

CHAPTER 2 - LITERATURE REVIEW

2.1. Company Introduction (PT DLN)

PT DLN is positioned as an industrial components provider in Indonesia with product lines centred on electric motors, gear reducers, and variable frequency drives (VFDs). The company's market development relies on accurate classification of prospective B2B customers by activity type so that sales campaigns and distributor enablement can be targeted at segments with a demonstrable fit (e.g., conveyors, packaging machinery, and automation cells). Because most candidate firms are discovered from public web sources (official portals, directories, and corporate sites), the practical challenge is not data scarcity but the reliable assignment of KBLI **codes** from short, noisy descriptions. Consistent five-digit KBLI labelling directly affects list quality, campaign precision, and follow-up cost. In this thesis, PT DLN serves as the case study context: the proposed agentic AI pipeline will be exercised on PT DLN's prospecting corpora to produce auditable KBLI predictions and short listable leads that can be actioned by marketing. The Indonesian statistical and regulatory environment defines KBLI as the national standard; therefore, the mapping logic, error handling, and acceptance thresholds are aligned to KBLI 2020's structure and notes (Badan Pusat Statistik, 2020).

2.1.1. Business Process

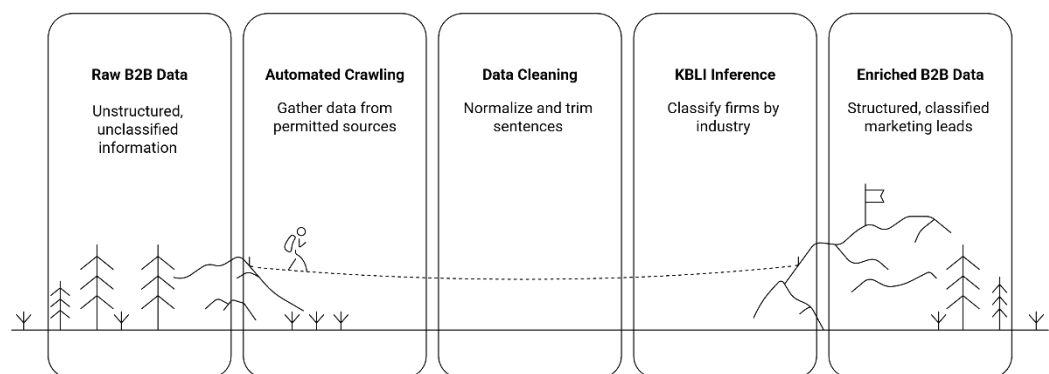


Figure 2. 1 B2B Data Enrichment Process

Lead generation proceeds in stages:

- Automated crawling of target B2B sources permitted by robot's directives and site terms;
- Cleaning and normalisation (language detection, romanisation, sentence trimming);
- **KBLI inference** at five-digit resolution;
- Consolidation across sources at firm level; and
- Hand-off to marketing for outreach.

Within this flow, steps 2–3 are the dominant sources of error in practice: Indonesian text exhibits affixation and code-mixing; firm pages are short, and many lack explicit KBLI mentions. Text classification under such conditions is known to benefit from strong linear baselines (TF-IDF+SVM) and pre-trained transformers adapted to Indonesian (Koto et al., 2020; Manning et al., 2009; Wilie et al., 2020). The pipeline thus requires both a lightweight baseline for breadth and a contextual model for depth.

2.1.2. Problem Detail

KBLI assignment is formalized by BPS as the reference taxonomy for economic activities; KBLI 2020 is the current edition and supersedes KBLI 2015, with an official two-way concordance to preserve comparability (Badan Pusat Statistik, 2021; Badan Pusat Statistik, 2020). However, firm-authored web text is heterogeneous and often under-specified. Indonesian morphology (prefixes, suffixes, infixes, and confixes) degrades sparse vector signals from very short documents unless handled with tokenization/stemming and sub word modelling (Adriani et al., 2007) (Manning et al., 2009). Prior studies that attempt automatic KBLI coding report non-trivial performance gaps at five digits and highlight gains from IndoBERT fine-tuning relative to traditional ML, with meaningful top-k use-cases for operator review (Amnur et al., 2025). These observations motivate the dual-track model selection and the agentic orchestration adopted in this work (Koto et al., 2020; Wilie et al., 2020)

2.2. Text Classification: Foundations and Evolution

2.2.1. Historical Development

Text classification, also known as text categorisation, refers to the task of assigning predefined category labels to textual documents based on their content. This task has a rich history in information retrieval and natural language processing, with applications spanning spam detection, sentiment analysis, topic categorisation, and industry classification (Liu et al., 2021).

The earliest approaches to text classification relied on rule-based systems where domain experts manually specified classification rules. While interpretable and controllable, these systems required substantial effort to develop and maintain, and often failed to generalise beyond the specific patterns anticipated by their designers.

The introduction of machine learning approaches in the 1990s marked a paradigm shift toward data-driven classification. Rather than manually specifying rules, machine learning systems learn classification patterns from labeled training examples. This approach proved more scalable and adaptable, though it introduced new challenges around training data acquisition and model selection.

2.2.2. Feature-Based Classification

The traditional machine learning approach to text classification involves two distinct phases: feature extraction and classification.

Feature Extraction: In the feature extraction phase, textual documents are transformed into numerical representations suitable for machine learning algorithms. The most widely adopted representation is the bag-of-words (BoW) model, wherein a document is represented as a vector of word frequencies or term presence indicators.

The bag-of-words model makes a simplifying assumption that word order and grammar can be ignored for classification purposes. While this assumption discards potentially useful information, the resulting representation has proven remarkably effective for many classification tasks.

The Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme enhances the bag-of-words representation by downweighing terms that appear frequently across the corpus while upweighting terms that are discriminative for particular documents.

The TF-IDF weight for term t in document d is calculated as:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \log(N / \text{DF}(t))$$

where $\text{TF}(t, d)$ is the frequency of term t in document d , N is the total number of documents, and $\text{DF}(t)$ is the number of documents containing term t .

Classification: For the classification phase, various machine learning algorithms have been applied. Naive Bayes classifiers apply Bayes' theorem with strong independence assumptions between features. Decision trees recursively partition the feature space based on discriminative features. k-Nearest Neighbours classifies based on similarity to labelled examples.

Support Vector Machines (SVM) have demonstrated particularly strong performance on text classification tasks (Joachims, 1998). The linear SVM finds a hyperplane that maximally separates documents of different classes in the feature space. The maximum margin principle provides good generalisation properties, while the kernel trick enables non-linear decision boundaries when needed.

2.2.3. Limitations of Feature-Based Approaches

Despite their effectiveness, traditional feature-based approaches exhibit several limitations:

- **Loss of Semantic Information:** The bag-of-words representation discards word order and syntactic structure, potentially losing discriminative information. The sentence "the company manufactures components for computers" and "the computer company manufactures components" would have identical bag-of-words representations despite different meanings.

- **Vocabulary Mismatch:** The reliance on exact lexical matching fails to capture semantic similarity between different words expressing similar meanings. A classifier trained on examples containing "manufacture" may fail to recognise similar activities described using "produce" or "fabricate".
- **Feature Engineering Burden:** Developing effective features requires domain expertise and may not generalise across different classification contexts. Features that work well for one domain may be irrelevant or misleading for another.
- **High Dimensionality:** The vocabulary size of natural language creates high-dimensional feature spaces. With vocabularies of hundreds of thousands of words, the curse of dimensionality can affect classifier performance and training efficiency.

2.2.4. International Classification Standards

Several international standards exist for industry classification:

- **International Standard Industrial Classification (ISIC):** Maintained by the United Nations Statistics Division, ISIC provides a globally harmonised framework with 21 sections, organised hierarchically to 4-digit detail.
- **North American Industry Classification System (NAICS):** Used by the United States, Canada, and Mexico, NAICS provides 20 sectors with 6-digit detail emphasising production processes.
- **Statistical Classification of Economic Activities in the European Community (NACE):** The European Union classification, closely aligned with ISIC, provides 21 sections with 4-digit detail.
- **Klasifikasi Baku Lapangan Usaha Indonesia (KBLI):** Indonesia's national classification, derived from ISIC but with additional Indonesia-specific categories.

2.2.5. KBLI

Klasifikasi Baku Lapangan Usaha Indonesia (KBLI) is the national classification maintained by BPS. KBLI 2020 refines the structure to five digits, aligns to ISIC Rev.4

to four digits, and provides elaborated explanatory notes and the official cross-walk to KBLI 2015 (Badan Pusat Statistik, 2020; Badan Pusat Statistik, 2021). The OSS portal exposes KBLI 2020 for licensing and risk-based business registration, making KBLI integral to compliance and to many public–private datasets (OSS, 2026). For text classification, KBLI notes are particularly valuable as canonical definitions against which candidate descriptions can be matched or retrieved for few-shot exemplars (Badan Pusat Statistik, 2020).

2.2.6. BPS

Klasifikasi Baku Lapangan Usaha Indonesia (KBLI) is the national classification maintained by BPS. KBLI 2020 refines the structure to five digits, aligns to ISIC Rev.4 to four digits, and provides elaborated explanatory notes and the official cross-walk to KBLI 2015 (Badan Pusat Statistik, 2020; Badan Pusat Statistik, 2021). The OSS portal exposes KBLI 2020 for licensing and risk-based business registration, making KBLI integral to compliance and to many public–private datasets (OSS, 2026). For text classification, KBLI notes are particularly valuable as canonical definitions against which candidate descriptions can be matched or retrieved for few-shot exemplars (Badan Pusat Statistik, 2020).

2.2.7. Hierarchical Classification Approaches

The hierarchical structure of industrial classifications suggests classification strategies that exploit this structure:

- **Top-Down Classification:** Predict coarse categories first, then fine categories conditioned on coarse predictions
- **Flat Classification:** Predict fine categories directly, ignoring hierarchy
- **Loss Function Hierarchy:** Flat classification with hierarchy-aware loss that penalises errors proportionally to taxonomic distance.

Research has shown that hierarchical approaches can improve accuracy while providing more interpretable errors that respect taxonomic structure.

2.2.8. Other Classifications

Three related taxonomies are often consulted for interoperability: ISIC Rev.4 as the global reference, NAICS 2022 as the North American system emphasizing digital economy updates, and NACE Rev.2 / Rev.2.1 as the European system with formal methodological notes (United States, 2022). These systems provide methodological guidance (e.g., “principal activity”), hierarchy depth, and examples that inform feature design and evaluation granularity when porting methods across jurisdictions (United States, 2022)

2.2.9. Comparison: KBLI vs ISIC vs NAICS vs NACE

As discussed above, the choice of taxonomy shapes annotation guidelines and error tolerance. Table 2.1 compares purpose, strengths, limits, and other criteria across KBLI, ISIC, NAICS, and NACE to clarify why KBLI remains the target for this thesis while insights from others inform method design (Badan Pusat Statistik, 2024)

Table 2. 1 Comparative overview of KBLI, ISIC, NAICS, and NACE (purpose, strengths, limits, criteria).

Classification System	Purpose	Strengths	Limitations	Core Classification Criteria
KBLI (Indonesia)	National standard for business activity classification and licensing in Indonesia	Official five-digit structure; fully integrated with OSS licensing; formal concordance between KBLI 2015 and 2020 revisions	Indonesia-specific terminology limits direct international comparability	Harmonised with ISIC Rev.4 at the four-digit level to ensure international alignment
ISIC Rev.4 (UN)	Global reference framework	Strong conceptual clarity; high international	Limited country-level granularity;	Methodological guidance based on the principal

Classification System	Purpose	Strengths	Limitations	Core Classification Criteria
	for economic activity classification	comparability; widely adopted as a base classification	requires national adaptations	economic activity of establishments
NAICS 2022	Standard classification system for North America (U.S., Canada, Mexico)	Incorporates digital-economy updates; detailed six-digit granularity; extensive explanatory examples	Region-specific structure reduces applicability outside North America	Six-digit hierarchical depth with periodic sectoral restructuring
	Official statistical classification system of the European Union	Detailed implementation guidelines; technology-intensity groupings; strong statistical harmonisation	EU-centric orientation limits global generalisation	Harmonisation across the European Statistical System (ESS)

Description. The table above 2.1 contrasts the four taxonomies along function and design. It highlights KBLI’s legal centrality in Indonesia, ISIC’s role in global comparability, NAICS’s digital updates, and NACE’s methodological guidance for EU statistics; these differences inform label definitions, training data curation, and evaluation splits (Badan Pusat Statistik, 2024; United States, 2022).

2.3. Distributed Representations and Word Embeddings

2.3.1. The Distributional Hypothesis

The distributional hypothesis, articulated by Harris (1954) and often summarised as "you shall know a word by the company it keeps" (Firth, 1957), provides the theoretical foundation for learning word representations from text corpora. This hypothesis posits that words appearing in similar contexts tend to have similar meanings.

Operationalizing this hypothesis through computational methods enables the automatic learning of semantic relationships from unannotated text. Rather than manually encoding semantic knowledge, representation learning approaches induce this knowledge from patterns of word co-occurrence in large corpora.

2.3.2. Neural Word Embeddings

The development of distributed word representations, also known as word embeddings, marked a significant advance in capturing semantic relationships. Word2Vec, introduced by (Mikolov et al., 2013), proposed two architectures for learning word embeddings: Continuous Bag-of-Words (CBOW) and Skip-gram.

The CBOW architecture predicts a target word from its surrounding context words. The Skip-gram architecture inverts this, predicting context words from a target word. Both architectures learn dense vector representations of words as a byproduct of the prediction task.

GloVe (Global Vectors for Word Representation), introduced by (Pennington et al., 2014), takes a different approach based on matrix factorisation of word co-occurrence statistics. The method combines the benefits of count-based methods (capturing global corpus statistics) with the advantages of prediction-based methods (capturing semantic relationships).

The resulting embeddings position semantically similar words close together in the vector space. Words with related meanings cluster together, and analogical relationships can be expressed as vector arithmetic (e.g., $\text{king} - \text{man} + \text{woman} \approx \text{queen}$).

2.3.3. Fast Text and Sub word Representations

Fast Text, developed by (Joulin et al., 2017), extends word embeddings to incorporate sub word information. Rather than learning a single embedding for each word, fast Text represents words as sums of their constituent character n-grams.

This approach provides several advantages:

- **Morphological Robustness:** Related words with shared morphological structure (e.g., "manufacture" and "manufacturing") have similar representations
- **Out-of-Vocabulary Handling:** Words not seen during training can still be represented through their component n-grams
- **Efficiency:** Simple averaging of n-gram embeddings enables fast inference
- For Indonesian text, the sub word approach is particularly relevant given the agglutinative morphology of the language. Words sharing the same root but differing in affixation receive related representations.

2.4. Transformer-Based Language Models

2.4.1. The Attention Mechanism

The attention mechanism, originally proposed for neural machine translation by (Bahdanau et al., 2016), enables models to selectively focus on relevant parts of the input. Rather than compressing the entire input into a fixed-size representation, attention allows the model to weight different input positions based on their relevance to the current processing step.

Self-attention, where a sequence attends to itself, enables each position to incorporate information from all other positions. This capability proves particularly powerful for capturing long-range dependencies in text that recurrent architectures struggle to model.

2.4.2. The Transformer Architecture

The Transformer architecture, introduced by (Vaswani et al., 2023) in the seminal paper "Attention is All You Need", revolutionised natural language processing. The architecture relies entirely on attention mechanisms without recurrence, enabling parallel processing of input sequences while capturing long-range dependencies.

The Transformer consists of an encoder and decoder, each comprising multiple layers of:

- **Multi-Head Self-Attention:** Allows the model to attend to different representation subspaces
- **Feed-Forward Networks:** Applies non-linear transformations to each position independently
- **Layer Normalisation and Residual Connections:** Stabilises training of deep networks

The parallel processing capability of Transformers enables more efficient training on modern hardware compared to sequential recurrent architectures. This efficiency has enabled training on increasingly large datasets, leading to substantial improvements in model capabilities.

2.4.1. BERT and Pre-training Paradigms

BERT (Bidirectional Encoder Representations from Transformers), developed by Devlin et al. (2019), introduced the paradigm of large-scale pre-training followed by task-specific fine-tuning.

BERT is pre-trained on two unsupervised tasks:

- **Masked Language Modeling (MLM):** Randomly mask 15% of input tokens and predict them from context
- **Next Sentence Prediction (NSP):** Predict whether two sentences are consecutive in the original text

The masked language modeling objective encourages the model to develop deep bidirectional understanding, incorporating both left and right context when predicting

masked tokens. This bidirectional capability distinguishes BERT from left-to-right language models.

Pre-training on massive text corpora (Wikipedia and BookCorpus, totaling over 3 billion words) enables BERT to learn general language representations. These representations capture syntactic patterns (part-of-speech, constituency structure), semantic relationships (synonymy, hypernymy), and world knowledge encoded in the training text.

2.4.3 IndoBERT and Indonesian Language Models

While multilingual BERT (mBERT) includes Indonesian in its pre-training corpus, models specifically pre-trained on Indonesian text have demonstrated superior performance. IndoBERT, introduced by (Koto et al., 2020), was pre-trained on approximately 4GB of Indonesian text from Wikipedia, news articles, and web content.

IndoBERT follows the BERT-base architecture with:

- 12 transformer layers
- 768 hidden dimensions
- 12 attention heads
- Approximately 110 million parameters

Evaluation on the IndoNLU benchmark demonstrated that IndoBERT outperforms mBERT on Indonesian tasks including:

- Sentiment analysis
- Named entity recognition
- Natural language inference
- Question answering

The performance advantage of monolingual models over multilingual alternatives reflects the benefit of focused pre-training on the target language, without the capacity dilution that occurs when a model must represent dozens of languages.

2.4.3. Sentence Embeddings

For classification tasks involving comparison against a taxonomy of predefined categories, sentence embedding approaches offer computational advantages over cross-encoder fine-tuning.

Sentence-BERT (SBERT), developed by (Reimers and Gurevych, 2019), modifies the BERT architecture to produce fixed-size sentence embeddings suitable for semantic similarity computation. The approach uses siamese or triplet network structures during fine-tuning.

Key advantages of sentence embeddings for classification:

- **Efficiency:** Category embeddings can be precomputed once and reused for all queries
- **Scalability:** New categories can be added by computing their embeddings without model retraining
- **Simplicity:** Classification reduces to nearest-neighbor search in embedding space

The paraphrase-multilingual-MiniLM-L12-v2 model extends sentence embeddings to 50+ languages including Indonesian. This model produces 384-dimensional sentence embeddings and has demonstrated strong performance on cross-lingual semantic similarity benchmarks.

2.5. Probability Calibration for Neural Networks

2.5.1. The Calibration Problem

The reliability of classification systems depends not only on accuracy but also on the quality of confidence estimates. A well-calibrated classifier produces confidence scores that accurately reflect the probability of correctness.

Formally, a classifier is calibrated if; $\mathbf{P}(Y = \hat{Y} \mid \mathbf{P}(Y = \hat{Y}) = p) = p$

Among all predictions made with confidence p , approximately p fraction should be correct.

(Guo et al., 2017) demonstrated that modern neural networks, despite achieving high accuracy, tend to be poorly calibrated. Specifically, these models exhibit overconfidence: expressing high confidence even for incorrect predictions.

This miscalibration has several causes:

- **Negative Log-Likelihood Training:** The standard cross-entropy loss encourages confident predictions regardless of correctness.
- **High Capacity:** Large models can memorise training examples, leading to extreme confidence on similar test examples.
- **Regularisation Effects:** Techniques like dropout affect calibration in complex ways.

2.5.2. Calibration Metrics

Expected Calibration Error (ECE) (Guo et al., 2017) provides a scalar summary of calibration quality. Predictions are partitioned into bins based on confidence, and ECE measures the average gap between confidence and accuracy:

$$ECE = \sum_{m=1}^M \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|$$

Where:

- M : The total number of bins used for partitioning the predictions.
- n : The total number of samples available in the dataset.
- B_m : The set of indices of samples whose prediction confidence falls within the m -th bin.
- $|B_m|$: The cardinality of the set B_m , representing the number of samples in bin m .
- $\text{acc}(B_m)$: The accuracy of bin B_m , calculated as the proportion of correct predictions within that specific bin.
- $\text{conf}(B_m)$: The average confidence of bin B_m , calculated as the mean of the predicted probabilities within that specific bin.

2.5.3. Temperature Scaling

Temperature scaling, proposed by Guo et al. (2017), offers a simple yet effective post-hoc calibration method. The technique introduces a single temperature parameter T that scales logits before softmax; $\hat{\mathbf{p}} = \text{softmax}(\mathbf{z} / T)$

Higher temperature values produce softer probability distributions, reducing overconfidence. Lower temperatures sharpen distributions. The temperature is optimised on validation data to minimise negative log-likelihood.

Key properties of temperature scaling:

- **Preserves Ranking:** The predicted class is unchanged; only confidence magnitudes are adjusted
- **Single Parameter:** Minimal overfitting risk compared to multi-parameter methods
- **Post-Hoc Application:** Applied after training without modifying the base model

2.6. Agentic AI Orchestration Systems

2.6.1. From Single-Pass to Iterative Systems

Traditional classification systems operate in a single pass: input is processed, prediction is generated, and processing terminates. Agentic AI systems, in contrast, can engage in iterative processing with multiple steps of reasoning and action.

Agentic systems are characterised by:

- **Autonomous Decision-Making:** The system decides which actions to take based on the current state
- **Tool Use:** The system can invoke external tools or APIs to gather information
- **Self-Correction:** The system can recognise errors and refine its outputs
- **Goal-Directed Behaviour:** Actions are taken to achieve specified objectives.

2.6.2. Agentic Classification Architecture

For classification tasks, agentic architecture involves:

1. **Initial Classification:** Generate prediction with confidence estimate
2. **Confidence Assessment:** Determine if confidence is sufficient for acceptance
3. **Re-Query Decision:** If confidence is low, decide whether additional information could help
4. **Information Gathering:** Execute targeted queries to gather additional context
5. **Aggregation:** Combine initial prediction with additional information for final decision
6. This architecture enables the system to balance efficiency (accepting confident predictions) with accuracy (gathering information for uncertain cases).

2.6.3. Selective Prediction

The concept of selective prediction, also known as prediction with a reject option, provides theoretical foundation for confidence-based routing. In selective prediction, the system abstains from predictions when confidence falls below threshold τ .

Key metrics for selective prediction:

- **Coverage:** Fraction of samples with predictions made (confidence $\geq \tau$)
- **Selective Accuracy:** Accuracy on samples where predictions are made
- **Risk-Coverage Curves:** Visualise trade-off between coverage and accuracy
- The effectiveness of selective prediction depends critically on calibration quality. With well-calibrated confidence, threshold τ directly controls the accuracy-coverage trade-off.

2.7. Research Design

This research adopts a design science research (DSR) framework, which is appropriate for developing and evaluating artefacts intended to solve identified problems (Hevner et al., 2004). The DSR paradigm emphasises the creation of innovative artefacts—in this case, an automated KBLI classification system—and their rigorous evaluation against specified objectives.

The research process follows an iterative design-build-evaluate cycle. In the design phase, system architecture and algorithmic approaches are specified based on literature review and domain requirements. In the build phase, these specifications are instantiated as working software components. In the evaluate phase, the implemented system is assessed against pre-registered performance targets using held-out test data. A Design Science Research (DSR) strategy is adopted because a concrete artefact—a KBLI assignment pipeline driven by agentic AI—is to be constructed and evaluated under realistic constraints. Within DSR, the relevance cycle links the environment (PT DLN’s lead-generation workflows and Indonesian public web data) to problem requirements; the rigor cycle draws on prior knowledge (short-text classification, Indonesian NLP, calibration, hierarchical evaluation); and the design cycle iterates artefact construction and assessment (Hevner et al., 2004).

The method is instantiated through six steps:

Table 2. 2 DSR process mapped to artefact, dataset, and evaluation outputs.

DSR Phase	Implementation in This Thesis
Problem Identification	Public business descriptions are short, noisy, and heterogeneous , often reflecting multiple activities per firm and frequently missing explicit KBLI labels , making reliable five-digit classification difficult
Objectives of a Solution	To produce reliable KBLI five-digit predictions with calibrated confidence scores and actionable textual evidence suitable for internal business decision-making
Design & Development	An agentic AI pipeline consisting of: (1) <i>Planner</i> (task orchestration), (2) <i>Retriever</i> (public web sourcing), (3) <i>Normaliser</i> (text cleaning and language handling), (4) <i>Dual Classifier</i> (lightweight triage + transformer-based semantic model), (5) <i>Calibrator</i> (temperature scaling), and (6) <i>Router</i> (model and confidence-based decision flow)
Demonstration	Applied to a public-web corporate corpus relevant to PT DLN , combined with curated KBLI labels for supervised fine-tuning and validation
Evaluation	Assessed using balanced accuracy, macro-F1, hierarchy-aware error distance, Expected Calibration Error (ECE), Brier score , and

DSR Phase	Implementation in This Thesis
	Coverage@τ, supplemented with ablation studies and statistical significance testing
Communication	Results communicated through management-oriented dashboards for operational insight and structured thesis chapters for academic dissemination

Description. The table summarises how abstract DSR steps are operationalised in this study. Each step is coupled to tangible artefacts (pipeline modules), data (PT DLN corpus), and predefined evaluation outputs, ensuring traceability from problem to metrics (Hevner et al., 2004).

The artefact is understood as a hybrid pipeline: polite public-web acquisition and text normalisation feed a two-track classifier (TF-IDF + Linear SVM; IndoBERT fine-tuning), followed by calibration and thresholded decision routing (auto-accept vs. human-in-the-loop). The design explicitly addresses Indonesian morphology, class imbalance, and label hierarchy while staying compatible with operational needs (balanced accuracy, latency, and interpretability)(Manning et al., 2009; Schick et al., 2023.) (Guo et al., 2017; Joachims, 1998; Koto et al., 2020).

2.8. Why KBLI Matters for PT DLN

For PT DLN, correct five-digit KBLI enables segment able outreach, distributor mapping, and resource allocation across Indonesian provinces. The OSS ecosystem and related taxonomies used by finance/regulatory bodies reinforce the operational need for consistent KBLI usage (Badan Pusat Statistik, 2024; OSS, 2026). Downstream market intelligence published by BPS also depends on consistent KBLI usage, creating an external validation reference for PT DLN’s pipeline (Badan Pusat Statistik, 2024; OSS, 2026).

2.9. Previous Studies

Recent Indonesian and international works address automatic industry coding and adjacent problems. (Amnur et al., 2025) compared SVM/Random Forest to IndoBERT

for five-digit KBLI on real and synthetic data and reported IndoBERT superiority with top-k metrics, deploying a REST API for scale (Amnur et al., 2025). (Silla and Freitas, 2011) implemented NLP pipelines for KBLI coding on provincial datasets and showed feasibility with simple models while flagging data quality issues documented a KBLI text classification pilot as statistical modernisation, highlighting annotation constraints and model governance in official statistics. Beyond Indonesia, (Bhattacharjee et al., 2021) presented hierarchical models to assign NAICS at varying digit depths using public signals (Bhattacharjee et al., 2021). An updated **ensemble approach** for Economic Census self-classification (BEACON) demonstrates production use of text classification with user-facing validation (Dumbacher et al., 2025) (Bhattacharjee et al., 2021; United States, 2022). Complementary research investigates text-based industry classification from websites or filings using BERT-family models and zero-shot variants, offering external baselines for architecture choices (Rizinski et al., 2023).

Table 2. 3 Critical review of prior studies on automatic industry coding.

Reference	Context / Dataset	Methodology	Key Contributions	Relevance to This Study
Amnur et al. (2025)	Indonesian companies mapped to KBLI 5-digit	Classical ML vs. IndoBERT; top-k evaluation; RESTful API	Demonstrated superior top-k accuracy of transformer models and feasibility of API-based deployment	Confirms effectiveness of transformer-based KBLI classification at fine granularity
Silitonga et al. (2023)	Provincial-level KBLI textual data	NLP-based classification with feasibility analysis	Emphasised data quality, annotation consistency, and regional variation	Highlights preprocessing and governance challenges in KBLI datasets
UN ESCAP (2023)	Official statistics pilot projects	Human-in-the-loop annotation;	Addressed annotation standards,	Provides governance rationale for AI-

Reference	Context / Dataset	Methodology	Key Contributions	Relevance to This Study
		governance-first AI	transparency, and institutional governance	driven industry classification
Bhattacharjee (2021)	NAICS hierarchical taxonomy	Multi-level hierarchical ML classification	Validated hierarchical prediction aligned with NAICS structure	Conceptually transferable to KBLI–ISIC hierarchical mapping
U.S. Census Bureau (2024)	Economic Census self-reporting	Ensemble learning (BEACON system)	Combined self-classification and ML ensembles at national scale	Benchmark for large-scale, production-grade industry classification
Jagrič & Herman (2024)	Corporate websites	BERT-based website text classification	Demonstrated strong performance using publicly available web text	Supports website-based enrichment for KBLI inference
Rizinski et al. (2023)	Multinational company descriptions	Zero-shot transformer models	Enabled industry classification without labeled training data	Supports low-supervision strategies for missing KBLI labels
Cuffe & Goldschlag (2019)	Large-scale U.S. firm text corpora	Text mining for NAICS assignment	Identified data scale and coverage requirements	Informs minimum data volume thresholds for KBLI intelligence

Reference	Context / Dataset	Methodology	Key Contributions	Relevance to This Study
Hoberg & Phillips (2016; 2018)	SEC filings	Text-based industry construction	Proposed external, market-driven industry benchmarks	Provides methodological basis for external validation
ASEAN / OJK (2023–2024)	Regional–national sector taxonomies	Cross-taxonomy alignment (ASEAN–OJK–KBLI)	Demonstrated operational interoperability of sector standards	Supports KBLI as a hub taxonomy for regional integration
This Thesis (2026)	Internal enterprise B2B datasets linked to KBLI	Agentic AI–based pipeline integrating web scraping, NLP (transformer-based classification), hierarchical KBLI–ISIC alignment, rule-based validation, and human-in-the-loop review	Proposes an end-to-end, autonomous KBLI market intelligence framework for business data enrichment, emphasizing deployment readiness, and internal usability	Extends prior work by shifting from standalone models to a production-oriented, agentic KBLI intelligence system tailored for internal enterprise decision-making

To further clarify the positioning of this research, Table 2.3 contrasts the framework developed in this thesis against frameworks and methodologies from seminal prior studies. While previous works primarily focused on feasibility studies or static classification tasks, this thesis introduces a novel active agentic loop and explicit trust calibration mechanisms designed specifically for production-grade B2B intelligence.

Table 2. 3. Comparison of Framework Features and Contributions							
Feature / Aspect	Current Thesis Work (2026)	Amnur et al. (2025)	Silitonga et al. (2023)	UN ESCAP (2023)	U.S. Census / Bhattacharjee (2021)	Jagrić & Herman (2024)	
Focus / Domain	KBLI 2020 (5-digit) for B2B Market Intelligence (PT DLN)	KBLI (5-digit) for General Classification	KBLI for Provincial Statistical Services	ISIC/KBLI for Official Statistics Pilot	NAICS (Hierarchical) for Economic Census	General Industry / Web Classification	
Core Framework	Agentic AI Pipeline (Hybrid + Re-query Loop)	Supervised Learning (Deep Learning vs ML)	Standard NLP Pipeline	Human-in-the-Loop Governance	Hierarchical ML Ensemble	Transfer Learning (BERT)	
Key Models	Hybrid: TF-IDF+SVM + IndoBERT (Fine-tuned)	IndoBERT vs SVM/Random Forest	Basic ML / NLP methods	Pilot / Rule-based / Manual assist	Hierarchical / 1 classifiers / Ensembles	BERT-based architectures	
Data Source	Public Data (noisy, short, mixed)	Web Data (noisy, code-mixed)	Curated / Synthetic Datasets	Provincial Government Data	Official Statistical Registers	Census responses / Admin Data	Corporate Websites
Orchestration	Agentic query: Triggers new search when confidence < τ	Re-pipeline (Single pass)	Fixed pipeline	Static processing	Manual review workflow	Batch processing	Static classification
Reliability	Calibrated (Platt/Temp)	Top-k Accuracy focus	Standard	Quality Assurance focus	Hierarchical 1	Standard Performance Metrics	

Feature / Aspect	Current Thesis Work (2026)	Amnur et al. (2025)	Silitonga et al. (2023)	UN ESCAP (2023)	U.S. Census / Bhattacharjee (2021)	Jagrič & Herman (2024)
	scaling) with Coverage@τ		Accuracy		Consistency	
Evaluation	Hierarchy-aware error, ECE, Business-centric metrics	F1-score, Accuracy, Top-k	Feasibility, Accuracy	Governance & Transparency	Multi-level & Accuracy	Precision / Recall
Deployment	Production Pipeline with Decision Routing (Automatic vs Review)	REST API prototype	Feasibility Study	Pilot Project	Production (BEACON system)	Research / Experimental
Main Contribution	Autonomous data enrichment with error recovery (agentic loop) and trust calibration.	Benchmarking Transformers vs Classical ML for KBLI.	Validating NLP feasibility for regional KBLI data.	Modernizing statistical coding with AI governance.	Validating hierarchical approaches for NAICS.	Demonstrating BERT effectiveness on web text.

As shown in Table 2.3, the primary differentiator of this work is the shift from static, single-pass classification to an agentic workflow that actively recovers from low-confidence states by seeking new evidence. Furthermore, the integration of hierarchy-aware metrics and reliability calibration addresses the specific operational risks of automated decision-making in a way that standard accuracy-focused baselines do not. This addresses the gap identified in previous feasibility studies regarding the 'missing KBLI' problem in sparse public data.

Description. The table consolidates methods (ML, transfer learning, zero-shot), datasets (KBLI-annotated official/crawled, NAICS labelled business records), and key

outcomes (top-k performance, deployment readiness). It informs model selection, annotation strategy, and evaluation granularity in this thesis (Amnur et al., 2025),(Bhattacharjee et al., 2021; United States, 2022) Method Selection: Indicators and Justification.

The selection of TF-IDF+SVM, fastText, and IndoBERT is justified by: (i) document length (very short firm descriptions), (ii) Indonesian morphology, (iii) label hierarchy depth, (iv) probability calibration needs for business rules, and (v) compute/latency constraints for large-scale triage. Sparse linear models excel with short text and support calibrated probabilities via Platt/temperature scaling for thresholded auto-acceptance (Joachims, 1998; Manning et al., 2009) . fastText adds subword robustness at low cost (Joulin et al., 2017). IndoBERT fine-tuning offers context sensitivity and demonstrably improves five-digit accuracy in Indonesian corpora (Amnur et al., 2025; Koto et al., 2020; Wilie et al., 2020). A hierarchy-aware reading of errors follows best practice to avoid over-penalising near-misses at parent levels (Silla and Freitas, 2011).

Table 2. 4 Model options vs indicators (short text, morphology, hierarchy, calibration, cost).

Model	Core Strengths	Handling of Indonesian Language	Label Structure	Probability Calibration	Computational Cost
Characteristics					
TF-IDF + Linear SVM	Strong performance on short and sparse text; highly interpretable; stable decision boundaries	Morphological variation handled via stemming and character/word n-grams	Flat (single-level) labels	Excellent when combined with Platt scaling or temperature scaling	Lowest (CPU-friendly, fast training and inference)

Model	Core Strengths	Handling of Indonesian Language Characteristics	Label Structure	Probability Calibration	Computational Cost
fastText	Robust subword embeddin gs; strong for short text and out-of-vocabular y (OOV) terms; simple deployme nt	Subword modeling naturally captures Indonesian affixes and variations	Flat (single-level) labels	Softmax probabilities; improved with temperature scaling	Very low (lightweight, production-efficient)
IndoBERT	Rich contextual semantics ; superior handling of paraphras e, ambiguity , and polysemy	Context-aware token representations tailored for Indonesian	Adaptable to hierarchic al KBLI via multi-head or level-wise training	Requires explicit calibration for reliable confidence scores	Highest (GPU-dependent, higher training and inference cost)

Description. The table 2.3 above maps each candidate to the indicators that matter in PT DLN’s setting. It motivates a two-stage design (baseline triage + contextual refinement) with explicit calibration for decision thresholds (Guo et al., 2017; Joachims, 1998).

2.10. Evaluation Considerations

Beyond micro/macro-F1, the study reports **top-k** ($k \in \{1, 3, 5\}$) to reflect reviewer workflows and **hierarchy-aware** metrics to avoid overstating harm of near-misses; balanced accuracy and confidence intervals are used to address class imbalance and variance (Brodersen et al., 2010; Silla and Freitas, 2011). Because transformer outputs can be over-confident, **temperature scaling** is applied on validation splits (Guo et al., 2017). The KBLI 2015↔2020 concordance supports sensitivity checks for label drift (Badan Pusat Statistik, 2024).

2.11. Challenges of Automated KBLI Classification

Automated KBLI classification presents several domain-specific challenges that distinguish it from standard text classification tasks:

- 1. Extreme granularity:** The KBLI 2020 taxonomy contains 1,497 five-digit subclasses. This extreme multiclass setting exceeds typical text classification benchmarks by an order of magnitude, creating severe data sparsity where many classes have very few or no training examples.
- 2. Hierarchical ambiguity:** Many business activities span multiple KBLI categories. For example, a company described as 'produsen dan distributor bahan kimia' (chemical manufacturer and distributor) could be classified under manufacturing (20111) or wholesale trade (46691). The correct assignment depends on the company's primary activity, which may not be clear from short descriptions.
- 3. Indonesian language complexity:** Business descriptions exhibit code-mixing between Indonesian and English, informal abbreviations (e.g., 'PT' for Perseroan Terbatas), and domain-specific jargon. The agglutinative morphology of Indonesian (e.g., 'memperdagangkan', 'perdagangan', 'pedagang' all derive from 'dagang') requires models capable of capturing morphological relationships.
- 4. Short and noisy descriptions:** Company registration descriptions are typically 10–50 words long, providing limited context for fine-grained classification. Many descriptions contain boilerplate text, addresses, or product lists rather than clear activity descriptions.

5. Class imbalance: The distribution of businesses across KBLI categories is highly skewed. Manufacturing (Section C) and wholesale/retail trade (Section G) dominate, while categories such as mining (Section B) and utilities (Section D) have minimal representation in typical business databases.

6. Semantic overlap between categories: Closely related KBLI codes (e.g., 46411 Wholesale of textiles vs. 47711 Retail of textiles) differ only in the business model (wholesale vs. retail), not in the product domain. Standard semantic similarity measures may fail to distinguish these codes because the textual descriptions of the categories themselves are nearly identical.

2.12. Research Gap and Contribution

The literature review reveals several gaps that this research addresses:

- **Indonesian KBLI Classification:** While IndoBERT has been evaluated on general NLP benchmarks, systematic evaluation for KBLI classification is lacking
- **Calibration for Indonesian Text:** Calibration methods have been studied primarily for English; their effectiveness for Indonesian industrial text is unknown
- **Agentic Classification Design:** The combination of calibrated confidence with agentic re-query for industry classification is novel
- **Hierarchy-Aware Evaluation:** Standard metrics fail to distinguish between minor and major classification errors

This research addresses these gaps through systematic development and evaluation of a calibrated, agentic classification system for Indonesian KBLI codes.

2.13. Chapter Summary

To close, the methods and frameworks are distilled into selection indicators for the final research design. As discussed above and formalised below, Table 2.4 functions as the chapter's decision artefact that will be referenced again in Chapter 3 when composing the research framework and success metrics (Badan Pusat Statistik, 2024; Guo et al., 2017; Koto et al., 2020; Wilie et al., 2020).