetri satu arah: transfer X → Reddit jauh lebih rapuh (ΔF1 0,345–0,400) daripada Reddit → X (0,135–0,240), dengan kegagalan berupa melewatkan sarkasme (false negative) alih-alih terlalu banyak menandainya (false positive). Secara praktis, pelatihan pada domain kaya konteks (Reddit) dan pemilihan IndoRoBERTa-base memberikan deteksi sarkasme lintas-platform paling andal, dengan mitigasi spesifik-arah saat penerapan.

Kata Kunci: Ketangguhan Lintas-Domain; IdSarcasm; IndoRoBERTa; Pemrosesan Bahasa Alami; Deteksi Sarkasme.

E-ISSN: 2580-1643

1. Introduction

Social media platforms such as X (formerly Twitter) and Reddit have become large-scale sources of public opinion, making them valuable data for sentiment analysis. However, analyzing sentiment on these platforms faces a major obstacle in the form of sarcasm an expression whose meaning is opposite to its literal meaning, so that it can reverse sentiment polarity and mislead automated systems. This challenge is even greater for Indonesian text, which is categorized as a low-resource language and exhibits platform-specific stylistic characteristics (Suhartono *et al.*, 2024); indeed, sarcasm detection in low-resource languages remains especially challenging because of the limited availability of annotated data and language-spesific tools (Khan *et al.*, 2024). The emergence of Transformer-based pretrained language models has substantially improved performance across many natural language processing (NLP) tasks. Nevertheless, high performance in one domain does not necessarily persist across domains a problem known as cross-domain robustness or generalization (Calderon *et al.*, 2024). When a model trained on one platform is deployed on another, distributional differences in writing style, text length, and contextual cues can degrade performance significantly. Sarcasm is a form of expression that conveys a meaning opposite to its literal meaning, often to mock or criticize. In NLP, sarcasm detection is commonly formulated as a binary text-classification task, namely determining whether a text is sarcastic or not.

The main challenge of sarcasm detection lies in its dependence on context, world knowledge, and implicit cues that are difficult to capture using surface-feature-based models alone. Approaches have evolved from classical feature-based and machine-learning methods such as the context-based LSTM approach by Khotijah *et al.* (2020) toward approaches based on Transformer pretrained language models that capture context more richly. For Indonesian specifically, recent work has explored hybrid deep-learning architectures, such as a multi-head attention CNN–BiGRU model for Indonesian–English code-mixed sarcasm (Alfan Rosid *et al.*, 2024) and hybrid IndoBERT–BiLSTM/BiGRU models for Indonesian sentiment and emotion classification (Ahmadian *et al.*, 2024); comparative studies on large social-media corpora likewise report that Transformer-based models such as BERT outperform classical machine-learning baselines for sarcasm detection (Šandor & Bagić Babac, 2024). Transformer-based pretrained language models are trained on large corpora to learn language representations and are subsequently fine-tuned for specific downstream tasks. BERT (Devlin *et al.*, 2019) introduced the masked language modeling mechanism, while RoBERTa (Liu *et al.*, 2019) refined BERT's pretraining strategy to produce stronger representations. For Indonesian, IndoBERT (Wilie *et al.*, 2020) was trained on the Indo4B corpus and has become a reference model for various Indonesian NLP tasks. IndoRoBERTa is a RoBERTa-type model trained on Indonesian corpora and is available in several size variants, including small and base. Differences in model size (number of parameters) can affect not only accuracy but also the ability to generalize to data outside the training distribution. The benchmark used in this study, IdSarcasm (Suhartono *et al.*, 2024), is designed specifically to evaluate the ability of language models to detect Indonesian sarcasm. It provides two datasets from two different domains X (Twitter) and Reddit each with standardized train, validation, and test splits. The datasets are released openly through Hugging Face by Wilson Wongso (account w11wo), one of the benchmark's authors. The availability of two domains within a single benchmark makes IdSarcasm particularly suitable for cross-domain robustness studies. Robustness refers to a model's ability to maintain performance when faced with data whose distribution differs from the training data, and domain shift is one of the main causes of performance degradation. (Calderon *et al.*, 2024) highlight that domain shift can degrade language-model performance across many testing conditions, while (Khoee *et al.*, 2024), in their survey, describe various techniques for improving domain generalization.

In the context of Indonesian sarcasm detection, (Sabrina *et al.*, 2026) investigated cross-domain transfer from Twitter to Reddit using XLM-RoBERTa, IndoBERT, and Multilingual BERT, with IndoBERT achieving the highest macro-F1 (0.7224) on the Reddit test set; their study showed that the sarcastic class tends to be difficult to classify due to class imbalance. Mulia *et al.* (2025) compared five loss functions (cross entropy, weighted cross entropy, focal loss, hybrid, and context-aware loss) on IndoBERTweet to address class imbalance; focal loss produced the highest macro-F1 (0.953), while context-aware loss yielded the best recall and representation separation, although the differences between methods were not statistically significant. Despite this encouraging progress, existing systems for Indonesian sarcasm detection still exhibit several limitations. First, most studies evaluate models only in an in-domain setting, where the training and testing data come from the same platform (Fitrianto & Editya, 2024; Rahman & Girsang, 2024), so the reported performance does not reflect a model's reliability when deployed on a

E-ISSN: 2580-1643

different platform. Second, the few cross-domain studies that do exist are limited to a single transfer direction—typically Twitter → Reddit—without examining the reverse direction (Sabrina *et al.*, 2026), so the possibility of a directional robustness asymmetry has never been tested. Third, efforts to improve performance have largely focused on handling class imbalance through loss-function engineering (Mulia *et al.*, 2025), whereas the effect of model capacity (size variants within the same model family) on cross-domain robustness remains unexplored. In addition, prior studies rarely control for confounding factors that can bias cross-domain measurements, such as the difference in dataset size between domains and domain-specific anonymization artifacts; as a consequence, a reported performance gap can be misinterpreted as a difference in sarcasm-detection ability when it actually reflects the model's ability to recognize the source domain. Based on these limitations, the research gap addressed by this study is the absence of a fully bidirectional cross-domain robustness evaluation of the IndoRoBERTa family (small vs. base) on the official IdSarcasm benchmark. Accordingly, this study aims to measure and compare the cross-domain robustness of IndoRoBERTa-small and IndoRoBERTa-base, with IndoBERT as a baseline, in order to determine which variant is most resistant to domain shift between X and Reddit. Robustness is operationalized as the difference in F1 score (ΔF1) between in-domain and cross-domain testing in both transfer directions. To ensure that the measured robustness reflects genuine sarcasm-detection ability rather than spurious shortcuts, the study additionally controls for domain-specific anonymization artifacts and for differences in dataset size, and verifies stability through Stratified 5-Fold cross-validation. The main contributions of this paper are threefold. First, it provides the first fully bidirectional cross-domain robustness evaluation of the IndoRoBERTa family on the official IdSarcasm benchmark. Second, it reveals a consistent directional asymmetry in robustness and explains its underlying linguistic mechanism through an error decomposition into false positives and false negatives. Third, it establishes through a data-size control test and artifact cleaning that the observed asymmetry is intrinsic to domain characteristics rather than an artifact of data imbalance or domain leakage.

2. Research Methodology

Research design

Before turning to each component in detail, it helps to see how the study fits together as a whole. The work is best read as a single thread running from an initial survey of the literature through to the conclusions, and Figure 1 traces that thread while marking the two points where we deliberately pause to protect the validity of the measurements: data cleaning and robustness measurement. The starting point was a review of prior work on Indonesian sarcasm detection, which is what brought the research gap into focus the absence of a bidirectional view of cross-domain robustness. With that gap framed, the two IdSarcasm datasets for X and Reddit were obtained from their official source, and a fair amount of effort went into understanding the data before any modeling began: exploring its characteristics and, just as importantly, removing the domain-specific anonymization artifacts that would otherwise let a model take a shortcut by recognizing the platform rather than the sarcasm itself.

Only after the data was trustworthy did the modeling start. Each of the three models IndoBERT together with IndoRoBERTa-small and IndoRoBERTa-base was fine-tuned across the four in-domain and cross-domain scenarios, and robustness was read directly from the resulting F1 gap (ΔF1) between in-domain and cross-domain testing in both directions. Because a single number can mislead, the analysis did not stop there: a data-size control test and a Stratified 5-Fold cross-validation were used to confirm that the pattern was neither an accident of one particular split nor a by-product of the difference in dataset sizes. The final step moved from measurement to interpretation, breaking the cross-domain errors down into false positives and false negatives so that the asymmetry could be explained rather than merely reported. The sub-sections that follow take up each of these stages in turn.

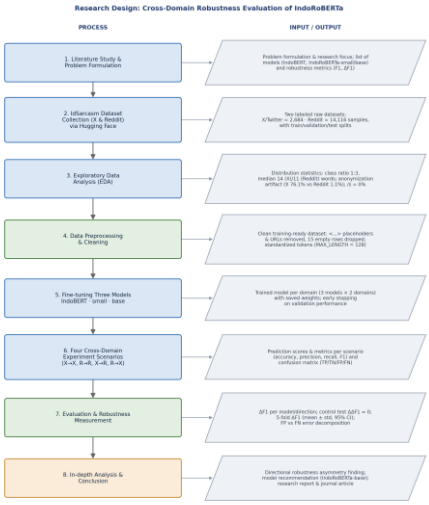

Figure 1. Overall research flow, from literature study to conclusion, with two validity-control stages

E-ISSN: 2580-1643

Dataset and preprocessing

This study uses two Indonesian datasets from the IdSarcasm benchmark (Suhartono *et al.*, 2024), released openly through Hugging Face by Wilson Wongso (account w11wo), one of the benchmark's authors. The first dataset is twitter_indonesia_sarcastic, sourced from the data of Khotijah *et al.* (2020), and the second is reddit_indonesia_sarcastic, collected from the r/indonesia subreddit using a procedure inspired by Ranti and Girsang (2020). Both datasets underwent the same standardization procedure: deduplication with MinHash LSH, personally identifiable information (PII) masking, and class balancing with a sarcastic-to-non-sarcastic ratio of 1:3 following SemEval-2022 Task 6 (iSarcasmEval) (Abu Farha *et al.*, 2022). The task is formulated as binary classification with labels 0 (non-sarcastic) and 1 (sarcastic). The detailed data split (balanced version) used in this study is shown in Table 1.

Table 1. Data Split (Balanced Version) of the two IdSarcasm domains

Domain	Train	Validation	Test	Total
X (w11wo/twitter_indonesia_sacrastic)	1.878	268	538	2.684
Reddit (w11wo/reddit_indonesia_sacrastic)	9.881	1.411	2.824	14.116

It should be noted that the two datasets are not equal in size the Reddit domain has roughly five times more data than the Twitter/X domain. Therefore, the cross-domain performance difference (ΔF1) is interpreted with care, because besides domain shift, the difference in the amount of training data may also affect the results. To mitigate this bias, the training configuration is kept identical across all scenarios, and a data-size control test is conducted (reported later in this section. Both datasets then pass through the same preprocessing and cleaning pipeline before modeling. The preprocessing stage includes tokenization using each model's built-in tokenizer and adjustment of the maximum token length as required by Transformer models. Before modeling, exploratory data analysis was performed on both datasets to ensure data quality, consistency, and comparability across domains. Both domains have a consistent class ratio approximately 25% sarcastic and 75% non-sarcastic (a 1:3 ratio). This consistency supports using F1-score as the main metric (rather than accuracy) and a Stratified K-Fold split that preserves the class proportion in each fold. The distribution statistics across domains are shown in Table 2.

Table 2. Class distribution and text length per domain

Domain	Total Samples	Sarcastic Ratio	Median Length
X	2.684	25%	14
Reddit	14.116	25%	11

The median text length in both domains (14 and 11 words) is far below the MAX_LENGTH = 128 token limit. Although Reddit comments can occasionally run longer than X posts, the overwhelming majority of texts in both domains still fall well within 128 tokens; this value was therefore chosen to cover virtually all samples while keeping memory use and training time efficient, so that truncation practically does not occur and causes no loss of information. Fixing the same maximum length across both domains also removes sequence length as a potential confound in the cross-domain comparison. The most important finding of this analysis is the presence of anonymization placeholder tokens (<username> and <link>) whose distribution is highly imbalanced across domains, as presented in Table 3.

Table 3. Proportion of rows with placeholder tokens

Domain	Rows with <username>/<link>
X	76.1%
Reddit	1.1%

These placeholders appear in 76.1% of the X-domain data but only 1.1% of the Reddit-domain data. This imbalance turns these tokens into a domain-specific artifact: a model can recognize the source domain solely from the presence of these tokens, without genuinely understanding sarcasm. In a cross-domain robustness study, such an artifact becomes a confound that can contaminate the measurement of ΔF1. Without cleaning, an increase in ΔF1 risks reflecting the model's ability to detect the domain rather than a decline in its ability to detect sarcasm. Cleaning these

E-ISSN: 2580-1643

placeholders is therefore not merely cosmetic but a prerequisite for the validity of the robustness measurement. The preprocessing function was applied identically across all experiments (main, control, and K-Fold cross-validation) to ensure comparability of results, with the following steps: (1) removal of anonymization placeholders all bracketed tokens such as <username> and <link> are stripped out; (2) URL removal—remaining web links are deleted; (3) character normalization—characters outside the set of letters, digits, whitespace, and basic punctuation are removed; (4) whitespace normalization multiple spaces are collapsed into a single space and leading/trailing spaces are trimmed; and (5) empty-row removal rows that become empty after cleaning are dropped (15 rows: 2 in X, 13 in Reddit). In addition, the explicit sarcasm marker /s common on Reddit was checked to anticipate label leakage; the /s token appeared in 0.0% of rows in both domains, so there is no risk of label leakage and no special handling was required.

Models, experimental setup, and evaluation

Three models are evaluated in this study, with IndoBERT serving as the comparison baseline and the cross-domain robustness of IndoRoBERTa-small and IndoRoBERTa-base as the main focus. All three models are fine-tuned by adding a classification head on top of the [CLS] token representation to produce binary classification output. To measure robustness, four experimental scenarios were designed covering both in-domain and cross-domain testing in both directions: (1) Train X → Test X (in-domain X); (2) Train X → Test Reddit (cross-domain, X → Reddit direction); (3) Train Reddit → Test Reddit (in-domain Reddit); and (4) Train Reddit → Test X (cross-domain, Reddit → X direction). The level of robustness is measured using the difference in F1 score (ΔF1) between the in-domain and cross-domain scenarios for each direction, as defined in Equation (1) and Equation (2). Here F1(A→B) denotes the F1 score obtained by training on domain A and testing on domain B.

$$\Delta F1\_X = F1(X \rightarrow X) - F1(X \rightarrow Reddit) \quad (1)$$

$$\Delta F1\_Reddit = F1(Reddit \rightarrow Reddit) - F1(Reddit \rightarrow X) \quad (2)$$

All models were implemented using the Hugging Face Transformers library on top of PyTorch, with a fixed random seed across runs to support reproducibility. To keep the comparison fair, the training configuration was made identical across all models and scenarios. The controlled hyperparameters were kept identical across all models: a learning rate of 2e-5, a batch size of 16, up to 10 training epochs with early stopping (patience of 3 epochs based on validation performance), a maximum token length of 128 tokens, a weight decay of 0.01, a warmup ratio of 0.1, and a fixed random seed of 42. The loss function used is binary cross entropy in line with the binary classification formulation; given the class imbalance (1:3), the handling of imbalance is a concern when interpreting the recall of the sarcastic class. Optimization uses the AdamW optimizer (Loshchilov & Hutter, 2019), an an adaptive gradient method with decoupled weight decay that is widely adopted for training Transformer models (Abdulkadirov *et al.*, 2023), and an early-stopping mechanism is applied based on validation performance to prevent overfitting. Finally, model performance is evaluated using accuracy, precision, recall, and F1-score. F1-score is taken as the main metric because it is more representative under imbalanced-data conditions, and the difference in F1-score (ΔF1) between in-domain and cross-domain testing is used as the primary indicator of model robustness.

3. Results and Discussion

3.1 Results

Main experimental results

The main experiment measures cross-domain robustness through the difference between in-domain and cross-domain performance (ΔF1) on the official IdSarcasm split for three models: IndoBERT, IndoRoBERTa-small, and IndoRoBERTa-base. The complete results are presented in Table 4.

Table 4. Main results: F1 and ΔF1 per direction

Model	Direction	F1 in-domain	F1 cross-domain	ΔF1
IndoBERT	X → Reddit	0.650	0.250	0.400
IndoBERT	Reddit → X	0.575	0.438	0.137
IndoRoBERTa-small	X → Reddit	0.604	0.237	0.367
IndoRoBERTa-small	Reddit → X	0.528	0.289	0.240
IndoRoBERTa-base	X → Reddit	0.643	0.298	0.345
IndoRoBERTa-base	Reddit → X	0.593	0.458	0.135

E-ISSN: 2580-1643

A consistent robustness asymmetry is evident: in the X → Reddit direction, all models suffer a large performance drop (ΔF1 0.345–0.400), whereas in the Reddit → X direction the drop is much smaller (ΔF1 0.135–0.240). In other words, models trained on Reddit generalize to X better than the reverse.

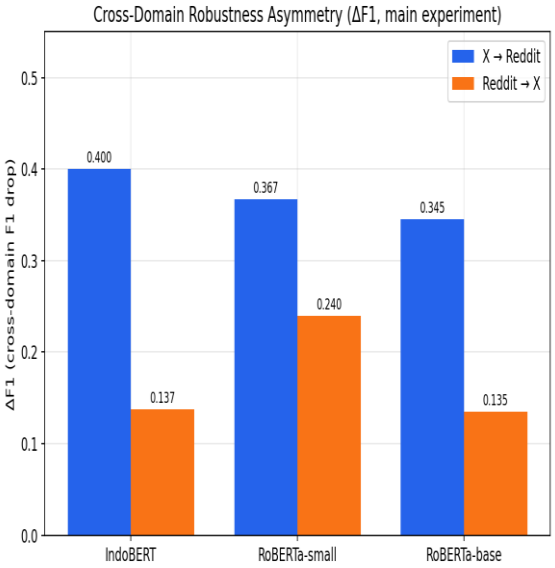

Figure 2. Asymmetry of ΔF1 between the two transfer directions for the three models

Figure 2 visualizes the imbalance in ΔF1 between transfer directions, while Figure 3 clarifies its cause: in the X → Reddit direction the cross-domain F1 collapses to the 0.24–0.30 range, indicating that the models struggle to recognize the longer, more contextual Reddit-style sarcasm. To compare degradation fairly across models with different in-domain starting points, the relative drop (relative drop = ΔF1 / in-domain F1) is used (Calderon *et al.*, 2024). In the X → Reddit direction the relative drop reaches 54–62% (IndoBERT 0.400/0.650 = 61.5%; small 0.367/0.604 = 60.8%; base 0.345/0.643 = 53.7%), while in the Reddit → X direction it is generally much smaller (IndoBERT 0.137/0.575 = 23.8%; base 0.135/0.593 = 22.8%). Even after normalization, the asymmetry pattern persists, reinforcing that this phenomenon is structural rather than merely an effect of performance scale. However, this normalization also reveals one important exception: in the Reddit → X direction, IndoRoBERTa-small experiences a relative drop of 45.5% (0.240/0.528) almost double that of IndoBERT (23.8%) and base (22.8%) and approaching the range of the fragile direction. In other words, even on the normalized metric, small is the least robust model, even in the transfer direction that should be safer consistent with the discussion of model capacity below. It should be noted that the relative drop can inflate when the in-domain F1 is small (small denominator); in this study all in-domain F1 values are in a moderate range (0.528–0.650), so the relative comparison remains reasonable to interpret. One important nuance for a fair interpretation: in the X → Reddit direction, IndoRoBERTa-small has the smallest absolute ΔF1 (0.367), but this is largely because it is also the lowest in-domain performer (F1 0.604), leaving it a narrower "room to fall". In fact, small's cross-domain F1 is the lowest among the three models (0.237 vs. IndoBERT 0.250 vs. base 0.298). Thus, the small ΔF1 of small is not a sign of true robustness but a consequence of its overall limited capacity. Conversely, IndoRoBERTa-base maintains the highest cross-domain F1 (0.298), indicating the best absolute generalization ability.

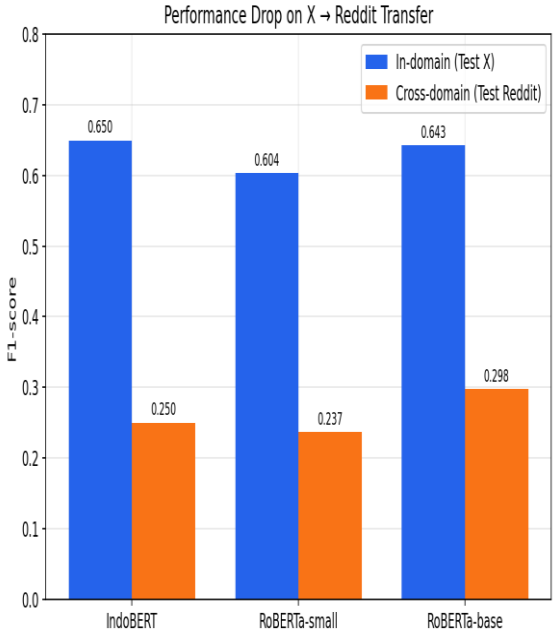

Figure 3. In-domain vs. cross-domain F1 across transfer scenarios

Data-size control test

Because the Reddit dataset (≈ 14,116 samples) is much larger than X (≈ 2,684 samples), a control test was conducted to ensure that the asymmetry is not merely a result of the difference in data amount. In this experiment, Reddit was subsampled to the size of X, and the Reddit → X ΔF1 was compared against the full-data condition (Table 5).

E-ISSN: 2580-1643

Table 5. Data-size control test (Reddit → X)

Model	ΔF1 main	ΔF1 control	ΔΔF1 (Control − Main)
IndoBERT	0.137	0.168	+0.030
IndoRoBERTa-small	0.240	0.201	−0.039
IndoRoBERTa-base	0.135	0.143	+0.008
Avarage			−0.000

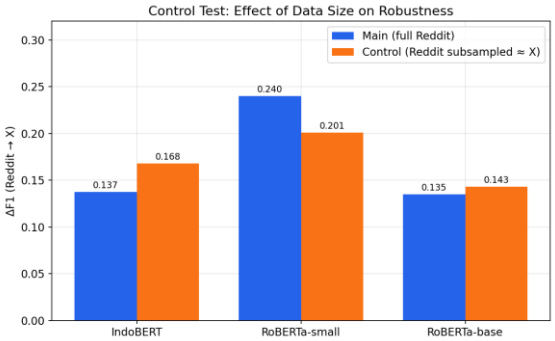

Figure 4. Data-size control test: ΔF1 under full vs. size-matched Reddit training data

The average ΔΔF1 ≈ 0 shows that shrinking Reddit to the size of X barely changes the robustness of Reddit-trained models. Therefore, data size is not the cause of the asymmetry; the asymmetry is intrinsic to the domain characteristics and the style of sarcasm.

K-fold cross-validation

To test the stability of the findings, Stratified 5-Fold cross-validation was performed on the X domain with cross-domain evaluation to Reddit. A summary of the in-domain metrics is presented in Table 6, and the cross-fold ΔF1 in Table 7.

Table 6. 5-Fold in-domain metrics on X (mean ± std)

Model	ΔF1 main	ΔF1 control	ΔΔF1 (Control − Main)
IndoBERT	0.137	0.168	+0.030
IndoRoBERTa-small	0.240	0.201	−0.039
IndoRoBERTa-base	0.135	0.143	+0.008
Avarage			−0.000

Table 7. Cross-fold ΔF1 (X → Reddit) vs. official split

Model	ΔF1 (mean ± std)	ΔF1 (official split)
IndoBERT	0.433 ± 0.017	0.400
IndoRoBERTa-small	0.334 ± 0.019	0.367
IndoRoBERTa-base	0.407 ± 0.018	0.345

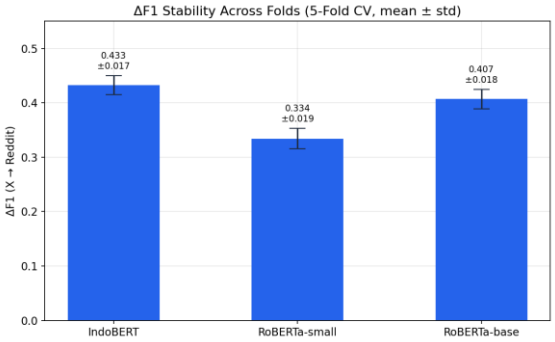

Figure 5. Distribution of ΔF1 (X → Reddit) across the five folds

The small standard deviation of ΔF1 (≈ 0.017–0.019) confirms that the X → Reddit robustness gap is stable across folds, not a coincidence of a single split. Assuming a normal distribution, the 95% confidence interval (≈ mean ± 1.96·std/√5) for ΔF1 is IndoBERT [0.418, 0.448], small [0.317, 0.351], and base [0.391, 0.423]; all lie far from zero, so the cross-domain gap is practically significant and cannot be explained by random fluctuation across folds. In addition, the in-domain F1 from the official split is consistent with the 5-fold average within ±0.01 (IndoBERT 0.650 vs. 0.661; small 0.604 vs. 0.601; base 0.643 vs. 0.642), validating the representativeness of the official split used in the main experiment. The precision–recall pattern in Table 6 is also informative: across all models, precision (0.665–0.760) is consistently higher than recall (0.549–0.590). This means that when a model predicts a text as sarcastic, the prediction is relatively trustworthy, but the model misses most of the actual sarcasm cases. This conservative behavior is consistent with the 1:3 class imbalance in IdSarcasm (the sarcastic class is the minority), which pushes the model toward the majority non-sarcastic class. Consequently, the seemingly high accuracy (0.818–0.849) can be misleading if used as a single measure, so the positive-class F1 remains the more representative main metric for this task (Rainio *et al.*, 2024).

Robustness asymmetry as a directional phenomenon

Across all three experiments, one result stands out: robustness is not symmetric, and the direction of transfer matters a great deal. Going from X to

E-ISSN: 2580-1643

Reddit costs the models far more (ΔF1 0.345–0.400, a relative drop of 54–62%) than going from Reddit to X (ΔF1 0.135–0.240, or 23–46%), and this held for every model and across every fold. What we want to stress is not the size of the drop but where it happens: a model trained on the short X domain struggles badly once it has to read the longer Reddit posts, while the opposite route holds up much better. Earlier sarcasm-detection studies usually report cross-domain degradation without separating the two directions (Chen *et al.*, 2024; Helal *et al.*, 2024; Oprea & Bâra, 2025), and the few cross-domain Indonesian studies look only at the X → Reddit route (Sabrina *et al.*, 2026). Treating direction as a variable in its own right therefore adds something genuinely new to how robustness is understood for Indonesian; it also complements domain-adaptation work on sarcasm in other languages, which has likewise had to bridge sharply different domains, such as tweets and hotel reviews (Ameur *et al.*, 2023).

Underlying mechanism

Why should the direction matter so much? The likely answer lies in how the two platforms actually use language. Sarcasm on Reddit tends to live in longer, more narrative comments that come with plenty of context, such as situational irony, explicit contrast, quotation marks, and the thread a comment is replying to (Bouazizi & Ohtsuki, 2022; Chen *et al.*, 2022). A model trained there learns representations of sarcasm that lean on context and travel reasonably well. X is the opposite: posts are short and dense, carry little context, and rely instead on cues tied to the platform itself, such as an informal register, abbreviations and slang, and lexical markers like hashtags that survive cleaning. What a model picks up from X is therefore shallower, and it falls apart against Reddit-style writing. Recall, too, that emojis and anonymization tokens (<username>, <link>) were stripped out during preprocessing (see the internal-validity discussion below), so the fact that the asymmetry still survives points away from these surface markers and toward a deeper structural difference between the domains, above all the length and contextual richness of the text. This lines up with the domain-generalization principle that causal or invariant features transfer better than features merely correlated with a domain (Khoee *et al.*, 2024).

Error analysis: false positives versus false negatives

To see what the asymmetry looks like at the level of individual mistakes, we split the cross-domain errors into false positives (FP) and false negatives (FN) using the confusion matrix from a single run on the official split. For context, sarcasm accounts for roughly 25% of each domain (X: 134/538; Reddit: 706/2,824), so comparing FP and FN counts across the two directions is fair.

Table 8. Confusion-matrix decomposition (X → Reddit)

Model	TP	TN	FP	FN	Recall	Precision	Pattern
IndoBERT	126	1.942	176	580	0.179	0.417	FN-dominant.
IndoRoBERTa-small	184	1.454	664	522	0.261	0.217	FP+FN
IndoRoBERTa-base	229	1.517	601	477	0.324	0.276	FP+FN

Table 9. Confusion-matrix decomposition (Reddit → X)

Model	TP	TN	FP	FN	Recall	Precision	Pattern
IndoBERT	88	224	176	580	0.179	0.417	FN-dominant.
IndoRoBERTa-small	45	271	664	522	0.261	0.217	FP+FN
IndoRoBERTa-base	79	272	601	477	0.324	0.276	FP+FN

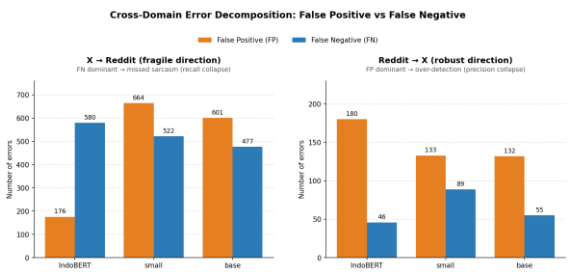

Figure 6. False positives vs false negatives per model and transfer direction

Figure 6 makes the contrast easy to see: in the X → Reddit direction the FN bars tower over everything else, because the model keeps missing sarcasm, whereas in Reddit → X it is the FP bars that stand out, because the model flags too much. Two mirror-image patterns show up consistently, and together they explain the asymmetry from the error side. The first pattern belongs to the move into Reddit. Here recall falls off a cliff, sliding from an in-domain 0.60–0.67 down to just 0.18–0.32, so most Reddit sarcasm slips past the model. IndoBERT is the starkest case: at a recall of 0.179 it misses 580 of the 706 real sarcasm cases (about 83%) and labels only 10.7% of comments positive (302/2,824), well under the 25% base rate, which amounts to severe under-prediction. Having learned from short, low-context X posts, the

E-ISSN: 2580-1643

model simply does not pick up the cues once the text turns long and contextual, and it defaults to guessing non-sarcastic. IndoRoBERTa-small and -base lose recall as well, but they pair that with a surge in false positives (precision 0.217–0.276), so their transfer into Reddit is noisy on both sides rather than merely cautious. The second pattern is the reverse. Moving to X, recall largely survives (0.34–0.66), but precision falls to 0.25–0.37 because the model now calls far too much of the X text sarcastic. IndoBERT records 180 false positives against just 46 false negatives and labels 49.8% of comments positive (268/538), nearly twice the 25% base rate, which is clear over-prediction. Trained on Reddit's context-heavy comments, it seems to over-trigger on sarcasm cues the moment the input becomes short and dense. By way of comparison, when a model is trained and tested on X alone the two error types stay fairly even (IndoBERT 53/44; small 51/54; base 39/52). The lopsided errors are therefore a product of domain shift, not something built into the models. This adds a further layer to the precision–recall discussion in the k-fold cross-validation: the asymmetry carries its own error signature, with the fragile direction (X → Reddit) failing by missing sarcasm (FN) and the sturdier one (Reddit → X) failing by over-flagging (FP). This also points to different fixes for each direction. A Reddit → X model would gain most from precision-oriented threshold calibration to rein in false positives, whereas an X → Reddit model needs its recall brought back up, for instance through class weighting, a focal or weighted loss, or domain adaptation, in order to cut down on false negatives (Calderon *et al.*, 2024; Mulia *et al.*, 2025).

Robustness versus absolute performance: the role of model capacity

These numbers also show why relative robustness (ΔF1) and absolute generalization (cross-domain F1) deserve to be kept apart (Calderon *et al.*, 2024). By average two-directional ΔF1 in the main experiment, IndoRoBERTa-base comes out the most robust (0.240) and IndoRoBERTa-small the least (0.304). The small model does post the lowest ΔF1 in the X → Reddit direction, but that figure is misleading: its in-domain F1 is already low (0.604), so it has little room left to fall. Judged by absolute cross-domain F1 it is in fact the weakest of the three (0.237), while base is the strongest (0.298). The lesson is that ΔF1 should never be read on its own, since a weak model can look "robust" simply because there was not much performance there to lose. Put together, the larger base model gives the best balance of competitive in-domain accuracy and the strongest cross-domain generalization. Two factors plausibly explain why IndoRoBERTa-small underperforms even relative to its already-modest size: with far fewer parameters and Transformer layers it has limited capacity to encode the subtle, context-dependent cues on which sarcasm depends, and the smaller pretraining footprint of the small variant yields weaker general-purpose Indonesian representations to fine-tune from. This is consistent with its behavior here, where it both starts lowest in-domain and transfers worst across domains.

Internal validity: data cleaning as a prerequiste

A word on validity is in order here. Every experiment was run only after the anonymization placeholders (<username>, <link>) had been cleaned out; before cleaning they appeared in 76.1% of the X data but in just 1.1% of the Reddit data. Left in place, those placeholders would have handed the model an easy shortcut for guessing the domain rather than detecting sarcasm (Khoee *et al.*, 2024), and ΔF1 would then track domain detection instead of robustness. Removing them shuts that leakage path and makes the robustness figures far more trustworthy. The data-size control test (Table 5) reinforces this. Cutting Reddit down to the size of X does nothing to remove the asymmetry (average ΔΔF1 ≈ 0). With both of the obvious confounds now accounted for, namely domain-signal leakage and the imbalance in sample counts, the asymmetry that remains can fairly be put down to the characteristics of the domains themselves.

Positioning and practical implications

Set against earlier work, these findings pushx the IdSarcasm line of research (Suhartono *et al.*, 2024) a step further. Most of that work concentrates on in-domain performance (Fitrianto & Editya, 2024; Rahman & Girsang, 2024); here we add the cross-domain robustness dimension and, with it, the transfer direction, which complements the cross-platform studies that came before (Sabrina *et al.*, 2026). Because the gap stays consistent across folds and the official split agrees with the 5-fold average, the results read as reproducible findings rather than methodological flukes. The high-precision, low-recall behavior we observe also fits the well-documented difficulty of classifying imbalanced data in sarcasm-detection tasks (Mulia *et al.*, 2025). There are practical lessons here too. For applications that have to work across platforms, training on the more context-rich domain (Reddit) is a safer bet for generalization than training on a short domain like X. And where the computing budget allows, the larger IndoRoBERTa-base is the model to reach for, since it balances raw performance against robustness better than the alternatives.

E-ISSN: 2580-1643

3.2 Discussion

The discussion surrounding sarcasm detection in the Indonesian language highlights that sarcasm poses a significant challenge in natural language processing, particularly on social media platforms such as X (formerly Twitter) and Reddit. Research by Suhartono *et al.* (2024) emphasizes that sarcasm can reverse the polarity of sentences, leading to difficulties in achieving accurate sentiment analysis. In this context, the study by Sabrina *et al.* (2026) underscores the importance of understanding the cross-domain robustness of language models like IndoRoBERTa, which were tested on the IdSarcasm dataset. This research noted an asymmetry in robustness, where the transfer from X to Reddit exhibited a greater performance drop compared to the transfer from Reddit to X. This finding aligns with the work of Mulia *et al.* (2025), which identified that class imbalance within the dataset could affect sarcasm detection outcomes. Thus, this study emphasizes the need for a more holistic approach to addressing sarcasm detection across various domains, including controlling for anonymization artifacts and conducting data-size control tests to ensure the validity of the results.

4. Conclusion

Based on the main experiment, the data-size control test, and the 5-fold cross-validation on the IdSarcasm benchmark, this study reaches several conclusions in line with the research questions posed. In the in-domain setting, all three models can detect sarcasm within the same domain with F1 in a moderate range. On the X domain, IndoBERT (F1 0.650) and IndoRoBERTa-base (0.643) are relatively comparable and superior to IndoRoBERTa-small (0.604), and a similar pattern holds on the Reddit domain (0.575; 0.593; 0.528). The 5-fold validation confirms this ranking with consistent in-domain X F1 values (IndoBERT 0.661; base 0.642; small 0.601).

When tested across domains, all models experience a clear F1 drop that follows a directional asymmetry: the X → Reddit direction is far more fragile (ΔF1 0.345–0.400; relative drop 54–62%) than Reddit → X (ΔF1 0.135–0.240). This gap is stable across folds, with all 95% confidence intervals lying far from zero, and it is not caused by the difference in data size, since the control test yields an average ΔΔF1 ≈ 0. The error analysis shows distinct failure modes by direction: X → Reddit fails through false negatives (a collapse in recall), whereas Reddit → X fails through false positives (a collapse in precision).

Taken together, IndoRoBERTa-base is the most robust variant while also having the best absolute generalization ability (the highest cross-domain F1), with the lowest average two-directional ΔF1 (0.240). IndoRoBERTa-small, by contrast, is the most fragile; its small absolute ΔF1 in a particular direction is an artifact of low in-domain performance rather than a sign of true robustness. Therefore, for cross-domain Indonesian sarcasm detection, IndoRoBERTa-base is recommended because it balances accuracy and cross-platform stability. The validity of these findings is reinforced by two control measures: cleaning the domain-specific anonymization artifacts, which closes the domain-signal leakage path, and the data-size control test, which rules out the imbalance in sample count as the cause of the asymmetry. This study still has several limitations. The main and control experiments are single-run on the official split; although validated with 5-fold on the X → Reddit direction, a fully bidirectional cross-validation would strengthen the variance estimate. The study is also limited to two domains (X and Reddit) and three Transformer-based models, so generalization to other domains, such as news comments or product reviews, and to large generative models has not been tested. In addition, the error analysis is still limited to a quantitative decomposition of false positives and false negatives, whereas a qualitative inspection of misclassified cases could deepen the understanding of the mechanism; and although removing the anonymization markers improves validity, it also discards some contextual signal, so alternative placeholder-handling strategies are worth exploring. For future research, it is recommended to perform a fully bidirectional cross-validation (X ↔ Reddit), expand the evaluation to a third domain and to large language models, conduct a qualitative error analysis of misclassified examples, and evaluate class-imbalance handling strategies (weighted, focal, or context-aware loss) together with domain adaptation to suppress false negatives in the X → Reddit direction.

5. References

Abdulkadirov, R., Lyakhov, P., & Nagornov, N. (2023). Survey of optimization algorithms in modern neural networks. *Mathematics, 11*(11), 2466. https://doi.org/10.3390/math11112466.

Abu Farha, I., Oprea, S. V., Wilson, S., & Magdy, W. (2022). SemEval-2022 Task 6: iSarcasmEval, intended sarcasm detection in English and Arabic. *Proceedings of the 16th International Workshop on Semantic Evaluation*

(SemEval-2022), 802–814. https://doi.org/10.18653/v1/2022.semeval-1.111.

Ahmadian, H., Abidin, T. F., Riza, H., & Muchtar, K. (2024). Hybrid models for emotion classification and sentiment analysis in Indonesian language. *Applied Computational Intelligence and Soft Computing*, 2024(1), 2826773. https://doi.org/10.1155/2024/2826773.

Alfan Rosid, M., Oranova Siahaan, D., & Saikhu, A. (2024). Sarcasm detection in Indonesian-English code-mixed text using multihead attention-based convolutional and bi-directional GRU. *IEEE Access, 12*, 137063–137079. https://doi.org/10.1109/ACCESS.2024.3436107.

Ameur, A., Hamdi, S., & Yahia, S. B. (2023). Domain adaptation approach for Arabic sarcasm detection in hotel reviews based on hybrid learning. *Procedia Computer Science, 225*, 3898–3908. https://doi.org/10.1016/j.procs.2023.10.385.

Bouazizi, M., & Ohtsuki, T. (2022). Sarcasm over time and across platforms: Does the way we express sarcasm change? *IEEE Access, 10*, 55958–55987. https://doi.org/10.1109/ACCESS.2022.3174862.

Calderon, N., Porat, N., Ben-David, E., Chapanin, A., Gekhman, Z., Oved, N., Shalumov, V., & Reichart, R. (2024). Measuring the robustness of NLP models to domain shifts. *arXiv*. https://doi.org/10.48550/arXiv.2306.00168.

Chen, W., Lin, F., Li, G., & Liu, B. (2024). A survey of automatic sarcasm detection: Fundamental theories, formulation, datasets, detection methods, and opportunities. *Neurocomputing, 578*, 127428. https://doi.org/10.1016/j.neucom.2024.127428.

Chen, W., Lin, F., Zhang, X., Li, G., & Liu, B. (2022). Jointly learning sentimental clues and context incongruity for sarcasm detection. *IEEE Access, 10*, 48292–48300. https://doi.org/10.1109/ACCESS.2022.3169864.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423.

Fitrianto, R. A., & Editya, A. S. (2024). Klasifikasi tweet sarkasme pada platform X menggunakan bidirectional encoder representations from transformers. *Jurnal Teknologi Dan Sistem Informasi Bisnis, 6*(3), 366–371. https://doi.org/10.47233/jteksis.v6i3.1344

Helal, N. A., Hassan, A., Badr, N. L., & Afify, Y. M. (2024). A contextual-based approach for sarcasm detection. *Scientific Reports, 14*(1), 15415. https://doi.org/10.1038/s41598-024-65217-8.

Khan, S., Qasim, I., Khan, W., Khan, A., Ali Khan, J., Qahmash, A., & Ghadi, Y. Y. (2024). An automated approach to identify sarcasm in low-resource language. *PLOS ONE, 19*(12), e0307186. https://doi.org/10.1371/journal.pone.0307186.

Khoee, A. G., Yu, Y., & Feldt, R. (2024). Domain generalization through meta-learning: A survey. *Artificial Intelligence Review, 57*(10), 285. https://doi.org/10.1007/s10462-024-10922-z.

Khotijah, S., Tirtawangsa, J., & Suryani, A. A. (2020). Using LSTM for context-based approach of sarcasm detection in Twitter. In *Proceedings of the 11th International Conference on Advances in Information Technology (IAIT 2020)*. https://doi.org/10.1145/3406601.3406624

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*. https://doi.org/10.48550/arXiv.1907.11692.

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *arXiv*. https://doi.org/10.48550/arXiv.1711.05101.

E-ISSN: 2580-1643

Mulia, B., Emily, M., Damario, V. A., & Hasani, M. F. (2025). Context-aware loss for Indonesian sarcasm detection: A comparative study with conventional imbalance-oriented objectives. *2025 5th International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, 882–888. https://doi.org/10.1109/ICICyTA68677.2025.11362728.

Oprea, S.-V., & Bâra, A. (2025). LLM-as-a-judge for sarcasm detection using supervised fine-tuning of transformers. *Journal of King Saud University Computer and Information Sciences, 37*(10), 357. https://doi.org/10.1007/s44443-025-00379-7.

Rahman, F., & Girsang, A. S. (2024). IndoBERTweet for sarcasm: Evaluating domain-adapted transformers for Indonesian Twitter sarcasm classification. *Journal of Logistics, Informatics and Service Science, 11*(2). https://doi.org/10.33168/JLISS.2024.0210.

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports, 14*(1), 6086. https://doi.org/10.1038/s41598-024-56706-x.

Ranti, K. S., & Girsang, A. S. (2020). Indonesian sarcasm detection using convolutional neural network. *International Journal of Emerging Trends in Engineering Research, 8*(9), 4952–4955. https://doi.org/10.30534/ijeter/2020/10892020.

Sabrina, A. L., Wijaya, D. C. C., & Meiliana. (2026). From Twitter to Reddit: Cross-domain Indonesian sarcasm detection with pretrained transformers. *2026 International Conference on Current Research in Artificial Intelligence and Data Science (ICCRAIDS)*, 1–5. https://doi.org/10.1109/ICCRAIDS67816.2026.11519738.

Šandor, D., & Bagić Babac, M. (2024). Sarcasm detection in online comments using machine learning. *Information Discovery and Delivery, 52*(2), 213–226. https://doi.org/10.1108/IDD-01-2023-0002.

Suhartono, D., Wongso, W., & Tri Handoyo, A. (2024). IdSarcasm: Benchmarking and evaluating language models for Indonesian sarcasm detection. *IEEE Access, 12*, 87323–87332. https://doi.org/10.1109/ACCESS.2024.3416955.

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. *arXiv*. https://doi.org/10.48550/arXiv.2009.05387.