Jurnal Komputer

Vol. 1, No. 1, October 2026

SAH PUBLISHER

# Phonetic Similarity Analysis of Indonesian Text: Comparing Custom Soundex, Jaro-Winkler, Cosine Similarity, and Euclidean Distance

**Ela Herawati**[1*]

[1]Universitas Sugeng Hartono, Sukoharjo, Indonesia

Email: ellaherwati05@gmail.com

**Abstract.** Indonesian text exhibits rich phonetic variation: regional spelling variants, old-to-new orthography changes such as tj→c and oe→u, and digraphs such as ng, ny, kh, and sy that standard phonetic algorithms designed for English fail to capture. This paper presents a comparative analysis of four phonetic similarity algorithms on Indonesian word pairs: (1) a Custom Soundex adapted to Indonesian phonology with 5-digit codes, digraph-aware encoding, old-spelling normalization, and silent-h stripping; (2) Jaro-Winkler similarity computed on grapheme-to-phoneme (G2P) representations, implemented as phoneme similarity via soft Levenshtein distance with phoneme-class substitution costs; (3) Cosine Similarity over positional phoneme vectors with a length-ratio penalty; and (4) Euclidean Distance over the same positional vectors. The evaluation uses a dataset of 10,000 Indonesian word pairs labeled Mirip (similar) or Beda (different), balanced 5,000/5,000 across phonetically similar and control pairs. Decision thresholds were tuned per method: identical codes for Soundex, 0.80 for Jaro-Winkler+G2P, 0.65 for Cosine Similarity, and a distance of 1.5 for Euclidean. Each algorithm classifies pairs with its tuned decision threshold. Custom Soundex ID achieved the best performance with 99.2% accuracy and an F1-score of 0.992, followed by Cosine Similarity (94.9%, F1 0.948), Jaro-Winkler+G2P (91.9%, F1 0.912), and Euclidean Distance (91.7%, F1 0.910). Confusion matrix analysis shows that phoneme-aware encoding resolves most false positives caused by coincidental prefix overlap and subword absorption. The main contribution is empirical evidence that a lightweight, linguistically informed phonetic code outperforms generic string similarity measures for Indonesian, together with a documented preprocessing pipeline of orthography normalization, G2P conversion, and silent-h stripping that is reusable for Indonesian fuzzy text matching.

Received August 2026 / Revised September 2026 / Accepted October 2026

*This work is licensed under a Creative Commons Attribution 4.0 International License.*

**Keywords:** phonetic similarity; Indonesian language; Soundex; Jaro-Winkler; grapheme-to-phoneme; fuzzy string matching

## INTRODUCTION

Phonetic algorithms map words to codes that group together spellings with similar pronunciation, enabling fuzzy matching in search, deduplication, and record linkage. The classic Soundex algorithm, patented in 1918 [1], encodes names by their initial letter followed by consonant-class digits, and its successors such as Double Metaphone [2] improved accuracy for English. Adapting such algorithms to other languages requires linguistic adjustments, as demonstrated for Indian languages [3] and for SMS text [4].

Indonesian (Bahasa Indonesia) poses specific challenges for phonetic matching. Its phonology includes nasal digraphs ng and ny, fricatives kh and sy, and diphthongs ai and au [5], none of which exist as single units in English-oriented encodings. The orthography also changed historically: the old spellings tj, dj, oe, nj, and sj were replaced by c, j, u, ny, and sy, so legacy text and modern text for the same word look different. Indonesian natural language processing has produced resources such as stemmers [6], [7] and the IndoNLU benchmark [8], but systematic comparison of phonetic similarity algorithms for Indonesian word pairs remains scarce.

String similarity measures offer a generic alternative. The Jaro [9] and Jaro-Winkler [10] measures reward common prefixes, Levenshtein edit distance [11] counts insertions, deletions, and substitutions, and surveys of approximate string matching [12] together with comparative studies [13] document their trade-offs. These measures underpin spelling correction [14], record linkage [15], and duplicate detection [16]. However, they operate on raw characters and know nothing about Indonesian phonology.

This paper compares four approaches on a balanced dataset of 10,000 Indonesian word pairs: a Custom Soundex redesigned for Indonesian phonology, Jaro-Winkler similarity applied to phoneme representations from a grapheme-to-phoneme (G2P) conversion, Cosine Similarity on positional phoneme vectors, and Euclidean Distance on the same vectors. The contributions are (1) a documented Indonesian phonetic preprocessing pipeline, (2) a Custom Soundex variant with digraph-aware encoding, and (3) an empirical comparison showing that linguistically informed encoding outperforms generic string similarity for Indonesian.

## METHODS

### Dataset

The experiments use a dataset of 10,000 Indonesian word pairs stored as dataset_fonetik_indo_v2.csv with four columns: Kata_1, Kata_2, Kategori, and Label. The label is Mirip (similar-sounding) for 5,000 pairs and Beda (different) for 5,000 pairs, giving a balanced binary classification task. The category column distinguishes fonetik_mirip (5,000 pairs constructed to sound alike, e.g., bank↔banq, suka↔suqa, dengar↔dengarh, naik↔naiq) from kontrol (5,000 pairs of unrelated words). Table 1 summarizes the composition.

Table 1. Dataset Composition

| Attribute | Value |
|---|---|
| Total word pairs | 10,000 |
| Label Mirip (similar) | 5,000 (50%) |
| Label Beda (different) | 5,000 (50%) |
| Category fonetik_mirip | 5,000 |
| Category kontrol | 5,000 |
| Example pairs | bank↔banq, rumah↔rumahh, banyak↔banyaq |

### Preprocessing Pipeline

All words pass through a three-stage preprocessing pipeline before analysis, illustrated in Fig. 1. First, old-spelling normalization maps legacy orthography to modern EYD spelling: tj→c, dj→j, nj→ny, sj→sy, oe→u, and ch→kh, plus phoneme synonyms q→k and x→ks, since q and k are phonetically identical in Indonesian. Second, silent-h stripping removes a word-final h when preceded by a consonant (bankh→bank, naikh→naik) while preserving rumah, sudah, and buah, where the h follows a vowel and is phonemic. Third, G2P conversion maps Indonesian digraphs to single phoneme tokens: ng→ŋ, ny→Ñ, kh→X, sy→ʃ, ai→ᾱ, and au→Ω. This follows the letter-to-sound tradition [17] with joint-sequence style mappings [18], simplified into deterministic rules sufficient for Indonesian's shallow orthography [19].

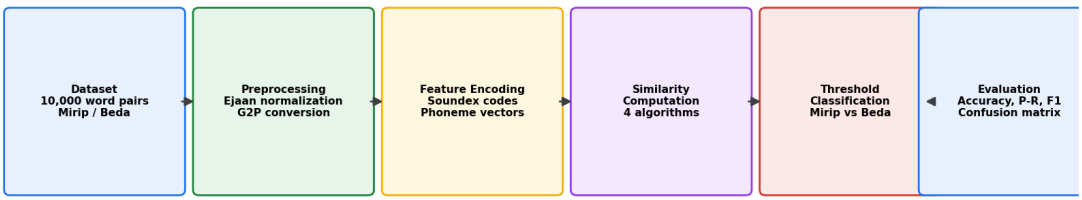


Figure 1. Experiment Pipeline: Preprocessing, Encoding, Similarity Computation, and Evaluation

**Custom Soundex Indonesia**

The Custom Soundex encodes each preprocessed word into a 5-digit code. The first character is kept as the initial letter, and the remaining consonants are mapped through an Indonesian consonant-class table: b/p/f/v → 1, d/t → 2, z/s/ʃ → 3, k/g/c/j/q → 4, l/r → 5, m/n → 6, ŋ (ng) → 7, Ñ (ny) → 8, and X (kh) → 4. Unlike classic Soundex [1], vowels are not aggressively discarded and the digraphs ng, ny, kh, and sy contribute their own classes, which raises precision for Indonesian [20]. Adjacent identical digits collapse, the code is padded or truncated to five characters, and two words are judged Mirip when their codes are identical. This design choice trades the graded scores of Double Metaphone [2] for a transparent binary decision, and it avoids the lexicon-scale retrieval concerns of approximate matching in large vocabularies [21] by operating on a fixed pair-classification task. Soundex-style codes have proven useful beyond names, including speech document indexing [22].

**Jaro-Winkler on G2P Representations**

The second method applies Jaro-Winkler similarity [9][10] to the G2P phoneme strings rather than the raw words. In the implementation this is realized as phoneme similarity: a soft Levenshtein distance [11] where substitution between phonemes of the same articulatory class (e.g., k ↔ g) costs 0.5 instead of 1, normalized by average phoneme length and subtracted from 1. This replaces the original Jaro-Winkler prefix bonus, which produced false positives such as gambar ↔ kamar scoring 0.86 despite different meanings; the phoneme similarity scores the same pair 0.73. A pair is classified Mirip when the score reaches the 0.80 threshold. Comparative studies of string metrics [13] motivate this kind of task-specific adaptation.

**Vector-Based Measures**

The third and fourth methods encode each word as a positional phoneme vector. The vector has three zones — global, first-third, and last-third of the word — over a phoneme alphabet, so anagrams such as baku and kuab produce different vectors where plain frequency vectors would give zero distance. Cosine Similarity [23] is computed on these vectors and multiplied by a length-ratio penalty min/max to suppress false positives from subword absorption (e.g., bang inside abang); term-weighting insights [24] support normalizing for length effects. The decision threshold is 0.65. Euclidean Distance on the same vectors classifies pairs as Mirip when the distance is at most 1.5. Character n-gram representations [25] offer a related alternative, but the positional phoneme vector keeps the representation phonologically grounded.

**Evaluation Protocol**

Each algorithm's predictions are compared against the dataset labels. Performance is reported with accuracy, precision, recall, and F1-score following standard classification measure definitions [26], and confusion matrices (TP, FN, FP, TN) are analyzed per algorithm [27]. Table 2 lists the decision thresholds and key hyperparameters used for every method.

Table 2. Decision Thresholds and Hyperparameters

| Algorithm | Parameter | Value |
|---|---|---|
| Custom Soundex ID | Code length | 5 digits |
| Custom Soundex ID | Decision rule | Identical codes → Mirip |
| Jaro-Winkler+G2P | Similarity threshold | ≥ 0.80 |
| Jaro-Winkler+G2P | Same-class substitution cost | 0.5 |
| Cosine Similarity | Similarity threshold | ≥ 0.65 |
| Cosine Similarity | Length-ratio penalty | min/max (active) |
| Euclidean Distance | Distance threshold | ≤ 1.5 |

| | | |
|---|---|---|
| Euclidean Distance | Vector zones | 3 (global, start, end) |

## RESULT AND DISCUSSION

### Overall Performance

Table 3 reports the classification metrics of the four algorithms on all 10,000 pairs, and Fig. 3 visualizes the accuracy comparison. Custom Soundex ID dominates with 99.2% accuracy, precision 0.997, recall 0.986, and F1-score 0.992. Cosine Similarity ranks second at 94.9% accuracy (F1 0.948), clearly ahead of Jaro-Winkler+G2P at 91.9% (F1 0.912) and Euclidean Distance at 91.7% (F1 0.910). All four exceed the 90% accuracy target, but the gap between the phonetic code and the generic measures is substantial: nearly eight percentage points separate Soundex from Jaro-Winkler.

Table 3. Classification Results on 10,000 Word Pairs

| Algorithm | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Custom Soundex ID | 99.2% | 0.997 | 0.986 | 0.992 |
| Cosine Similarity | 94.9% | 0.969 | 0.928 | 0.948 |
| Jaro-Winkler+G2P | 91.9% | 0.997 | 0.841 | 0.912 |
| Euclidean Distance | 91.7% | 0.997 | 0.837 | 0.910 |




Figure 2. Confusion Matrices of the Four Algorithms (TP, FN, FP, TN)

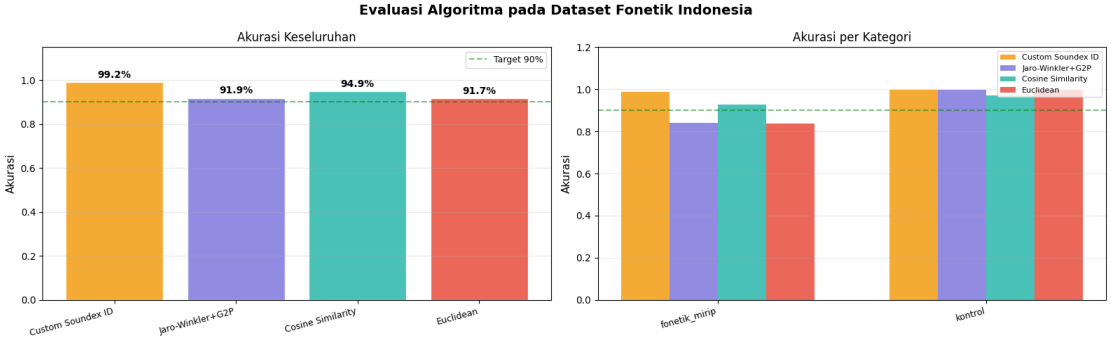



Figure 3. Accuracy Comparison: Overall and per Category

## Confusion Matrix Analysis

Fig. 2 shows the confusion matrices. Custom Soundex ID makes only 83 errors in total (69 FN, 14 FP). The dominant error pattern for Jaro-Winkler+G2P and Euclidean Distance is false negatives (797 and 816): their conservative thresholds miss many genuinely similar pairs, which explains their high precision (0.997) paired with low recall (~0.84). Cosine Similarity shows the opposite tendency with 148 false positives, the price of its bag-of-phonemes character despite the length-ratio penalty. In spelling correction research [14], the same precision-recall tension appears when thresholds move; here the fixed thresholds reveal each method's natural operating point.

### Per-Category Analysis

On fonetik_mirip pairs, Custom Soundex ID reaches 98.6% accuracy, Cosine Similarity 92.8%, Jaro-Winkler+G2P 84.1%, and Euclidean Distance 83.7%. On kontrol pairs, all methods score 97.0–99.8%, confirming that distinguishing unrelated words is easy and that the benchmark's difficulty lies entirely in recognizing true phonetic similarity. The result supports the duplicate-detection literature [16]: blocking-style phonetic codes [28] are most valuable exactly where surface strings diverge but sounds coincide.

### Discussion

Three findings stand out. First, encoding Indonesian-specific phonology (digraphs, old spelling, silent h) into the Soundex code beats generic similarity measures by a wide margin, echoing the lesson from non-English Soundex adaptations [3]. Second, the length-ratio penalty and positional vectors were necessary repairs: without them, Cosine Similarity absorbed subwords and Euclidean Distance equated anagrams. Third, Jaro-Winkler's prefix emphasis is actively harmful for Indonesian pairs with coincidental shared prefixes, justifying its replacement with phoneme-class soft matching. Limitations include the synthetic construction of the fonetik_mirip pairs, which may not reflect the full distribution of real-world spelling variation, and the absence of cross-validation; future work should test on naturally occurring Indonesian noisy text and compare against neural phonetic embeddings.

## CONCLUSION

This paper compared four phonetic similarity algorithms for Indonesian on 10,000 labeled word pairs. Custom Soundex ID, redesigned around Indonesian phonology with digraph-aware encoding, orthography normalization, and silent-h stripping, achieved 99.2% accuracy and F1-score 0.992, outperforming Cosine Similarity (94.9%), Jaro-Winkler+G2P (91.9%), and Euclidean Distance (91.7%). Confusion matrix and per-category analyses show that phoneme-aware encoding resolves the false positives and false negatives that limit generic string measures. The documented preprocessing pipeline and threshold settings provide a reusable baseline for Indonesian fuzzy text matching in search, deduplication, and record linkage. For Indonesian text, monolingual models such as IndoBERT have been shown to outperform multilingual counterparts on classification tasks [29], and recent work on trademark-name similarity explicitly combines lexical, phonetic, and semantic signals [30], supporting the value of the phonology-aware representations explored here.

## REFERENCES


[1] R. C. Russell, "Index," U.S. Patent 1 261 167, Apr. 2, 1918.

[2] L. Philips, "The double metaphone search algorithm," C/C++ Users Journal, vol. 18, no. 6, pp. 38-43, Jun. 2000.

[3] V. Gautam, A. Pipal, and M. Arora, "SoundEx algorithm revisited for Indian language," in Proc. Int. Conf. Innovative Computing and Communications (ICICC 2018), Lecture Notes in Networks and Systems, vol. 56, Singapore: Springer, 2019. DOI: 10.1007/978-981-13-2354-6_6

[4] D. Pinto, D. Vilarino, Y. Aleman, H. Gomez, N. Loya, and H. Jimenez-Salazar, "The Soundex phonetic algorithm revisited for SMS text representation," in Text, Speech and Dialogue (TSD 2012), Lecture Notes in Computer Science, vol. 7499, Berlin: Springer, pp. 47-55, 2012. DOI: 10.1007/978-3-642-32790-2_5

[5] H. Lapoliwa, A Generative Approach to the Phonology of Bahasa Indonesia, Pacific Linguistics, Series D, no. 34, Canberra: Australian National University, 1981. DOI: 10.15144/PL-D34

[6] J. Asian, H. E. Williams, and S. M. M. Tahaghoghi, "Stemming Indonesian," in Proc. 28th Australasian Computer Science Conf. (ACSC 2005), Conferences in Research and Practice in Information Technology, vol. 38, Newcastle, Australia, pp. 307-314, 2005.

[7] M. Adriani, J. Asian, B. Nazief, S. M. M. Tahaghoghi, and H. E. Williams, "Stemming Indonesian: A confix-stripping approach," ACM Trans. Asian Language Information Processing, vol. 6, no. 4, pp. 1-33, Dec. 2007. DOI: 10.1145/1316457.1316459

[8] B. Wilie et al., "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," in Proc. 1st Conf. Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th Int. Joint Conf. Natural Language Processing (AACL-IJCNLP 2020), Suzhou, pp. 843-857, 2020. DOI: 10.18653/v1/2020.aacl-main.85

[9] M. A. Jaro, "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida," J. Amer. Statistical Association, vol. 84, no. 406, pp. 414-420, Jun. 1989. DOI: 10.1080/01621459.1989.10478785

[10] W. E. Winkler, "String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage," in Proc. Section on Survey Research Methods, American Statistical Association, pp. 354-359, 1990.

[11] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," Soviet Physics Doklady, vol. 10, no. 8, pp. 707-710, 1966.

[12] G. Navarro, "A guided tour to approximate string matching," ACM Computing Surveys, vol. 33, no. 1, pp. 31-88, Mar. 2001. DOI: 10.1145/375360.375365

[13] W. W. Cohen, P. Ravikumar, and S. E. Fienberg, "A comparison of string distance metrics for name-matching tasks," in Proc. IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03), Acapulco, Mexico, pp. 73-78, 2003.

[14] K. Kukich, "Techniques for automatically correcting words in text," ACM Computing Surveys, vol. 24, no. 4, pp. 377-439, Dec. 1992. DOI: 10.1145/146370.146380

[15] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," J. Amer. Statistical Association, vol. 64, no. 328, pp. 1183-1210, Dec. 1969. DOI: 10.1080/01621459.1969.10501049

[16] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate record detection: A survey," IEEE Trans. Knowledge and Data Engineering, vol. 19, no. 1, pp. 1-16, Jan. 2007. DOI: 10.1109/TKDE.2007.250581

[17] A. W. Black, K. Lenzo, and V. Pagel, "Issues in building general letter to sound rules," in Proc. 3rd ESCA/COCOSDA Workshop on Speech Synthesis, Jenolan Caves, Australia, pp. 77-80, 1998.

[18] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," Speech Communication, vol. 50, no. 5, pp. 434-451, May 2008. DOI: 10.1016/j.specom.2008.01.002

[19] A. Deri and K. Knight, "Grapheme-to-phoneme models for (almost) any language," in Proc. 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), Berlin, pp. 399-408, 2016. DOI: 10.18653/v1/P16-1038

[20] D. Holmes and M. C. McCabe, "Improving precision and recall for Soundex retrieval," in Proc. Int. Conf. Information Technology: Coding and Computing (ITCC 2002), pp. 22-26, 2002. DOI: 10.1109/ITCC.2002.1000354

[21] J. Zobel and P. Dart, "Finding approximate matches in large lexicons," Software: Practice and Experience, vol. 25, no. 3, pp. 331-345, Mar. 1995. DOI: 10.1002/spe.4380250307

[22] H. Raghavan and J. Allan, "Using Soundex codes for indexing names in ASR documents," in Proc. Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL 2004, Boston, MA, pp. 22-27, 2004.

[23] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," Communications of the ACM, vol. 18, no. 11, pp. 613-620, Nov. 1975. DOI: 10.1145/361219.361220

[24] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information Processing & Management, vol. 24, no. 5, pp. 513-523, 1988. DOI: 10.1016/0306-4573(88)90021-0

[25] G. Kondrak, "N-gram similarity and distance," in String Processing and Information Retrieval (SPIRE 2005), Lecture Notes in Computer Science, vol. 3772, Berlin: Springer, pp. 115-126, 2005. DOI: 10.1007/11575832_13

[26] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information Processing & Management, vol. 45, no. 4, pp. 427-437, Jul. 2009. DOI: 10.1016/j.ipm.2009.03.002

[27] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874, Jun. 2006. DOI: 10.1016/j.patrec.2005.10.010

[28] P. Christen, Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, Data-Centric Systems and Applications, Berlin: Springer, 2012. DOI: 10.1007/978-3-642-31164-2

[29] A. Hapsari and Suyahman, "Comparative Analysis of Trademark Class Identification Using IndoBERT and Multilingual BERT," J. Artificial Intelligence and Legal Technology, vol. 1, no. 1, pp. 1-8, 2025.

[30] M. A. Pratama and Suyahman, "Semantic Analysis of Trademark Names Using Large Language Models," J. Artificial Intelligence and Legal Technology, vol. 1, no. 1, pp. 15-26, 2025.