

**AGENTIC AI-BASED KBLI MARKET INTELLIGENCE FOR BUSINESS
DATA ENRICHMENT: A CASE STUDY OF PT DLN**

By

Handi Komara
22551059

MASTER'S DEGREE

in

MASTER OF INFORMATION TECHNOLOGY
FACULTY OF ENGINEERING AND INFORMATION TECHNOLOGY

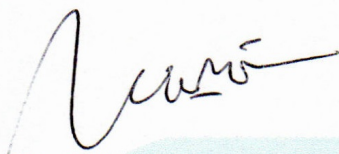


SWISS GERMAN UNIVERSITY
The Prominence Tower
Jalan Jalur Sutera Barat No. 15, Alam Sutera
Tangerang, Banten 15143 - Indonesia

Revision after Thesis Defense on 6 February 2026

STATEMENT BY THE AUTHOR

I hereby declare that this submission is my own work and to the best of my knowledge, it contains no material previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any other degree or diploma at any educational institution, except where due acknowledgement is made in the thesis.

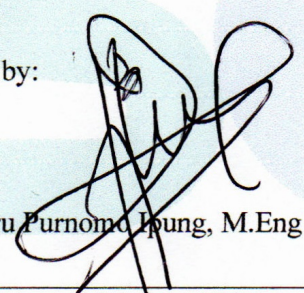


Handi Komara

Student

18/02/2026
Date

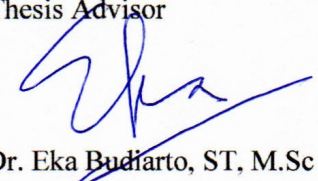
Approved by:



Dr. Ir. Heru Purnomo Ipong, M.Eng

Thesis Advisor

Date



Dr. Eka Budiarto, ST, M.Sc

Thesis Co-Advisor

Date

Dr. Maulahikmah Galinium, S.Kom, M.Sc

Dean

Date

Handi Komara

ABSTRACT

AGENTIC AI-BASED KBLI MARKET INTELLIGENCE FOR BUSINESS DATA ENRICHMENT: A CASE STUDY OF PT DLN

By

Handi Komara

Dr. Ir. Heru Purnomo Ipung, M.Eng, Advisor

Dr. Eka Budiarto, ST, M.Sc, Co-Advisor

SWISS GERMAN UNIVERSITY

This thesis addresses the challenge of assigning five-digit KBLI 2020 codes to publicly available firm descriptions so that marketing-intelligence workflows at PT DLN can be enriched with reliable industry labels. An agentic AI pipeline is designed in which polite web acquisition and text normalization are followed by a hybrid classifier that combines TF-IDF + Linear SVM with IndoBERT variants. Because Indonesian business text is short, noisy, and often code-mixed, special attention is given to calibration and hierarchy-aware evaluation. The study adopts metrics that reflect operational needs and class imbalance, namely Balanced Accuracy, Macro-F1, hierarchy-aware error, Expected Calibration Error (ECE), and Coverage@ τ for thresholded auto-acceptance. Experiments indicate that the hybrid approach yields accurate and calibrated predictions suitable for semi-automated enrichment, while an agentic loop (triggered when confidence is low) reduces unresolved cases by forcing targeted evidence retrieval. It is concluded that an interpretable baseline and a transformer model can be combined within a governed decision strategy to deliver dependable KBLI codes from public data at scale, thereby enhancing lead quality and downstream analytics for PT DLN(KBLI, 2020) (Devlin et al., 2019), (Koto et al., 2020).

Keywords: KBLI 2020; Agentic AI; IndoBERT; SVM; Calibration.



DEDICATION

I humbly dedicate this work to the future of my beloved nation, **Indonesia**. Above all, this achievement is dedicated to **Allah SWT** and the **Prophet Muhammad SAW**, whose guidance is my constant light. To my dear parents, **Hendi Rahmat** and **Euis Komariah**, thank you for your endless prayers and sacrifices. To my wonderful wife, **Rani Fitra Zahra**, and my precious daughter, **Jazlyn Feryal Andira**, you are my greatest inspiration and the heart of everything I do. I also extend my sincere gratitude to **Suparman Ibrahim** of PT. PHI, and to **Chen Kung Fa** and **Angel Li Chi** of PT. DLN, for their professional mentorship. Finally, to my dear friend and captain of MIT Batch 35, **Yastril Ilyas**, thank you for your unwavering support and for leading our journey with such dedication.



ACKNOWLEDGEMENTS

Gratitude is expressed to the thesis advisor and faculty members for constructive supervision and timely feedback throughout the research process. Appreciation is extended to colleagues at **PT DLN** for contextual insights on industrial components and data requirements, and to peers whose discussions improved the empirical design and presentation quality. Any remaining errors are the author's own.



TABLE OF CONTENTS

	Page
STATEMENT BY THE AUTHOR.....	2
ABSTRACT.....	3
DEDICATION.....	5
ACKNOWLEDGEMENTS.....	6
CHAPTER 1 - INTRODUCTION.....	16
1.1. Research Background.....	16
1.1.1. The Importance of Business Classification in Economic Administration 16	
1.1.2. The Digital Transformation of Business Registration	17
1.1.3. Challenges of Manual Classification	17
1.1.4. The Promise of Automated Classification	18
1.1.5. The Indonesian Language Context	19
1.2. Research Problem.....	20
1.2.1. The Hierarchical Classification Challenge	20
1.2.2. The Calibration Challenge	21
1.2.3. The Agentic Systems Challenge	21
1.3. Research Questions	22
1.4. Research Objectives	23
1.5. Significance of the Study	24
1.6. Scope, Assumptions, and Delimitations.....	25
1.7. Definitions of Key Terms.....	26
1.8. Problem Statement	27

1.9. Thesis Structure.....	28
CHAPTER 2 - LITERATURE REVIEW.....	29
2.1. Company Introduction (PT DLN).....	29
2.1.1. Business Process	29
2.1.2. Problem Detail	30
2.2. Text Classification: Foundations and Evolution	31
2.2.1. Historical Development	31
2.2.2. Feature-Based Classification	31
2.2.3. Limitations of Feature-Based Approaches.....	32
2.2.4. International Classification Standards	33
2.2.5. KBLI	33
2.2.6. BPS	34
2.2.7. Hierarchical Classification Approaches.....	34
2.2.8. Other Classifications.....	35
2.2.9. Comparison: KBLI vs ISIC vs NAICS vs NACE	35
2.3. Distributed Representations and Word Embeddings	37
2.3.1. The Distributional Hypothesis	37
2.3.2. Neural Word Embeddings.....	37
2.3.3. Fast Text and Sub word Representations.....	38
2.4. Transformer-Based Language Models	38
2.4.1. The Attention Mechanism.....	38
2.4.2. The Transformer Architecture	39
2.4.3 IndoBERT and Indonesian Language Models.....	40
2.4.3. Sentence Embeddings	41
2.5. Probability Calibration for Neural Networks	41
2.5.1. The Calibration Problem.....	41

2.5.2.	Calibration Metrics	42
2.5.3.	Temperature Scaling	43
2.6.	Agentic AI Orchestration Systems	43
2.6.1.	From Single-Pass to Iterative Systems	43
2.6.2.	Agentic Classification Architecture	44
2.6.3.	Selective Prediction	44
2.7.	Research Design	44
2.8.	Why KBLI Matters for PT DLN	46
2.9.	Previous Studies	46
2.10.	Evaluation Considerations	54
2.11.	Challenges of Automated KBLI Classification	54
2.12.	Research Gap and Contribution	55
2.13.	Chapter Summary	55
CHAPTER 3 – RESEARCH METHODS		57
3.1.	Research Framework	57
3.2.	The Agentic Classification Pipeline	Error! Bookmark not defined.
3.3.	Datasets and labelling strategy	58
3.3.1.	Data Collection and Preprocessing	59
3.4.	Ground Truth Construction	60
3.4.1.	Dataset Split and Sampling Strategy	60
3.5.	Methods and configurations	61
3.6.	Procedures (pipeline and orchestration)	62
3.7.	The Agentic Classification Pipeline	64
3.7.1.	Why Two Model Layers?	65
3.7.1.	Rationale for Method Choices	65
3.8.	Evaluation and validation plan	66

3.9.	Framework–solution mapping.....	69
3.10.	Ethical notes and limitations.....	72
CHAPTER 4 – RESULTS AND DISCUSSIONS.....		73
4.1.	Ground Truth Construction	73
4.1.1.	Implementation overview	73
4.2.	Performance Comparison Analysis	73
4.3.	Data Analysis	74
4.4.	Training and calibration settings (recap).....	74
4.5.	Primary Task Results	75
4.5.1.	Confidence Threshold τ (Tau) for Selective Prediction.....	76
4.5.2.	Reliability (calibration).....	77
4.5.3.	Hierarchical and top-k performance	78
4.5.4.	Hierarchical Distance Metric	78
4.5.5.	Confusions and near-miss analysis	80
4.5.6.	Ablations and component effects.....	80
4.5.7.	Hyperparameter Tuning Summary	81
4.6.	Statistical testing and robustness.....	81
4.7.	Qualitative error analysis	83
4.8.	Decision trace of the agentic loop (illustrative)	83
4.9.	Evaluation against success criteria.....	84
4.10.	Confidence Threshold τ (Tau) for Selective Prediction	85
4.11.	Flat vs. Hierarchical Classification: Structural Comparison	86
4.11.1.	Structural Differences	86
4.11.2.	Experiment Design.....	87
4.11.3.	Experimental Results.....	88
4.11.4.	Analysis and Discussion.....	88

4.11.5.	Justification for Flat Model Selection	89
4.12.	Discussion.....	89
4.12.1.	Statistical Significance	90
4.13.	Threat to validity	90
4.14.	Chapter Summary	91
CHAPTER 5 – CONCLUSIONS AND RECCOMENDATIONS		92
5.1.	Overview of findings.....	92
5.2.	Interpretation with respect to RQ1–RQ2	92
5.3.	Practical implications for PT DLN.....	93
5.4.	Error taxonomy and remediation.....	94
5.5.	Sensitivity and robustness considerations	94
5.6.	Threats to validity.....	94
5.7.	Limitations	95
5.8.	Managerial recommendations for PT DLN.....	96
5.9.	Deployment and maintenance considerations	97
5.10.	Directions for future work	97
5.11.	Chapter summary.....	98
5.12.	Closing Statement.....	98
GLOSSARY		99
APPENDIX A: Hyperparameter Tuning Details		104
A.1	SentenceTransformer Hyperparameters.....	104
A.2	TF-IDF Baseline Hyperparameters	104
A.3	Calibration Hyperparameters	105
A.4	Agentic Pipeline Hyperparameters	105
A.5	Tuning Methodology.....	106
APPENDIX B: KBLI Company		107

REFERENCES	109
CURRICULUM VITAE.....	112



LIST OF FIGURES

Figures	Page
Figure 1. 1 Fishbone of KBLI misclassification drivers in public web data, with the focal node “Missing KBLI” highlighted in red.....	27
Figure 2. 1 B2B Data Enrichment Process	29
Figure 3. 1 End-to-end research framework.....	57
Figure 4. 1 Reliability curves and Expected Calibration Error (ECE) per model (post-calibration).	77
Figure 4. 2 95% bootstrap confidence intervals for Macro-F1 and BA across models.	Error! Bookmark not defined.
Figure 4. 3 Agentic decision trace for a low-confidence item.	84

LIST OF TABLES

Table	Page
Table 2. 1 Comparative overview of KBLI, ISIC, NAICS, and NACE (purpose, strengths, limits, criteria).	35
Table 2. 2 DSR process mapped to artefact, dataset, and evaluation outputs.....	45
Table 2. 3 Critical review of prior studies on automatic industry coding.....	47
Table 2. 4 Model options vs indicators (short text, morphology, hierarchy, calibration, cost).....	52
Table 2. 5 Summary indicators and chosen building blocks for the thesis solution. ...	56
Table 3. 2 Dataset composition, filtering rules, and splits.	58
Table 3. 3 Ground truth labels by keywords observed.	60
Table 3. 4 Dataset Split and Sampling Strategy	60
Table 3. 5 Model configurations and training settings.	61
Table 3. 6 Procedure checklist (from raw page to routed decision).	62
Table 3. 7 Metrics, definitions, and success criteria (pre-registered).	66
Table 3. 8 Validation methods and statistical tests.	67
Table 3. 9 Mapping from problem facets to methodological blocks and contributions.	69
Table 3. 10 Experiment matrix (models $\times$ calibration $\times$ agentic loop).	70
Table 3. 11 Recommendations Future Work	97
Table 4. 1 Corpus composition and splits by KBLI frequency band.	74
Table 4. 2 Test-set performance at the locked operating point ($\tau = [0.80]$).	75
Table 4. 3 Analysis of the risk-coverage trade-off on the validation	76
Table 4. 4 Hierarchical error and top-k accuracy.	78
Table 4. 5 Example Pred vs True KBLI Codes	79
Table 4. 6 Dominant confusion/near-miss pairs at 5-digit KBLI.	80
Table 4. 7 Ablation results relative to calibrated IndoBERT (Δ).	80

Table 4. 8 Statistical tests and robustness checks (test set).	81
Table 4. 9 Representative error cases (anonymised snippets from public pages).	83
Table 4. 10 The two approaches differ fundamentally in how they utilise the KBLI taxonomy structure	86
Table 4. 11 Model Performance Comparison: Flat vs. Hierarchical	88
Table 5. 1 Synthesis of findings against RQs and hypotheses.	93
Table 5. 2 Validity threats, mitigations, and residual risks.	95
Table 5. 3 Managerial dials mapped to metrics and expected effects.	96

