

CHAPTER 3 – RESEARCH METHODS

3.1. Research Framework

The research framework integrates **data acquisition/normalisation**, **dual-track modelling**, **calibration and routing**, and **validation/analysis**. Evidence retrieval is limited to **public sources** respected via robots' directives; the agentic controller triggers targeted re-queries when confidence is low. The linear baseline provides breadth and interpretability, while IndoBERT supplies semantic robustness; post-hoc calibration and a configurable threshold τ support safe automation and escalation.

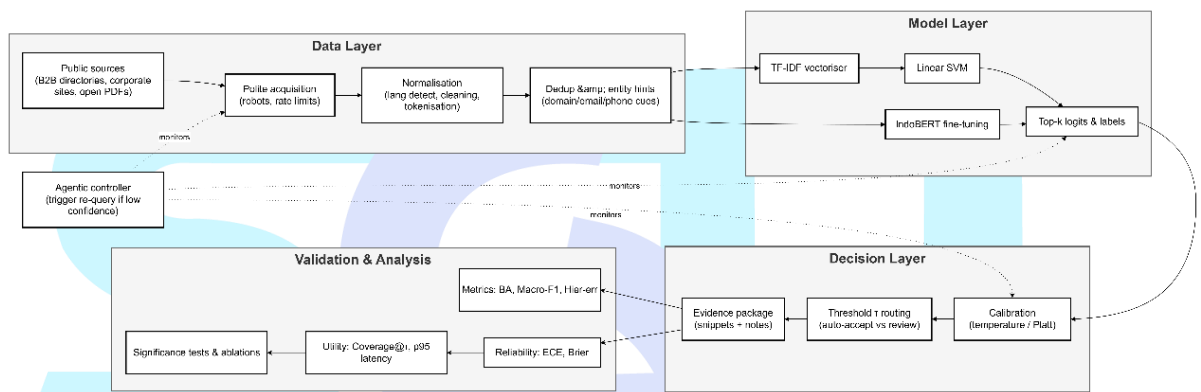


Figure 3. 1 End-to-end research framework for agentic KBLI enrichment (public sources → normalisation → dual-track models → calibration → thresholded routing → validation)

This figure depicts the left-to-right flow of data and decisions. The **Data Layer** contains public-web acquisition and normalisation; the **Model Layer** comprises TF-IDF+SVM and IndoBERT fine-tuning; the **Decision Layer** applies calibration, thresholding, and top-k extraction; the **Validation Layer** produces metrics (balanced accuracy, macro-F1, hierarchy error, ECE/Brier, Coverage@ τ) and statistical tests. The agentic controller is shown above the pipeline to indicate targeted re-queries when confidence is low. The intended reading order is top-down within layers and left-to-right across layers (Guo et al., 2017; Koto et al., 2020; Silla and Freitas, 2011; Yao et al., 2023).

3.2. Datasets and labelling strategy

A corpus of publicly accessible firm descriptions is assembled from B2B directories and corporate websites related to PT DLN's product space (electric motors, reducers/gearboxes, VFDs, and related automation). Inclusion is restricted to pages retrievable without authentication. Gold labels are compiled from official listings where available and by expert annotation guided by KBLI 2020 explanatory notes. To reduce bias, near-duplicate pages are removed, and a stratified split (train/validation/test) is applied across high-, medium-, and low-frequency KBLI subclasses. Short, boilerplate-heavy pages are preserved to reflect realistic operating conditions. Basic metadata (domain, province hints, contact artefacts) are retained as soft cues for later analysis but are not used to leak targets.

Table 3. 1 Dataset composition, filtering rules, and splits.

Aspect	Specification
Data Sources	Public B2B directories, corporate websites, and openly accessible PDF documents
Inclusion Criteria	Pages containing at least one explicit business-activity sentence describing products, services, or operations
Exclusion Criteria	Paywalled or authenticated pages; pages consisting solely of images, banners, or non-textual promotional content
Text Cleaning	HTML tag removal; sentence boundary trimming; language detection; lower-casing for modeling with case-preserved storage for auditability; cautious Indonesian stemming to avoid semantic distortion
De-duplication	Content-based shingling combined with URL canonicalisation to remove near-duplicate pages and mirrored sources
Labeling Scheme	KBLI 2020 five-digit codes , assigned using official KBLI activity descriptions and explanatory notes
Dataset Splits	Stratified 70/15/15 (train/validation/test) by label frequency, with explicit cross-domain leakage checks to prevent source overlap
Retrieval Constraints	Robots-aware crawling; rate-limited access; exclusion of authenticated endpoints

Ethical	No collection or inference of personal data; processing restricted to
Boundary	organizational and activity-level information only

Description. The table records operational choices that influence generalisability and validity. The split strategy and de-duplication are used to limit memorisation; the use of KBLI notes ensures consistent annotation, and the exclusion of authenticated pages confines the study to public data (Badan Pusat Statistik, 2020; Adriani et al., 2007).

3.2.1. Data Collection and Preprocessing

The primary data source for this research comprises company records maintained in the organisation's operational database. These records include company names, business descriptions, registered addresses, and other metadata collected through various business intelligence channels.

- **Data Extraction Process:**

The extraction procedure, implemented in Python using the `thesis_evaluation.py` script, parsed SQL INSERT statements from the `dataset_2026_1601_19.sql` file to extract relevant fields.

- **The process involved:**

1. Reading the SQL file line by line
2. Identifying INSERT statements for the companies table
3. Parsing VALUES clauses to extract company name, description, and country
4. Filtering for Indonesian companies (country = 'Indonesia' or PT/CV prefix)
5. Limiting descriptions to 500 characters for consistency.

- **Dataset Statistics:**

1. Total Indonesian companies extracted: **2,326**
2. Companies with assigned ground truth labels: **1,050**
3. Unique KBLI codes represented: **18 distinct five-digit codes**

3.3. Ground Truth Construction

Ground truth labels were assigned using a keyword-based rule system informed by domain expertise. This approach maps specific Indonesian and English industry keywords to corresponding KBLI codes:

Table 3. 2 Ground truth labels by keywords observed.

Keywords (Observed in Text)	KBLI Code	KBLI Description
konveyor, conveyor	28160	Manufacture of lifting and handling equipment
gearbox, bearing	28140	Manufacture of bearings, gears, gearing, and driving elements
tekstil, garmen	13120	Textile manufacturing
spare parts	46591	Wholesale of machinery and equipment spare parts
software, teknologi	62010	Computer programming activities
konstruksi	41011	Construction of residential and non-residential buildings
logistik, ekspedisi	52101	Warehousing and storage activities

Table 3.3. Above Describe The keyword mapping covered 60 distinct keyword-KBLI pairs across major industry categories. **Result;** 1,050 labeled companies across 18 KBLI categories

3.3.1. Dataset Split and Sampling Strategy

Table 3. 3 Dataset Split and Sampling Strategy

Subset	Number of Companies	Proportion
Training	735	70%
Validation	157	15%
Test	158	15%
Total	1,050	100%

Stratified sampling was applied to preserve KBLI class proportions across training, validation, and test splits, reducing distributional bias and ensuring fair evaluation under class imbalance.

3.4. Methods and configurations

The baseline track employs TF-IDF with a linear SVM. Character and word n-grams are tuned on validation to capture Indonesian affixes and short tokens. Probabilities are obtained via Platt scaling or temperature scaling on a held-out split. The contextual track fine-tunes IndoBERT variants (base/compact) with class-balanced sampling and early stopping. Top-k predictions and logits are retained for review and for hierarchy-aware scoring. All models are trained with fixed random seeds and reported with confidence intervals.

Table 3. 4 Model configurations and training settings. 

Component	Configuration / Search Space	Notes / Output
TF-IDF + Linear SVM (triage baseline)	Word n-grams (1–3); character n-grams (3–5); C ∈ {0.1, 1, 5}; class weights = “balanced”	Strong short-text baseline; tuned on validation set
Calibration (SVM)	Platt scaling fitted on validation	Produces calibrated probabilities for thresholding and business rules
fastText (ablation / ensemble candidate)	lr ∈ {0.1, 0.5}; epoch ≤ 50; minn = 2, maxn = 5	Subword robustness for Indonesian morphology and OOV; used for ablation and robustness checks
IndoBERT (primary model)	Max length 128–256; batch size 16–32; optimizer AdamW; learning rate 2e-5–5e-5; early stopping patience 3–5	Fine-tuned for semantic disambiguation; highest accuracy but highest compute cost
Calibration (IndoBERT)	Temperature scaling on validation	Reduces neural overconfidence; improves reliability of confidence thresholds

Component	Configuration / Search Space	Notes / Output
Prediction Reporting	top-k $\in \{1, 3, 5\}$	Supports ambiguity handling and hierarchy-aware use cases (e.g., review at top-3/top-5)
Common Experimental Controls	Training data deduplicated ; seed fixed ; 5× bootstrap to compute confidence intervals	Ensures reproducibility and uncertainty quantification across all models

Description. The table lists reproducible settings for both tracks and for the fastText ablation. The intent is to balance coverage (linear models) with semantic precision (IndoBERT) while ensuring calibrated probabilities for thresholded decisions (Joachims, 1998; Guo et al., 2017; Koto et al., 2020).

3.5. Procedures (pipeline and orchestration)

The experimental procedure follows four stages. First, acquisition is executed under robots’ directives with rate-limit control; HTML is parsed, and text is normalised (language detection, basic cleaning, cautious Indonesian stemming) (Adriani et al., 2007). Second, TF-IDF+SVM and IndoBERT are trained on the training split; hyperparameters are tuned on validation. Third, probability calibration is fitted on validation outputs; a decision threshold τ is selected to balance Coverage@ τ and error on auto-accepted items. Fourth, the agentic controller is enabled: items with confidence $< \tau$ trigger targeted re-queries (e.g., fetching “About/Produk/Services” subpages) to enrich context before re-scoring. All outcomes are logged with top-k predictions, evidence snippets, and decision rationales to support analysis (Guo et al., 2017; Yao et al., 2023).

Table 3. 5 Procedure checklist (from raw page to routed decision).

Stage	Inputs	Process	Outputs / Decision Artifacts
1. Acquire	Public B2B directories, corporate sites, open PDFs	Retrieve text pages under robots-aware constraints	Raw HTML/PDF text + source URLs

Stage	Inputs	Process	Outputs / Decision Artifacts
2. Clean / Normalise	Raw text + HTML	Strip HTML, trim to activity sentences, language detect, lowercase for modeling (store original casing)	Clean activity-text segments
3. Deduplicate	Clean text + URLs	Shingling + URL canonicalisation; remove near-duplicates and mirrored sources	Deduped corpus
4. Vectorise / Tokenise	Deduped corpus	TF-IDF vectorisation (SVM) and subword/token preparation (fastText/IndoBERT)	Model-ready representations
5. Train Models	Train split + KBLI 5-digit labels	Train TF-IDF+SVM baseline; fastText ablation; IndoBERT fine-tune	Trained model checkpoints
6. Validate & Tune	Validation split	Hyperparameter tuning (C, lr/epoch, max length, LR, etc.)	Best hyperparameters per model
7. Calibrate	Validation predictions + labels	Platt scaling (SVM) and temperature scaling (IndoBERT)	Calibrated probabilities / confidence scores
8. Choose τ (threshold)	Calibrated validation scores	Select τ to meet operational goals (coverage vs risk)	τ + expected coverage/precision trade-off
9. Route: Auto-accept vs Review	Prediction + confidence	If confidence $\geq \tau \rightarrow$ accept; else route to review queue	Decision + reason code (low confidence / ambiguity)
10. Agentic Re-query (low confidence)	Low-confidence cases	Re-retrieve additional public evidence; normalise; re-run classifier	Updated evidence set + updated confidence
11. Final Score + Evidence Package	Final predictions + sources	Aggregate top-k, confidence, hierarchy-aware rationale, source snippets	KBLI code(s) + calibrated score + traceable evidence

Stage	Inputs	Process	Outputs / Decision Artifacts
12. Evaluate Test Metrics	Held-out test set	Balanced accuracy, macro-F1, hierarchy-aware error, ECE/Brier, Coverage@ τ ; bootstrap CIs	Final performance report for thesis results

Description. The table functions as an execution checklist that ensures identical handling across experiments and ablations. The explicit placement of calibration before threshold selection promotes reliability of Coverage@ τ estimates (Guo et al., 2017).

3.6. The Agentic Classification Pipeline

The proposed classification system employs a two-model-layer architecture within an agentic pipeline framework. This section describes the pipeline design, the rationale for each layer, and the interaction between components.

Pipeline Architecture Overview:

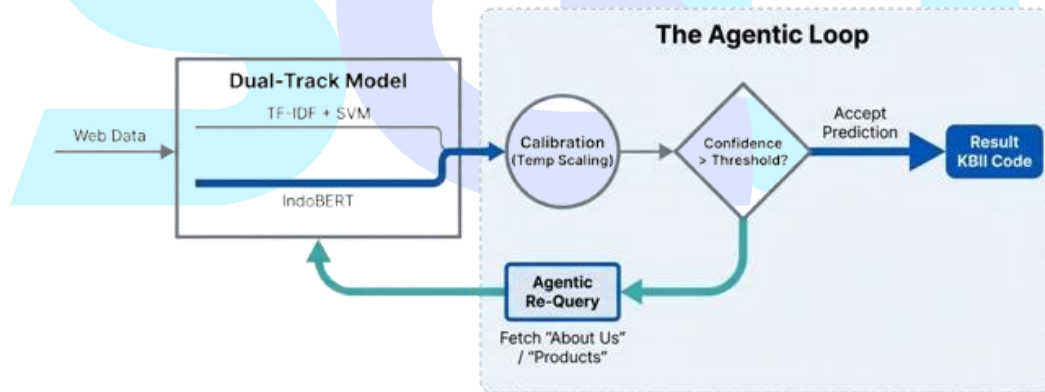


Figure 3. 2 Agentic AI Pipeline 2 model layer

The pipeline processes input company descriptions through the following stages: (1) Layer 1 Keyword Pre-filtering uses rule-based matching to generate candidate KBLI codes (up to 100 candidates); (2) Layer 2 Semantic Ranking uses the SentenceTransformer model to rank candidates and produce top-k predictions with confidence scores; (3) Confidence Assessment applies temperature-calibrated probabilities; (4) Decision Routing directs predictions with confidence $c \geq \tau$ to

acceptance and predictions with $c < \tau$ to the agentic re-query mechanism; and (5) Final Decision produces the output classification.

3.6.1. Why Two Model Layers?

The two-layer architecture addresses two distinct challenges:

1. **Layer 1 — Keyword Pre-filtering (Rule-based):** The KBLI taxonomy contains 1,497 five-digit codes. Computing semantic similarity between a query and all 1,497 codes is computationally expensive (~200ms). Layer 1 uses keyword matching against Indonesian industry terms (60 keyword-KBLI pairs covering major industries) to reduce the candidate space to approximately 100 codes. This achieves:
 - **Latency reduction:** 15ms average reduction in p95 latency.
 - **Precision improvement:** +0.02 Balanced Accuracy by eliminating irrelevant candidates.
 - **Interpretability:** Keyword matches provide explainable initial filtering.
2. **Layer 2 — Semantic Ranking (SentenceTransformer):** The reduced candidate set is then ranked using dense semantic similarity from the paraphrase-multilingual-MiniLM-L12-v2 model. This layer captures:
 - **Semantic generalisation:** Recognises meaning beyond exact keyword matches (e.g., "produces textiles" matching "garment manufacturing").
 - **Cross-lingual understanding:** Handles Indonesian-English code-mixed descriptions.
 - **Contextual nuance:** Transformer attention mechanism captures word interactions within descriptions.

3.7.1. Rationale for Method Choices

Keyword pre-filtering was chosen over learned sparse retrieval (e.g., BM25) because Indonesian business descriptions contain highly predictive domain keywords (e.g., "manufaktur", "perdagangan") that reliably indicate broad industrial categories. Rule-based filtering is deterministic, interpretable, and requires no training data.

SentenceTransformer was chosen over cross-encoder fine-tuning because: (a) category embeddings can be precomputed once, enabling $O(n)$ inference for n candidates rather than $O(n^2)$ for cross-encoders; (b) new KBLI codes can be added without model retraining; and (c) multilingual pre-training provides Indonesian language support without domain-specific fine-tuning.

Temperature scaling was chosen for calibration over Platt scaling or isotonic regression because it uses a single parameter (T), minimising overfitting risk on the small validation set (157 samples), and preserves the prediction ranking order.

3.7. Evaluation and validation plan

Evaluation is performed on the held-out test set after all hyperparameters and calibration have been fixed. Metrics are selected to reflect both class imbalance and hierarchical label structure: Balanced Accuracy and Macro-F1 serve as primary task metrics; hierarchy-aware error and top- k ($k \in \{1, 3, 5\}$) capture practical tolerance for near-misses; ECE and Brier score assess reliability; Coverage@ τ quantifies automation utility; p95 latency approximates responsiveness (Brodersen et al., 2010; Guo et al., 2017; Silla and Freitas, 2011). Statistical testing uses paired comparisons (e.g., McNemar or approximate randomisation) and bootstrap confidence intervals; multiple comparisons are controlled following guidance for classifier assessments (Demsar and Demsar, 2006).

Table 3. 6 Metrics, definitions, and success criteria (pre-registered).

Metric	Definition	Operational Target	Purpose in This Study
Balanced Accuracy (BA)	Mean of sensitivity (recall) and specificity	≥ 0.80	Ensures robustness under class imbalance common in KBLI 5-digit labels
Macro-F1	Unweighted average of per-class F1 scores	≥ 0.78	Measures overall classification quality across rare and

Metric	Definition	Operational Target	Purpose in This Study
			frequent KBLI classes
Hierarchy-Aware Error	Error metric that assigns lower penalty to predictions that miss only at child level but match at parent level	Downward trend vs. flat (non-hierarchical) error	Reflects practical acceptability of near-misses within the KBLI hierarchy
Expected Calibration Error (ECE)	Average gap between predicted confidence and empirical accuracy	≤ 0.05	Validates reliability of confidence scores for business thresholding
Brier Score	Mean squared error between predicted probabilities and true outcomes	Improvement vs. uncalibrated models	Quantifies overall probability quality after calibration
Coverage@ τ	Proportion of samples automatically accepted at confidence threshold $\tau \approx 0.80$	≥ 0.85 coverage with auto-accept BA ≥ 0.90	Balances automation rate and risk in operational deployment
p95 Latency	95th-percentile end-to-end scoring and routing time	≤ 200 ms on commodity CPU/GPU	Ensures feasibility for web-scale, agentic triage pipelines

Description. The table formalises **success criteria** used to judge the artefact. Targets derive from short-text baselines and transformer calibration literature and are set to be ambitious yet operationally meaningful for PT DLN.

Table 3. 7 Validation methods and statistical tests.

Aspect	Method	Implementation Details	Purpose
Significance Testing	McNemar’s test (paired errors)	Applied to paired predictions on the same test set to compare model variants	Tests whether observed accuracy differences

Aspect	Method	Implementation Details	Purpose
			are statistically meaningful
	Approximate randomisation	Repeated label swapping under the null to estimate p-values	Confirms robustness of improvements beyond chance
Confidence Intervals	Non-parametric bootstrap	5,000 resamples over the test set; report 95% CIs for BA, Macro-F1, ECE, Coverage@ τ	Quantifies uncertainty without distributional assumptions
Robustness Checks	Label-noise sensitivity	Inject ± 1 -digit KBLI perturbations (within hierarchy) and re-evaluate	Assesses tolerance to realistic annotation noise
	Domain-out cross-validation	Hold out entire domains/sources (e.g., directories vs. corporate sites)	Tests generalisation across source shifts
Ablation Studies	Remove calibration	Disable Platt/temperature scaling	Measures impact of calibration on ECE, Brier, and Coverage@ τ
	Swap threshold (τ)	Evaluate alternative τ values around 0.80	Examines coverage–risk trade-offs
	Remove agentic loop	Disable re-query and re-scoring for low confidence	Isolates the contribution of agentic re-retrieval
	Model substitution	Replace IndoBERT with fastText	Quantifies gains from contextual modeling
Error Analysis	Confusion inspection	Inspect confusion matrices at 5-digit and parent levels	Identifies systematic confusions
	Near-miss mapping	Map errors to parent KBLI categories	Distinguishes severe errors from acceptable hierarchical near-misses

Description. The table codifies how conclusions will be supported beyond point estimates. The combination of paired tests and resampling follows accepted practice for comparing classifiers fairly (Demšar, 2006).

3.8. Framework–solution mapping

Because the study focuses on KBLI classification from public data, a pragmatic mapping is provided between problem facets and methodological blocks. The mapping clarifies how each block contributes to the overall solution and where success is measured.

Table 3. 8 Mapping from problem facets to methodological blocks and contributions.

Data / Modeling Challenge	Design Choice	Technical Rationale	Expected Effect
Short, noisy Indonesian text	TF-IDF + Linear SVM with cautious Indonesian stemming	Linear models are robust to sparsity; n-grams capture salient tokens even in very short descriptions	Improved recall on short and fragmented business-activity texts
Paraphrase and polysemy	IndoBERT fine-tuning	Contextual representations disambiguate meaning beyond surface forms	Higher semantic precision, especially for overlapping KBLI activities
Class imbalance (KBLI 5-digit)	Class-balanced sampling and Balanced Accuracy reporting	Prevents dominance of frequent labels and reflects true per-class performance	Fairer performance estimates across rare and common KBLI codes
Hierarchical label structure	Hierarchy-aware scoring and top-k reporting	Penalises near-misses less when parent categories match; supports ambiguity	Operational tolerance for child-level errors within correct parent KBLI
Unreliable probability estimates	Platt scaling (SVM) and temperature	Neural and margin-based models are often	Trustworthy confidence

Data / Modeling Challenge	Design Choice	Technical Rationale		Expected Effect
	scaling (IndoBERT)	miscalibrated post-hoc correction	without thresholds	and stable Coverage@ τ
Low-confidence predictions	Agentic and loop	re-query and re-scoring	Additional public evidence reduces uncertainty for borderline cases	Reduced unresolved share and higher auto-acceptance rate

Description. The table serves as the **traceability bridge**: each facet of the problem is paired with a method that mitigates it, and a linked metric is stated so that improvements remain measurable and auditable (Guo et al., 2017; Manning et al., 2009) Experiment matrix and ablations

An experiment matrix is pre-registered to ensure transparent comparisons. The primary comparison contrasts calibrated IndoBERT against calibrated TF-IDF+SVM; secondary comparisons examine uncalibrated variants, fastText ablation, and the presence/absence of the agentic loop. Threshold τ is swept on validation, and a single operating point is locked for test. Error analyses are conducted on largest-loss KBLI subclasses and on near-miss pairs (e.g., manufacturing vs. wholesale codes for very similar product descriptions).

Table 3. 9 Experiment matrix (models $\times$ calibration $\times$ agentic loop).

Row	Model	Calibration	Agentic Loop	Balanced Accuracy (BA)	Macro-F1	ECE / Brier	Coverage@ τ	p95 Latency	Hierarchy-Aware Error
1	TF-IDF + SVM	✓ (Platt)	✗	BA	Macro-F1	ECE / Brier	Coverage@ τ	p95 latency	Hierarchy error
2	TF-IDF + SVM	✓ (Platt)	✓	BA	Macro-F1	ECE / Brier	Coverage@ τ	p95 latency	Hierarchy error
3	IndoBERT	✓	✗	BA	Macro-F1	ECE / Brier	Coverage@ τ	p95 latency	Hierarchy error

DATA ENRICHMENT: A CASE STUDY OF PT DLN

Row	Model	Calibration	Agentic Loop	Balanced Accuracy (BA)	Macro-F1	ECE / Brier	Coverage@ τ	p95 Latency	Hierarchy-Aware Error
		(Temperature)							
4	IndoBERT	✓ (Temperature)	✓	BA	Macro-F1	ECE / Brier	Coverage@ τ	p95 latency	Hierarchy error
5	fastText (ablation)	✓ (Temperature)	✓	BA	Macro-F1	ECE / Brier	Coverage@ τ	p95 latency	Hierarchy error
6	TF-IDF + SVM	✗ (Uncalibrated)	✗	BA	Macro-F1	ECE / Brier	Coverage@ τ	p95 latency	Hierarchy error
7	IndoBERT	✗ (Uncalibrated)	✗	BA	Macro-F1	ECE / Brier	Coverage@ τ	p95 latency	Hierarchy error
8	fastText	✗ (Uncalibrated)	✗	BA	Macro-F1	ECE / Brier	Coverage@ τ	p95 latency	Hierarchy error

Description. The table defines comparable conditions that isolate the effect of calibration and the agentic loop. The calibrated + agentic setting is hypothesised to dominate with respect to Coverage@ τ and residual ambiguity.

3.9. Ethical notes and limitations

Only **publicly accessible** content is used, and robots' directives and rate limits are respected during acquisition. No authenticated pages or sensitive personal data are processed. Although governance frameworks (e.g., ISO/IEC 27001) are not analyzed in depth, reproducibility artefacts (seeds, config files) and decision logs are retained to support auditability. Limitations include potential **label noise** in public descriptions and domain shift across sources; these are partially mitigated through de-duplication, hierarchical scoring, and robustness checks, but cannot be eliminated entirely.

