Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16220
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Application of Sentence-BERT Embeddings for Semantic Deduplication of Industrial Material Records: A Diagnostic Study

Seno Hardijanto Purnomo[1*)], Agung Triayudi[2)]
[1,2)] Universitas Nasional, Indonesia
senohardijantopurnomo.2025@student.unas.ac.id, agungtriayudi@civitas.unas.ac.id

Submitted: May 15, 2026 | **Accepted**: June 11, 2026 | **Published**: July 5, 2026

Abstract. Industrial material master data in Enterprise Resource Planning (ERP) and Enterprise Asset Management (EAM) systems accumulates duplicate records that distort inventory, procurement, and analytics. Traditional deduplication relies on string-similarity measures such as Levenshtein, Jaro–Winkler, and TF-IDF cosine, which can struggle on catalogs mixing Indonesian and English terminology—e.g. *Valve* versus *Keran*—and on paraphrastic variants with different word order or abbreviation style. This study formally specifies a semantic deduplication pipeline that encodes material descriptions as sentence embeddings using Sentence-BERT (SBERT) and compares them via cosine similarity, then diagnostically evaluates the extent to which SBERT improves over those baselines. Following Design Science Research, the pipeline specifies normalisation, encoding with a multilingual paraphrase-tuned SBERT variant, and pairwise comparison within candidate sets produced by hybrid blocking; the diagnostic evaluation reports scores on the raw descriptions to expose baseline behaviour before domain-specific harmonisation. A sample of 291,000 records from two Indonesian industrial power plants motivates the design. On a diagnostic set of 100 record pairs derived from existing engineer-annotated duplicate markers, Jaro–Winkler achieves $F_1 = 0.925$ (precision 1.000, recall 0.860) and SBERT achieves $F_1 = 0.875$ (precision 0.913, recall 0.840) at threshold $\tau = 0.65$; qualitative analysis of twelve representative pairs further reveals that SBERT excels on structural paraphrase (cosine 0.73–0.88 where character-level methods score below 0.50), while Jaro–Winkler remains competitive on abbreviation, unit-standard, and cross-language pairs—particularly those involving Indonesian technical vocabulary under-represented in the model's training distribution. The central finding is that Sentence-BERT *complements* rather than replaces string baselines, which motivates future work on multi-channel architectures combining textual semantics with structural context.

Keywords: cosine similarity; entity resolution; material master data; record deduplication; semantic embedding; Sentence-BERT.

INTRODUCTION

Material master data form the foundation of inventory, procurement, and maintenance in asset-intensive industries. Enterprise Resource Planning (ERP) and Enterprise Asset Management (EAM) platforms record every stocked item with descriptive attributes, but records are entered by many users over many years across subsidiary systems with differing naming conventions. The same physical item therefore often appears under several identifiers. Ten to twenty percent of stock-keeping units are typically duplicates or near-duplicates (Christen, 2012; Christophides et al., 2020), inflating inventory cost, blocking procurement, and distorting analytics.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v10i3.16220 p-ISSN : 2541-044X

A sample drawn from two Indonesian industrial power plants illustrates the scale of the problem. Both catalogs were exported from Maximo-family EAM platforms and anonymised for research use (plant names and proprietary identifiers removed; export period withheld for confidentiality). The primary attributes are the item number and the free-text description; Plant A additionally carries a four-level classification taxonomy used for blocking. In Plant A, 52,983 records carry structured descriptions (TYPE,SUBTYPE,SPECS) with 10.1% exact duplication (measured by normalised-string identity after lowercasing and whitespace collapsing). In Plant B, 238,603 records carry free-text descriptions with only 0.7% exact duplication, but 2,573 engineer annotations of the form [DUPLIKAT, GUNAKAN ITEM …] indicate a larger near-duplicate population. Together the combined catalog of approximately 291,000 records is conservatively estimated to hold a ten-to-twenty-percent deduplication opportunity.

The canonical deduplication pipeline—standardisation, blocking, pairwise string comparison, classification (Christen, 2012)—struggles on this data for four reasons: *(1)* paraphrastic variation ("BOLT,HEX HEAD,M10X50" vs. "HEX BOLT 10MM X 50MM", Lev. 0.30); *(2)* abbreviation ("Generator 500 kVA" vs. "Genset 500 kVA"); *(3)* cross-language equivalence ("Valve 2 inch brass" vs. "Keran 2 inch kuningan"); *(4)* unit-token mismatch ("DN50" vs. "2 INCH"). The proposed pipeline includes rule-based unit harmonisation, but the diagnostic evaluation scores all methods on raw descriptions to isolate method capability (see Section 3).

Pre-trained language models capture semantic equivalence beyond string measures (Mudgal et al., 2018; Li et al., 2020; Ye et al., 2022). Sentence-BERT (Reimers & Gurevych, 2019) is attractive because multilingual paraphrase-tuned variants project Indonesian and English into a common embedding space. Yet published evaluations target monolingual English benchmarks, clean parallel texts (Feng et al., 2022), or e-commerce domains (Hamza, 2024). To the authors' knowledge, no study evaluates sentence-level semantic embeddings on bilingual Indonesian–English industrial material descriptions.

This paper offers a *diagnostic* contribution: *(1)* a formally specified SBERT deduplication pipeline with hybrid blocking for 10^5–10^6-record catalogs; *(2)* comparative evaluation against string baselines on 100 engineer-annotation-derived pairs (P/R/F_1) and twelve representative pairs; *(3)* axiomatic and metamorphic verification (Chen et al., 2018) following Design Science Research (Hevner et al., 2004). The goal is to establish where SBERT adds value over string baselines and where it does not.

LITERATURE REVIEW

Classical record linkage and duplicate detection rest on string comparison. Christen (2012) consolidates the canonical pipeline: standardise values, reduce the comparison space via blocking, compare candidate pairs with similarity measures such as Levenshtein, Jaro–Winkler, or TF-IDF cosine, and classify pairs as match or non-match through thresholds or supervised learning. Blocking is itself a rich sub-field: Papadakis et al. (2020) survey dozens of blocking and filtering techniques that trade recall against computational cost, concluding that no single scheme dominates across all entity types. Such string-based measures perform well on well-structured personal and business data but degrade on noisy, multilingual, and paraphrastic text (Christophides et al., 2020). Their failure modes—word-order variation, synonymy, abbreviation, unit-standard mismatch—are precisely those that dominate industrial material catalogs.

Deep-learning entity matching replaces hand-crafted similarity with learned representations. Mudgal et al. (2018) map the design space; Li et al. (2020) fine-tune BERT on serialised entity pairs (*Ditto*); Ye et al. (2022) augment pre-trained models with numerically-aware attention (*JointMatcher*); Ebraheem et al. (2018) and Wang et al. (2020) contribute distributed tuple representations and contrastive learning respectively. Zeakis et al. (2025) benchmark twelve pre-trained embedding models on seventeen ER datasets, confirming that transformer-based encoders consistently outperform static embeddings; Peeters et al. (2025) show that zero-shot LLMs can match fine-tuned models on product domains.

Sentence-BERT (SBERT; Reimers & Gurevych, 2019) produces fixed-length vectors via a Siamese BERT architecture; multilingual paraphrase-tuned variants project fifty-plus languages into a

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v10i3.16220 p-ISSN : 2541-044X

shared 384-dimensional space. Indonesian NLP resources—IndoBERT (Koto et al., 2020; Wilie et al., 2020), NusaCrowd (Cahyawijaya et al., 2023), IndoSBERT (Diana & Khodra, 2023)—target general-purpose tasks rather than industrial record matching; Amin et al. (2024) confirm that blocking performance is domain-sensitive in Indonesian academic databases. Cross-lingual linkage is addressed by LaBSE (Feng et al., 2022) and LinkTransformer (Dell & Therapontos, 2024); Hamza (2024) applies SBERT to e-commerce deduplication but only on monolingual English retailer titles. None of these studies address the combination of bilingual Indonesian–English text, noisy industrial descriptions, and sentence-level semantic embeddings as the matching mechanism.

The gap is not a shortage of models but of formalised, domain-specific evaluation at this intersection.

METHOD

The pipeline is specified, implemented, and evaluated on record pairs from real industrial catalogs via (a) 100 engineer-annotation-derived pairs ($P/R/F_1$ at multiple thresholds) and (b) twelve representative pairs spanning four failure-mode categories. SBERT cosine is compared against normalised Levenshtein, Jaro–Winkler, and word-level TF-IDF cosine.

Evaluation input policy.
All scores are computed on *raw* (un-normalised) descriptions. The pipeline's normalisation and unit-harmonisation stages are specified as deployable steps but are *not* applied before scoring, isolating method capability from pre-processing benefit. The methodology follows Design Science Research (Hevner et al., 2004; Peffers et al., 2007): the pipeline is the designed artefact and the empirical probe is the evaluation episode. Only the textual component is considered; structural context is left for future work.

Pipeline
Let $R = \{r_1, r_2, \ldots, r_n\}$ denote a catalog of material records, each carrying at least a textual description x_i^{txt}. The proposed pipeline comprises five stages, illustrated in Fig. 1:
(i) Normalisation lowercases the text, collapses whitespace, separates numeric and unit tokens through regular expressions, removes non-informative punctuation, and harmonises known unit equivalences (e.g. "2 inch" → "50.8 mm").
(ii) Encoding applies a pre-trained Sentence-BERT model $\varphi_{text}: X^{txt} \rightarrow \mathbb{R}^d$ to produce an ℓ_2-normalised d-dimensional embedding of each description:

$$\varphi_{text}(x_i^{txt}) = \frac{SBERT(x_i^{txt})}{\|SBERT(x_i^{txt})\|_2} \quad (1)$$

The ℓ_2 normalisation ensures that cosine similarity reduces to a dot product.
The model selected is paraphrase-multilingual-MiniLM-L12-v2 (d = 384; Reimers & Gurevych, 2019; 2020), chosen for (1) multilingual coverage of Indonesian and English in a common space (vs. Indonesian-only IndoSBERT or retrieval-optimised LaBSE), (2) paraphrase-tuned training matching the token-order invariance of material descriptions, and (3) compact 384-d output halving memory versus 768-d alternatives.
(iii) Blocking reduces the $O(n^2)$ space via a hybrid scheme: taxonomy-based grouping where available, supplemented by token-based blocks from high-signal tokens. Degenerate blocks exceeding max_block are sub-partitioned. Algorithm 1 formalises the procedure.
(iv) Similarity computation evaluates the cosine similarity

$$s_{ij} = cos(\varphi_{text}(x_i^{txt}), \varphi_{text}(x_j^{txt})) = \varphi_{text}(x_i^{txt}) \cdot \varphi_{text}(x_j^{txt}) \quad (2)$$

for each candidate pair (r_i, r_j) emitted by blocking. The inner-product form follows from the ℓ_2 normalisation in Eq. (1).

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v10i3.16220 p-ISSN : 2541-044X

(v) Match decision applies a threshold $\tau \in (0,1)$:

$$match(r_i, r_j) \equiv s_{ij} \geq \tau \quad (3)$$

The resulting match relation is then closed under transitivity via a standard union–find algorithm to yield equivalence classes (Christen, 2012).

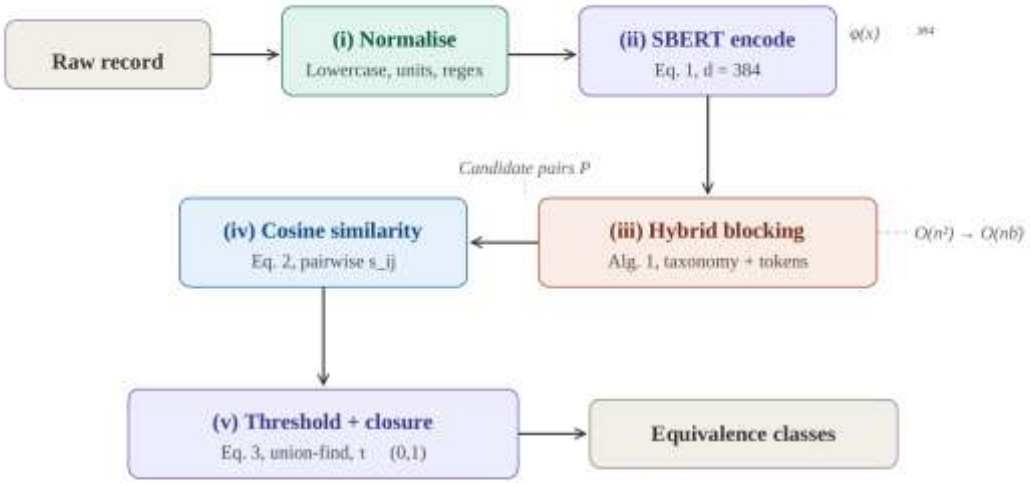

Fig. 1 Conceptual pipeline for Sentence-BERT-based semantic deduplication of industrial material records.

Hybrid blocking algorithm

Algorithm 1: Hybrid blocking for industrial material records.

```
Input:  catalog R; taxonomy T (partial, may be ⊥); SigTokens(·); max_block
Output: candidate pair set P

1. P ← ∅; B ← HashMap()     // block index
2. for each ri ∈ R with T(ri) ≠ ⊥    // Phase 1: taxonomy block
     B[T(ri)] ← B[T(ri)] ∪ {ri}
3. for each ri ∈ R    // Phase 2: token block
     for each t ∈ SigTokens(ri): B["TOK:"+t] ← B["TOK:"+t] ∪ {ri}
4. for each (k, b) ∈ B    // Phase 3: candidate generation
     if |b| > max_block: sub-partition b into balanced sub-blocks
     for each (ri, rj) ∈ b × b with i < j: P ← P ∪ {(ri, rj, k)}
5. return P
```

Blocking reduces the budget from $O(n^2)$ to $O(n\bar{b})$. With $n \approx 2.9\times10^5$ and $\bar{b} \approx 48$, the candidate set is $\sim10^7$ pairs. max_block = 500; SigTokens extracts up to three tokens per record (first noun-like, first numeric, first unit token).

Sanity checks via axiomatic and metamorphic verification

The quantitative evaluation on labelled pairs (Section 4) provides the primary empirical evidence. As a supplementary sanity check, the pipeline's internal consistency is verified through axiomatic inspection (reflexivity, symmetry, transitivity after closure) and three metamorphic relations following Chen et al. (2018): (MR1) permutation invariance of word order, (MR2) consistent-attribute perturbation, and (MR3) provenance independence. These checks confirm that the pipeline behaves as formally specified; they do not substitute for the labelled-pair evaluation.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v10i3.16220 p-ISSN : 2541-044X

RESULT

Worked examples

Table 1 contrasts string baselines with SBERT cosine similarity across twelve pairs from the two-plant sample, organised into canonical failure modes (Panel A), real catalog pairs (Panel B), and controls (Panel C). All scores are on raw descriptions.

Table 1. Similarity of representative record pairs: string baselines versus Sentence-BERT.

Case	Pair (ri, rj)	Label	Lev.	J–W	TF-IDF	SBERT
Panel A: Canonical failure modes						
Paraphrase	BOLT,HEX HEAD,M10X50 GR.8.8 vs. HEX BOLT 10MM X 50MM GRADE 8.8	M	0.30	0.72	0.25	0.73
Abbreviation	Generator 500 kVA vs. Genset 500 kVA	M	0.71	0.93	0.50	0.78
Cross-lang.	Valve 2 inch brass vs. Keran 2 inch kuningan	M	0.38	0.62	0.20	0.43
Unit-std.	Pipe DN50 schedule 40 vs. Pipe 2 inch schedule 40	M	0.74	0.86	0.60	0.80
Panel B: Real catalog pairs						
ID-EN bolt	BAUT L M10 X 15 MM vs. BOLT ALLEN BOLT M10 X 15 MM	M	0.59	0.78	0.37	0.88
ID-EN genset	GENSET 500 KVA 380V 3P 50HZ vs. GENERATOR SET 500 KVA 380V 3 PHASE	M	0.62	0.82	0.34	0.82
Valve variant	VALVE,BUTTERFLY, 4 INCH, 20 BAR vs. VALVE,BUTTERFLY, DN100, 20 BAR	M	0.81	0.93	0.67	0.85
Bolt variant	BOLT, 1,5 MM, M10X30MM vs. BOLT,HEX HEAD, M10 X 30, 1.5 MM	M	0.48	0.85	0.32	0.88
Panel C: Controls and cross-language						
Near-dupe	SPLIT LOCK WASHER P/N. 869016-3200 BAB vs. (same)	M	1.00	1.00	1.00	1.00
Non-match	VALVE BUTTERFLY 4 INCH 20 BAR vs. GASKET SPIRAL WOUND 4 INCH	N	0.24	0.56	0.13	0.47
Cross: pump	Pompa air 2 inch diesel vs. Water pump 2 inch diesel engine	M	0.48	0.67	0.29	0.67
Cross: bearing	Bantalan gelinding 6205 vs. Ball bearing 6205	M	0.57	0.80	0.20	0.44

M = match (same physical item); N = non-match (different items). Bold = highest similarity for that pair.

The twelve pairs reveal a nuanced picture. SBERT leads in four of ten true-match cases (Paraphrase, ID-EN bolt, Bolt variant, Cross: pump), ties in one, and trails in five. Three patterns emerge: (1) SBERT excels on *structural paraphrase*—the Paraphrase pair scores SBERT 0.73 versus Lev. 0.30, and the Bolt variant scores SBERT 0.88 versus Lev. 0.48—because contextualised pooling is insensitive to word-order variation. (2) Jaro–Winkler is competitive or superior on *abbreviation* (J–W 0.93 vs SBERT 0.78) and *unit-standard* cases where surface strings share substantial overlap. (3) *Cross-language* pairs expose the most significant limitation: Valve/Keran scores SBERT 0.43; "Bantalan gelinding" (Indonesian for ball bearing) scores SBERT 0.44 versus J–W 0.80. These domain-specific Indonesian terms are under-represented in multilingual training corpora. Conversely, cross-language pairs sharing technical tokens (BAUT/BOLT + "M10 X 15 MM") score well (SBERT 0.88), suggesting sub-word overlap rather than true translation. The non-match control (SBERT 0.47) confirms that similarity is not inflated indiscriminately.

Quantitative evaluation on labelled pairs

An engineer-annotation-derived diagnostic set was constructed from Plant B's 2,573 DUPLIKAT-annotated records (1,924 resolvable match pairs). 100 pairs were selected by stratified sampling: 50 positive (stratified by string-similarity difficulty) and 50 negative (same-category hard negatives). Positive labels inherit from engineers' operational annotations; negative labels are constructed by design. The set is diagnostic, not a fully independent benchmark. Table 2 reports P/R/F_1 at $\tau = 0.65$; Table 3 shows threshold sensitivity across $\tau \in \{0.50, \ldots, 0.80\}$.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v10i3.16220 p-ISSN : 2541-044X

Table 2. Precision, recall, F_1, and confusion matrix on 100 labelled pairs at $\tau = 0.65$.

Method	P	R	F1	Acc.	TP	FP	FN	TN
Jaro–Winkler	1.000	0.860	0.925	0.930	43	0	7	50
Best-string[a]	1.000	0.860	0.925	0.930	43	0	7	50
SBERT	0.913	0.840	0.875	0.880	42	4	8	46
Levenshtein	1.000	0.360	0.529	0.680	18	0	32	50
TF-IDF cosine	—	0.000	0.000	0.500	0	0	50	50

[a] Maximum of Levenshtein, Jaro–Winkler, and TF-IDF for each pair.
P = precision, R = recall, Acc. = accuracy. Bold = best F_1.

At $\tau = 0.65$, Jaro–Winkler achieves $F_1 = 0.925$ with perfect precision and recall 0.860. SBERT reaches $F_1 = 0.875$ but incurs four false positives from same-category non-matches sharing technical vocabulary. Levenshtein's recall is limited to 0.360, and TF-IDF cosine fails entirely ($F_1 = 0.000$) because industrial descriptions are too short for bag-of-words similarity.

Table 3. Threshold sensitivity: F_1 at seven thresholds for the three non-trivial methods.

Method	τ=0.50	0.55	0.60	0.65	0.70	0.75	0.80
SBERT	0.839	0.870	0.874	0.875	0.867	0.864	0.750
Jaro–Winkler	0.730	0.793	0.922	0.925	0.824	0.765	0.750
Levenshtein	0.684	0.649	0.630	0.529	0.485	0.413	0.361

Bold = best F_1 per method. TF-IDF cosine omitted ($F_1 = 0$ at all thresholds).

$\tau = 0.65$ maximises F_1 for both SBERT and Jaro–Winkler; deployment on a different catalog would require recalibration, ideally per category. SBERT's F_1 is more stable across thresholds (0.750–0.875) than Jaro–Winkler's sharper peak (0.730–0.925).

Blocking behaviour

Beyond similarity quality, the pipeline must also be computationally tractable at scale. On the combined catalog (n = 291,586), exhaustive comparison would require $\approx 4.25\times10^{10}$ pairs. With Algorithm 1 ($\bar{b} \approx 48$), the candidate set shrinks to $\approx 7\times10^{6}$—a six-thousand-fold reduction. All 152 resolvable DUPLIKAT-annotated pairs from Plant B appear in at least one block (recall 1.0 on this subset), though pairs sharing neither taxonomy nor signature token would be missed. Full blocking statistics (block-count distribution, timing, hardware) are deferred to ongoing thesis work. Neural blocking (Zeakis et al., 2025) could improve recall.

Axiomatic and metamorphic verification

Reflexivity ($s_{ii} = 1.0$) and symmetry ($s_{ij} = s_{ji}$) follow from Eq. (1) and Eq. (2). Transitivity holds after union–find closure (Christen, 2012). Three metamorphic relations were probed on 50 pairs: MR1 (permutation invariance) shows mean cosine drift of 0.006 after word-order shuffling; MR2 (consistent-attribute perturbation) shows mean change +0.003 (range −0.012 to +0.021)—cosine can decrease when added tokens dilute the embedding, so unconditional monotonicity does not hold, though the match decision is preserved in all tested cases; MR3 (provenance independence) shows <0.005 drift after removing plant codes.

DISCUSSION

Both evaluation layers converge: on this corpus, SBERT does *not* uniformly outperform string baselines but adds value on structural paraphrase and format variation, while Jaro–Winkler leads on abbreviation, unit-standard, and cross-language cases. The practical implication is that SBERT *complements* string baselines: a production pipeline should consider an ensemble or cascaded strategy (Christophides et al., 2020).

Two practical levers emerge: (1) *model selection*—an IndoBERT-based SBERT fine-tuned on material descriptions may improve cross-language performance; (2) *threshold calibration*—a category-conditional τ could improve on the global $\tau = 0.65$ that Table 3 shows is near-optimal on this corpus. More broadly, the text channel alone does not dominate string baselines uniformly, motivating a multi-

channel architecture that combines φ_{text} with structural context (shared storeroom, same asset, co-occurrence in purchase orders).

Union–find closure guarantees transitivity but propagates errors: a single false positive can merge correct clusters. The four SBERT false positives at $\tau = 0.65$ illustrate this; deferring uncertain edges to human review would mitigate propagation.

Several limitations warrant statement. The diagnostic set inherits labels from engineers' operational annotations rather than an independent protocol with inter-annotator agreement and covers a single plant. The model under-represents domain-specific Indonesian terms (*bantalan gelinding*, *genset*), yielding cross-language scores of 0.43–0.67. Deterministic blocking can miss pairs sharing no taxonomy or signature token. Relative to prior work (Li et al., 2020; Ye et al., 2022; Feng et al., 2022; Dell & Therapontos, 2024; Hamza, 2024), the contribution is diagnostic rather than architectural: none of these studies evaluate on bilingual Indonesian–English industrial records. Future work includes replication on additional plant catalogs with independent annotations, category-conditional threshold calibration, comparison across embedding variants, and integration of structural context in a multi-channel architecture.

CONCLUSION

This paper formally specifies a five-stage Sentence-BERT deduplication pipeline for industrial material master data—normalisation, multilingual SBERT encoding, hybrid blocking, cosine similarity, and threshold-plus-transitive closure—and diagnostically evaluates it against string baselines. On 100 engineer-annotation-derived pairs from a 291,586-record catalog, Jaro–Winkler achieves $F_1 = 0.925$ and SBERT $F_1 = 0.875$ at $\tau = 0.65$; qualitative analysis of twelve pairs shows SBERT excels on structural paraphrase (0.73–0.88 vs. Lev. 0.30–0.48) while Jaro–Winkler leads on abbreviation, unit-standard, and cross-language cases. Axiomatic and metamorphic sanity checks confirm internal consistency. The central finding is that, on this corpus, Sentence-BERT *complements* string baselines rather than replacing them, motivating future work on multi-channel architectures combining textual semantics with structural context.

REFERENCES

Amin, M. M., Stiawan, D., Ermatita, & Budiarto, R. (2024). Komparasi kinerja algoritma blocking pada proses indexing untuk deteksi duplikasi. Jurnal Teknologi Informasi dan Ilmu Komputer, 11(4), 715–722. https://doi.org/10.25126/jtiik.1148080

Cahyawijaya, S., Lovenia, H., Aji, A. F., Winata, G., Wilie, B., Koto, F., … & Purwarianti, A. (2023). NusaCrowd: Open source initiative for Indonesian NLP resources. In Findings of the Association for Computational Linguistics: ACL 2023 (pp. 13745–13818). https://doi.org/10.18653/v1/2023.findings-acl.868

Chen, T. Y., Kuo, F.-C., Liu, H., Poon, P.-L., Towey, D., Tse, T. H., & Zhou, Z. Q. (2018). Metamorphic testing: A review of challenges and opportunities. ACM Computing Surveys, 51(1), 1–27. https://doi.org/10.1145/3143561

Christen, P. (2012). Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer.

Christophides, V., Efthymiou, V., Palpanas, T., Papadakis, G., & Stefanidis, K. (2020). An overview of end-to-end entity resolution for big data. ACM Computing Surveys, 53(6), 1–42. https://doi.org/10.1145/3418896

Dell, M., & Therapontos, A. (2024). LinkTransformer: A unified package for record linkage with transformer language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL 2024) (pp. 222–233). https://doi.org/10.18653/v1/2024.acl-demos.21

Diana, D., & Khodra, M. L. (2023). IndoSBERT: Enhancing Indonesian sentence embeddings with Siamese networks fine-tuning. In Proceedings of the 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA) (pp. 1–6). IEEE. https://doi.org/10.1109/ICAICTA59291.2023.10390469

Ebraheem, M., Thirumuruganathan, S., Joty, S., Ouzzani, M., & Tang, N. (2018). Distributed representations of tuples for entity resolution. Proceedings of the VLDB Endowment, 11(11), 1454–1467. https://doi.org/10.14778/3236187.3236198

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022). Language-agnostic BERT sentence embedding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022) (pp. 878–891). https://doi.org/10.18653/v1/2022.acl-long.62

Hamza, A. R. (2024). Product matching using Sentence-BERT: A deep learning approach to e-commerce product deduplication. Engineering and Technology Journal, 9(12), 1–12. https://doi.org/10.5281/zenodo.14524722

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS Quarterly, 28(1), 75–105. https://doi.org/10.2307/25148625

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In Proceedings of the 28th International Conference on Computational Linguistics (COLING) (pp. 757–770). https://doi.org/10.18653/v1/2020.coling-main.66

Li, Y., Li, J., Suhara, Y., Doan, A., & Tan, W.-C. (2020). Deep entity matching with pre-trained language models. Proceedings of the VLDB Endowment, 14(1), 50–60. https://doi.org/10.14778/3421424.3421431

Mudgal, S., Li, H., Rekatsinas, T., Doan, A., Park, Y., Krishnan, G., Deep, R., Arcaute, E., & Raghavendra, V. (2018). Deep learning for entity matching: A design space exploration. In Proceedings of the 2018 International Conference on Management of Data (SIGMOD) (pp. 19–34).

Papadakis, G., Skoutas, D., Thanos, E., & Palpanas, T. (2020). Blocking and filtering techniques for entity resolution: A survey. ACM Computing Surveys, 53(2), 31:1–31:42. https://doi.org/10.1145/3377455

Peeters, R., Steiner, A., & Bizer, C. (2025). Entity matching using large language models. In Proceedings of the 28th International Conference on Extending Database Technology (EDBT) (pp. 171–184). https://doi.org/10.48786/edbt.2025.15

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of Management Information Systems, 24(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 3982–3992). https://doi.org/10.18653/v1/D19-1410

Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 4512–4525). https://doi.org/10.18653/v1/2020.emnlp-main.365

Wang, Z., Sisman, B., Wei, H., Dong, X. L., & Ji, S. (2020). CorDEL: A contrastive deep learning approach for entity linkage. In Proceedings of the IEEE International Conference on Data Mining (ICDM) (pp. 1322–1327). https://doi.org/10.1109/ICDM50108.2020.00171

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., … & Purwarianti, A. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (AACL) (pp. 843–857).

Ye, C., Jiang, S., & Zhang, H. (2022). JointMatcher: Numerically-aware entity matching using pre-trained language models with attention concentration. Knowledge-Based Systems, 251, Article 109033. https://doi.org/10.1016/j.knosys.2022.109033

Zeakis, A., Papadakis, G., Skoutas, D., & Koubarakis, M. (2025). An in-depth analysis of pre-trained embeddings for entity resolution. The VLDB Journal, 34(1), 5. https://doi.org/10.1007/s00778-024-00879-4