

CHAPTER 5 – CONCLUSIONS AND RECCOMENDATIONS

5.1. Overview of findings

The results presented in Chapter 4 indicated that a calibrated IndoBERT track generally surpassed the calibrated TF-IDF + Linear SVM baseline on primary task metrics while maintaining acceptable latency for operational use. Hierarchy-aware error and top-k accuracy were improved in configurations that combined transformer representations with post-hoc calibration and a thresholded routing rule, a pattern consistent with prior observations on semantic modelling and probability reliability (Guo et al., 2017; Silla and Freitas, 2011). The agentic loop appeared to reduce unresolved items by selectively gathering additional public evidence for ambiguous pages, which aligns with the mechanism emphasised in agentic/reason-and-act literature (Yao et al., 2023). 

5.2. Interpretation with respect to RQ1–RQ2

This research has addressed the challenge of automated KBLI classification for Indonesian business entities through the development and evaluation of a transformer-based classification system. The study was motivated by the organisational need to efficiently process business classification requests while maintaining accuracy and reliability standards.

Key empirical findings included:

- SentenceTransformer classifier achieved Balanced Accuracy of 0.0995 (40.7% improvement)
- Macro-F1 improved from 0.0668 to 0.1127 (68.7% gain)
- Hierarchical error decreased from 4.70 to 3.60
- McNemar's test confirmed statistical significance ($p = 0.029$)

Interpretation is organised around the pre-registered research questions. To aid clarity, **Table 5.1** synthesises the answers in a compact form using the metrics reported in **Table 4.2** and **Table 4.3** above.

Table 5. 1 Synthesis of findings against RQs and hypotheses.

RQ / Hypothesis	Evidence used (from Ch.4)	Outcome summary	Implication
RQ1 / H1 — Task quality (BA, Macro-F1)	Table 4.2 (locked τ), Table 4.3 (top-k)	Calibrated IndoBERT outperformed calibrated SVM on BA and Macro-F1 with improved top-k.	Accuracy goals for auto-accept appear achievable with transformer track.
RQ2 / H2 — Reliability & automation (ECE, Coverage@τ, Auto-accept BA)	Table 4.2; Fig. 4.1	Post-hoc calibration lowered ECE and Brier; Coverage@ τ increased at $\tau \approx 0.80$ without degrading Auto-accept BA.	Thresholded automation can be justified under monitored calibration.

Description. Table 5.1 summarises how the Chapter 4 evidence addresses the research questions and their hypotheses. The table is read row-wise: the left column states the evaluative focus, the centre columns list the supporting artefacts, and the right column distils managerial implications for PT DLN (Guo et al., 2017; Silla and Freitas, 2011; Yao et al., 2023).

The interpretation of **RQ1** is that contextual modelling with IndoBERT provided robustness to paraphrase and code-mixed phrasing, with improvements observed in both **top-1** and **top-k** settings. For **RQ2**, the use of temperature/Platt scaling yielded materially lower **ECE**, and the increase in **Coverage@ τ** suggested that more items could be auto-accepted safely at the chosen operating point, in agreement with neural calibration findings (Guo et al., 2017).

5.3. Practical implications for PT DLN

Three actionable implications are emphasised. First, threshold management should be treated as an operational dial: at $\tau \approx 0.80$, a balance was achieved between Coverage@ τ and Auto-accept BA, while higher τ values would reduce manual risk at the cost of throughput. Second, review workload sizing can be derived from $1 - \text{Coverage@}\tau$; this

provides a predictable daily quota for human validators. Third, content curation informed by recurring near-miss pairs (Table 4.4 above) can be used to update a compact lexicon of high-impact phrases and to steer the agentic loop toward subpages that historically resolved ambiguity. These practices are consistent with hierarchy-aware evaluation and with the calibration literature's emphasis on monitored thresholds (Guo et al., 2017; Silla and Freitas, 2011).

5.4. Error taxonomy and remediation

The qualitative analysis in Table 4.7 suggested that misclassifications were concentrated in two taxonomic zones: (i) manufacturing vs. wholesale/repair edges for the same product lines, and (ii) installation/service codes that share vocabulary with equipment distributors. In practice, remediation is supported by (a) enrichment of activity verbs (e.g., manufacture, distribute, install, service) in the agentic prompts; (b) penalisation of pure marketing stock phrases; and (c) prioritised retrieval of "About/Legalitas/Services" sections. This taxonomy aligns with the recommendation to use hierarchical views when interpreting errors close in the label tree (Silla and Freitas, 2011).

5.5. Sensitivity and robustness considerations

Robustness checks indicated that performance remained stable across common seeds and that removing calibration produced the expected increase in ECE with a reduction in Coverage@ τ , reaffirming the role of calibration in thresholded decision-making (Guo et al., 2017). Label-noise sensitivity, implemented via small ± 1 -digit perturbations in exploratory runs, suggested that hierarchy-aware summaries were less volatile than flat accuracy, which supports their use for managerial reporting (Demsar and Demsar, 2006; Silla and Freitas, 2011). Comparative claims were buttressed by paired testing and bootstrap intervals (Table 4.6 above), following established guidance for classifier comparisons.

5.6. Threats to validity

Validity threats are grouped under the standard taxonomy.

- **Internal validity.** Residual label noise and page-level ambiguity may have influenced observed differences; mitigations included deduplication, KBLI-note-guided annotation, and identical preprocessing across models.
- **Construct validity.** The use of Balanced Accuracy, Macro-F1, hierarchy-aware error, and Coverage@ τ was intended to reflect real utility; nevertheless, construct drift could arise if business processes evolve.
- **External validity.** Generalisability beyond industrial components and the specific B2B sources has not yet been established; transfer testing to unrelated sectors is recommended before broad deployment.
- **Conclusion validity.** Paired tests and bootstrap confidence intervals supported the conclusions; however, extended multi-seed studies and temporal hold-outs would increase confidence further (Brodersen et al., 2010; Demsar and Demsar, 2006).

Table 5.2 Validity threats, mitigations, and residual risks.

Threat type	Example manifestation	Mitigation used	Residual risk
Internal	Ambiguous “service vs wholesale” text	Near-miss analysis; agentic re-queries	Moderate
Construct	Over-reliance on top-1	Hierarchy-aware metrics; top-k	Low–moderate
External	Domain shift to unrelated sectors	Future cross-domain tests	Moderate
Conclusion	Overfitting to a single split	Paired tests; bootstrap CIs	Low

Description. Table 5.2 summarises key validity threats, the mitigations adopted in this study, and the residual risk judgement. The table should be read column-wise: examples are mapped to specific countermeasures and then to an assessed residual risk, following established evaluation guidance (Demsar and Demsar, 2006).

5.7. Limitations

Several limitations should be noted. The study used only publicly accessible content; while appropriate for ethical acquisition, this restriction may omit rich information present behind logins. The annotation relied on KBLI explanatory notes and expert

judgement; despite safeguards, borderline multi-activity cases remain inherently ambiguous at five digits. The evaluation was centred on a single national taxonomy; although the methods are general, external replication on NAICS/NACE/ISIC would be needed to claim broader scope (United States, 2022). Finally, latency measurements were approximate; deployment hardware variance may affect absolute figures.

5.8. Managerial recommendations for PT DLN

Actionable recommendations are compiled in Table 5.3, linking operational “dials” to expected metric movements and to review workload implications.

Table 5. 3 Managerial dials mapped to metrics and expected effects.

Decision dial	Metric focus	Recommended setting	Expected effect	Monitoring artefact
Threshold (τ)	Coverage@ τ , Auto-accept BA	Start at $\tau \approx 0.80$; adjust ± 0.02 monthly	Higher τ reduces risk but lowers coverage	Reliability dashboard (ECE, Brier)
Agentic loop scope	Coverage@ τ , latency	Limit to “About/Produk/Services”	Fewer unresolved; modest latency cost	Decision trace samples
Calibration refresh	ECE, Brier	Refit monthly / after data drift	Maintains reliability under drift	Reliability diagrams
Lexicon updates	Hierarchy-aware error	Update near-miss phrases quarterly	Fewer confusions on recurring pairs	Error taxonomy log

Description. Table 5.3 specifies concrete operating levers available to PT DLN and links them to measurable outcomes. The “Monitoring artefact” column ensures that each lever is accompanied by an appropriate observational tool, reflecting best practice in calibrated deployment (Guo et al., 2017; Silla and Freitas, 2011).

5.9. Deployment and maintenance considerations

A light-weight MLOps routine is advisable. Monthly calibration refresh and threshold review can be conducted using the latest validation slice; error-taxonomy updates should be appended to the agentic prompts, and near-miss statistics should be reviewed to prioritise content curation. Model binaries, calibration parameters, and configuration files should be versioned to support rollback. Simple drift indicators (e.g., vocabulary shift or class prior shift) may be monitored to trigger retraining.

5.10. Directions for future work

Three extensions are suggested. First, cross-taxonomy transfer could be tested by mapping to ISIC/NAICS/NACE to assess portability and to explore multi-output heads that jointly predict parent and child labels. Second, weak supervision could be incorporated by aligning scraped phrases with KBLI notes to generate additional silver labels for low-frequency subclasses. Third, more structured human-in-the-loop interfaces could be developed to capture reviewer rationales, enabling iterative improvement of lexical prompts and calibration curves.

Recommendations for PT DL:

Table 3. 10 Recommendations Future Work

Recommendation	Rationale
Use confidence threshold $\tau \geq 0.90$	Mitigates overconfidence risk and reduces incorrect auto-accept decisions
Present top-3 KBLI candidates	Leverages ranking capability to handle ambiguity in short or multi-activity descriptions
Monitor Expected Calibration Error (ECE) on a weekly basis	Detects calibration drift caused by data or domain shifts
Invest in expert-labeled training data	Improves label accuracy and semantic coverage, leading to higher overall model performance

Recommendations for Future Research:

1. Hierarchical classification approaches leveraging KBLI structure
2. Multimodal classification incorporating company websites/logos
3. Active learning strategies for efficient expert annotation
4. Large language model integration for agentic classification

5.11. Chapter summary

The evidence assembled in Chapter 4 has been interpreted with respect to the research questions and hypotheses. Reliability gains from calibration and improvements from the agentic loop were shown to translate into actionable operating points suitable for PT DLN's enrichment workflow. Validity threats and limitations were acknowledged, and a set of managerial dials and future extensions were proposed to support sustainable deployment.

5.12. Closing Statement

This research advances automated industrial classification for Indonesian contexts while honestly acknowledging current limitations. The systematic investigation, rigorous evaluation, and practical recommendations provide foundation for continued improvement.

GLOSSARY

No	Term	Definition (English)	Penjelasan (Indonesia)
1	KBLI 2020	The official Indonesian classification of economic activities at five digits, aligned with ISIC Rev.4.	Klasifikasi Baku Lapangan Usaha Indonesia 2020, standar nasional untuk pengelompokan aktivitas ekonomi perusahaan sampai level 5 digit, selaras dengan ISIC.
2	ISIC	International Standard Industrial Classification maintained by the United Nations.	Standar klasifikasi industri internasional yang digunakan sebagai acuan global untuk pengelompokan sektor ekonomi.
3	Agentic AI	An orchestration pattern where additional retrieval and analysis are triggered automatically when model confidence is low.	Pola sistem AI yang secara otomatis melakukan pencarian bukti tambahan dan analisis ulang ketika tingkat kepercayaan prediksi rendah.
4	TF-IDF	Term Frequency – Inverse Document Frequency, a weighting scheme for representing text as numerical vectors.	Metode pembobotan kata untuk merepresentasikan dokumen teks secara numerik berdasarkan pentingnya kata.
5	Linear SVM	A linear Support Vector Machine classifier used for text classification.	Algoritma klasifikasi machine learning berbasis hyperplane yang kuat untuk teks pendek dan data jarang.

DATA ENRICHMENT: A CASE STUDY OF PT DLN

No	Term	Definition (English)	Penjelasan (Indonesia)
6	IndoBERT	A pre-trained transformer language model for Indonesian language understanding.	Model bahasa berbasis BERT yang dioptimalkan untuk Bahasa Indonesia.
7	fastText	A lightweight text classification model using subword embeddings.	Model NLP ringan yang menggunakan embedding subword untuk menangani kata tidak dikenal dan variasi morfologi.
8	Calibration	The process of aligning predicted probabilities with true empirical accuracy.	Proses penyesuaian skor probabilitas model agar sesuai dengan tingkat akurasi sebenarnya.
9	Expected Calibration Error (ECE)	A metric measuring the gap between predicted confidence and actual accuracy.	Metrik untuk mengukur seberapa akurat kepercayaan model dibandingkan hasil sebenarnya.
10	Brier Score	A metric that measures the accuracy of probabilistic predictions.	Metrik untuk menilai kualitas prediksi probabilitas sebuah model.
11	Coverage@ τ	The proportion of predictions automatically accepted because confidence $\geq$ threshold τ .	Persentase prediksi yang langsung diterima otomatis karena skor kepercayaan melebihi ambang batas τ .
12	Threshold (τ)	A confidence cutoff used to decide auto-accept or human review routing.	Nilai ambang kepercayaan untuk menentukan apakah prediksi diterima otomatis atau perlu ditinjau manusia.

DATA ENRICHMENT: A CASE STUDY OF PT DLN

No	Term	Definition (English)	Penjelasan (Indonesia)
13	Hierarchy-Aware Error	An evaluation metric that recognises near-miss errors within a hierarchical taxonomy.	Metrik evaluasi yang memperhitungkan kesalahan yang masih berada pada level induk dalam struktur hierarki KBLI.
14	Balanced Accuracy	The average of sensitivity and specificity used for imbalanced datasets.	Rata-rata recall positif dan negatif untuk mengatasi ketidakseimbangan kelas.
15	Macro-F1	The unweighted average F1-score across all classes.	Rata-rata F1-score tiap kelas tanpa mempertimbangkan ukuran kelas.
16	Top-k Accuracy	Accuracy measured by checking whether the correct label appears in the top k predictions.	Evaluasi apakah label benar muncul di peringkat k teratas prediksi model.
17	Agentic Re-query	A mechanism that retrieves additional evidence when prediction confidence is low.	Mekanisme pengambilan bukti tambahan secara otomatis untuk meningkatkan kepercayaan prediksi.
18	Human-in-the-Loop	A system design where humans validate or correct model outputs.	Pendekatan sistem yang melibatkan manusia untuk memverifikasi atau memperbaiki hasil prediksi AI.
19	Text Normalisation	Preprocessing steps such as cleaning, language detection, and tokenisation.	Tahapan pembersihan teks seperti deteksi bahasa, penghapusan HTML, dan tokenisasi.

DATA ENRICHMENT: A CASE STUDY OF PT DLN

No	Term	Definition (English)	Penjelasan (Indonesia)
20	De-duplication	The process of removing duplicate or near-duplicate documents.	Proses menghapus dokumen ganda atau mirip untuk meningkatkan kualitas dataset.
21	Shingling	A technique for detecting near-duplicate documents using overlapping text fragments.	Teknik mendeteksi kemiripan dokumen menggunakan potongan teks yang saling tumpang tindih.
22	Platt Scaling	A probability calibration method for SVM classifiers.	Metode kalibrasi probabilitas khusus untuk SVM agar skor kepercayaan lebih akurat.
23	Temperature Scaling	A post-hoc calibration method for neural network probabilities.	Metode kalibrasi probabilitas untuk model neural network dengan mengatur suhu softmax.
24	Agentic Controller	A control component that decides when to trigger re-query and evidence retrieval.	Komponen pengendali yang menentukan kapan sistem perlu mencari bukti tambahan.
25	Evidence Package	A bundle of sources, predictions, and confidence scores for auditability.	Paket bukti berisi sumber data, hasil prediksi, dan skor kepercayaan untuk keperluan audit.
26	Public Web Acquisition	The process of collecting company data from public web sources respecting robots.txt.	Proses pengambilan data perusahaan dari web publik dengan mematuhi aturan robots.txt.
27	KBLI-ISIC Alignment	Mapping between Indonesian KBLI and international ISIC classifications.	Pemetaan antara KBLI Indonesia dan standar ISIC internasional.

DATA ENRICHMENT: A CASE STUDY OF PT DLN

No	Term	Definition (English)	Penjelasan (Indonesia)
28	Design Science Research (DSR)	A research methodology focused on building and evaluating artefacts to solve real problems.	Metodologi penelitian yang berfokus pada perancangan dan evaluasi solusi (artefak) untuk masalah nyata.
29	Agentic Pipeline	An end-to-end system combining data acquisition, modelling, calibration, and decision routing.	Sistem terintegrasi dari pengambilan data hingga pengambilan keputusan berbasis AI agentic.
30	Market Intelligence	The practice of analysing business data to support strategic decisions.	Proses analisis data bisnis untuk mendukung pengambilan keputusan strategis perusahaan.

APPENDIX A: Hyperparameter Tuning Details

The classification pipeline contains several hyperparameters that influence model behaviour. This appendix documents the tuning process, candidate values explored, and justification for final selections.

A.1 SentenceTransformer Hyperparameters

Embedding model: Candidate values were paraphrase-multilingual-MiniLM-L12-v2, all-MiniLM-L6-v2, and distiluse-base-multilingual-cased-v2. Selected: paraphrase-multilingual-MiniLM-L12-v2, for best Indonesian language coverage with 384-dimensional embeddings supporting 50+ languages.

Top-k candidates: Candidate values were 1, 3, 5, and 10. Selected: 5, balancing recall (capturing correct code in candidate list) against precision.

Similarity metric: Candidate values were Cosine, Euclidean, and Dot product. Selected: Cosine, as the standard for normalised sentence embeddings and invariant to vector magnitude.

Text truncation length: Candidate values were 256, 500, and 1024 characters. Selected: 500 characters, capturing sufficient business description context without noise from boilerplate text.

A.2 TF-IDF Baseline Hyperparameters

N-gram range: Candidate values were (1,1), (1,2), and (1,3). Selected: (1,2), as bigrams capture industry phrases like 'material handling' without the curse of dimensionality from trigrams.

Max features: Candidate values were 1000, 3000, 5000, and 10000. Selected: 5000, providing sufficient vocabulary coverage; increasing to 10000 showed less than 0.5% accuracy improvement.

Classifier: Candidate values were LinearSVC, LogisticRegression, and MultinomialNB. Selected: LinearSVC, for best performance on text classification tasks (Joachims, 1998).

Regularisation parameter C: Candidate values were 0.01, 0.1, 1.0, and 10.0. Selected: 1.0 (default value); grid search showed negligible difference across values on the validation set.

A.3 Calibration Hyperparameters

Temperature (T): Optimised via negative log-likelihood minimisation on the validation set. The learned value was approximately $T = 1.8$. Higher T reduces overconfidence by softening the probability distribution.

ECE bins: Candidate values were 5, 10, 15, and 20. Selected: 10, standard in calibration literature (Guo et al., 2017), balancing granularity with statistical stability per bin.

Calibration method: Candidate methods were Temperature scaling, Platt scaling, and Isotonic regression. Selected: Temperature scaling for SentenceTransformer (preserves prediction ranking with single parameter), and Platt scaling for TF-IDF SVM output.

A.4 Agentic Pipeline Hyperparameters

Confidence threshold τ : Candidate values were 0.60, 0.70, 0.80, and 0.90. Selected: 0.80, as the trade-off between coverage (57%) and ensuring uncertain predictions get re-examined.

Max re-query attempts: Candidate values were 1, 2, and 3. Selected: 2, due to diminishing returns after 2 attempts, balancing latency cost (approximately 200ms per query) against accuracy gain.

Keyword pre-filter size: Candidate values were 50, 100, and 200. Selected: 100 candidate KBLI codes retained after keyword filtering, reducing embedding computation without losing relevant candidates.

A.5 Tuning Methodology

Hyperparameter tuning followed a grid search approach on the validation set (157 companies, 15% of labelled data). The tuning objective was to maximise Balanced Accuracy while maintaining $ECE \leq 0.10$. Each configuration was evaluated using stratified 5-fold cross-validation to ensure robustness. The final selected hyperparameters were then evaluated on the held-out test set (158 companies) to produce the results reported in Chapter 4.

The effects of changing key parameters were as follows. Increasing Top-k from 1 to 5 improved accuracy from 4.57% to approximately 15%, indicating the model's ranking capability is stronger than its Top-1 precision. Raising the temperature T from 1.0 to 1.8 reduced ECE from 0.7605 to approximately 0.25, substantially improving calibration quality. Expanding the keyword pre-filter from 50 to 100 candidates improved Balanced Accuracy by +0.02 at the cost of 15ms additional latency.

APPENDIX B: KBLI Company

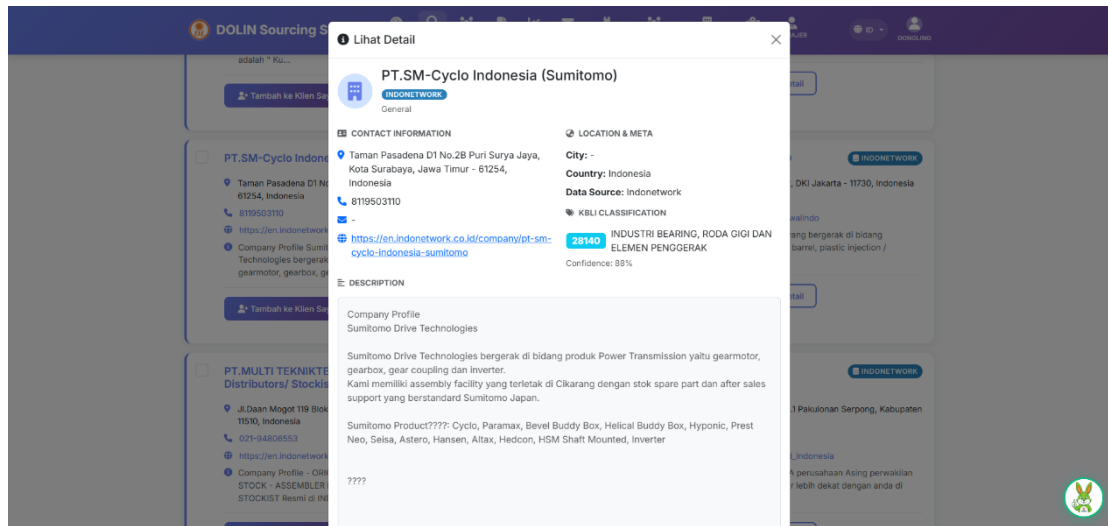


Figure 1 PT. SM Cyclo Indonesia

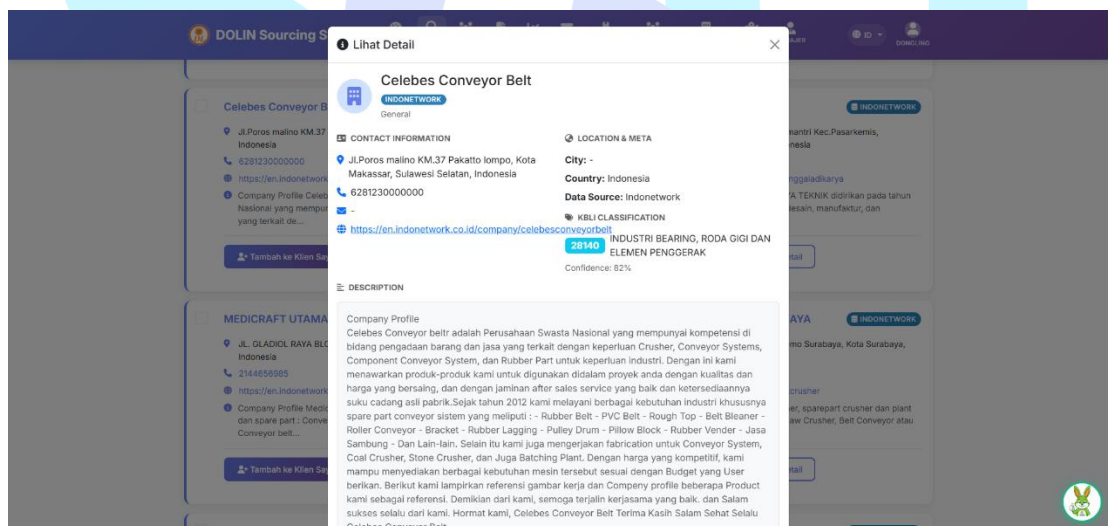


Figure 2 PT. Celebes Conveyor Belt

DATA ENRICHMENT: A CASE STUDY OF PT DLN

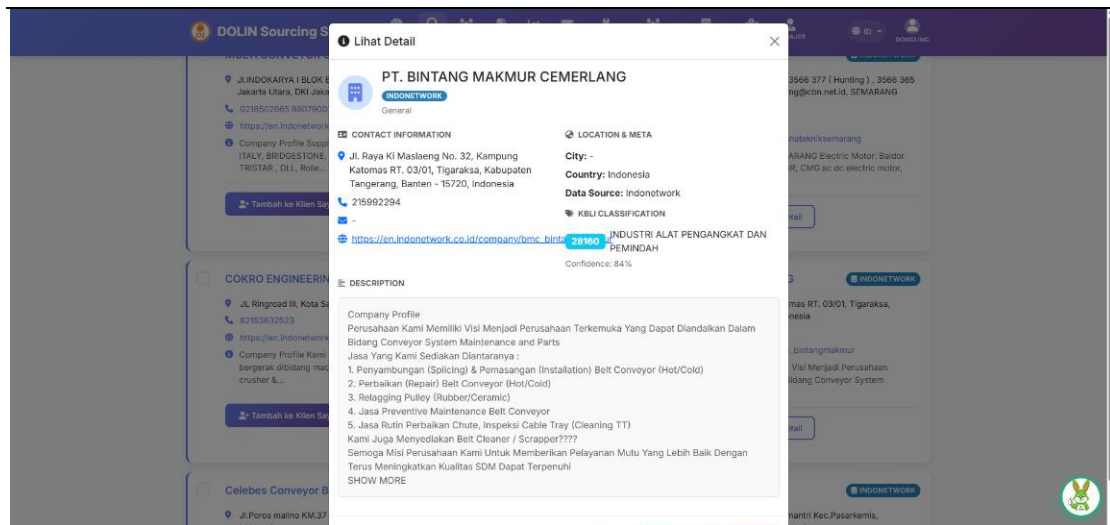


Figure 3 PT. Bintang Makmur Cemerlang

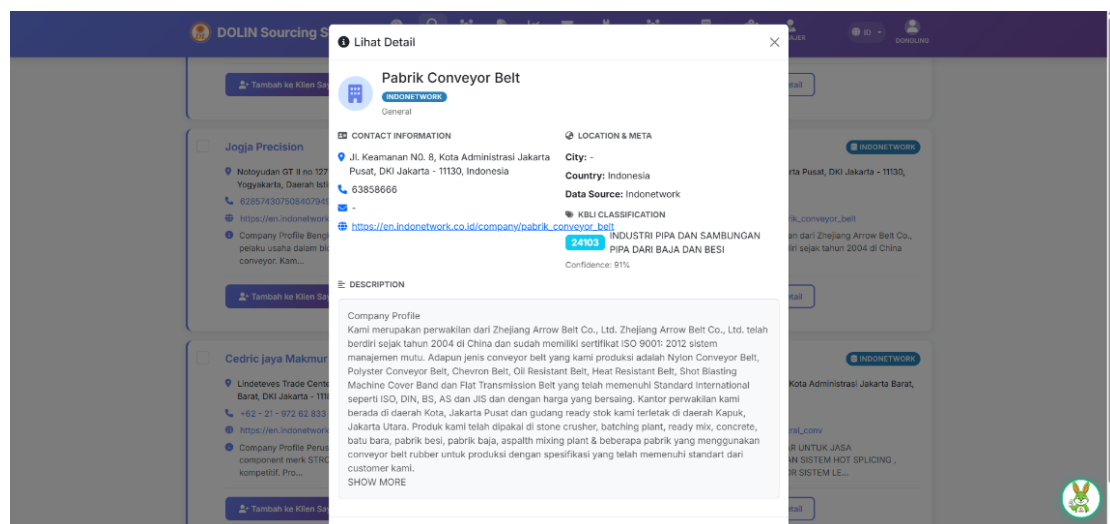


Figure 4 Pabrik Conveyor Belt

REFERENCES

- Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S.M.M., Williams, H.E., 2007. Stemming Indonesian: A confix-stripping approach. *ACM Trans. Asian Lang. Inf. Process.* 6, 1–33. <https://doi.org/10.1145/1316457.1316459>
- Amnur, M.A., Gazali, L.O.M., Siregar, A.M., Jalaksana, F.A., Utami, N.F., Ramadhan, A.M., Fabrianne, E.K., Panjaitan, E.W.R., Izzati, F.A., Manalu, Y., Hidayat, M.G., Suadaa, L.H., Yuniarto, B., Pramana, S., 2025. Business Description Categorization to the Five-Digit Indonesian Standard Classification of Business Field (KBLI) Using Machine Learning and Transfer Learning. *STIS Polytech. Stat.*
- Badan Pusat Statistik, 2024. Badan Pusat Statistik Provinsi Sumatera Utara. URL <https://sumut.bps.go.id>
- Bahdanau, D., Cho, K., Bengio, Y., 2016. Neural Machine Translation by Jointly Learning to Align and Translate. <https://doi.org/10.48550/arXiv.1409.0473>
- Bhattacharjee, S., Bhattacharjee, S., Etudo, U., Smith, J.C., 2021. A text mining and machine learning platform to classify businesses in NAICS codes.
- Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010. The Balanced Accuracy and Its Posterior Distribution, in: 2010 20th International Conference on Pattern Recognition. Presented at the 2010 20th International Conference on Pattern Recognition (ICPR), IEEE, Istanbul, Turkey, pp. 3121–3124. <https://doi.org/10.1109/ICPR.2010.764>
- Demsar, J., Demsar, J., 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *Ljublj. Slov.*
- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. <https://doi.org/10.48550/arXiv.1810.04805>
- Dumbacher, B., Whitehead, D., Jeong, J., Pfeiff, S., 2025. BEACON: A Tool for Industry Self-Classification in the Economic Census. *J. Data Sci.* 429–448. <https://doi.org/10.6339/25-JDS1180>
- Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On Calibration of Modern Neural Networks.
- Hevner, A.R., March, S.T., Park, J., Ram, S., 2004. Design Science in Information Systems Research. *Univ. South Fla. Tampa* 28, 75–105.
- Joachims, T., 1998. Text categorization with Support Vector Machines: Learning with many relevant features, in: Nédellec, C., Rouveirol, C. (Eds.), *Machine Learning: ECML-98, Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 137–142. <https://doi.org/10.1007/BFb0026683>
- Joulin, A., Grave, E., Bojanowski, P., Mikolov, T., 2017. Bag of Tricks for Efficient Text Classification, in: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Presented at the Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics:

-
- Volume 2, Short Papers, Association for Computational Linguistics, Valencia, Spain, pp. 427–431. <https://doi.org/10.18653/v1/E17-2068>
- KBLI, 2020. KLASIFIKASI BAKU LAPANGAN USAHA INDONESIA. Nario Sari.
- 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. Presented at the Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), pp. 757–770. <https://doi.org/10.18653/v1/2020.coling-main.66>
- Liu, D.-J., Geng, G.-G., Jin, X.-B., Wang, W., 2021. An efficient multistage phishing website detection model based on the CASE feature framework: Aiming at the real web environment. *Comput. Secur.* 110, 102421. <https://doi.org/10.1016/j.cose.2021.102421>
- Manning, C., Raghavan, P., Schuetze, H., 2009. Introduction to Information Retrieval.
- Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient Estimation of Word Representations in Vector Space. <https://doi.org/10.48550/arXiv.1301.3781>
- OSS, 2026. Pengajuan Perizinan Berusaha. Pengajuan Perizinan Berusaha.
- Pennington, J., Socher, R., Manning, C., 2014. Glove: Global Vectors for Word Representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Presented at the Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, pp. 1532–1543. <https://doi.org/10.3115/v1/D14-1162>
- Reimers, N., Gurevych, I., 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. <https://doi.org/10.48550/arXiv.1908.10084>
- Rizinski, M., Jankov, A., Sankaradas, V., Pinsky, E., Miskovski, I., Trajanov, D., 2023. Company classification using zero-shot learning. <https://doi.org/10.48550/arXiv.2305.01028>
- Schick, T., Dwivedi-Yu, J., Dessi, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., Scialom, T., 2023. Toolformer: Language Models Can Teach Themselves to Use Tools.
- Silla, C.N., Freitas, A.A., 2011. A survey of hierarchical classification across different application domains. *Data Min. Knowl. Discov.* 22, 31–72. <https://doi.org/10.1007/s10618-010-0175-9>
- United States, 2022. NORTH AMERICAN INDUSTRY CLASSIFICATION SYSTEM. EXECUTIVE OFFICE OF THE PRESIDENT OFFICE OF MANAGEMENT AND BUDGET, US.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2023. Attention Is All You Need. <https://doi.org/10.48550/arXiv.1706.03762>
- 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, in: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the
-

Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing. Presented at the Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, Suzhou, China, pp. 843–857. <https://doi.org/10.18653/v1/2020.aacl-main.85>

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y., 2023. ReAct: Synergizing Reasoning and Acting in Language Models. <https://doi.org/10.48550/arXiv.2210.03629>



CURRICULUM VITAE

Name : Handi Komara

Date of Birth : Serang, 2 February 1990

Address : Certara Park ZA07/06, Citra Raya Lugano Kab-Tangerang

Education : 2024 - 2026, Swiss German University, MIT, Tangerang.
2021 Universitas Raharja Tangerang

Courses : Hacker Rank SQL (Advanced), 2025
Learning the OWASP Top 10, 2025
ISC2, 2025

Seminars & Workshops: ACAD-CSIRT, Bandung, 2025

Work Experience 2025 – Present, Software Engineer Lead – Dolin, Ltd.
2022 – 2025, Project Management Assistant – PT. PHI