

CHAPTER 1 - INTRODUCTION

1.1. Research Background

1.1.1. The Importance of Business Classification in Economic Administration

The classification of business activities constitutes a fundamental pillar in the administration and analysis of economic systems worldwide. In Indonesia, the Klasifikasi Baku Lapangan Usaha Indonesia (KBLI, 2020), which translates to the Indonesian Standard Industrial Classification, serves as the official framework for categorising business entities according to their primary economic activities (BPS, 2020). This classification system, which was most recently updated in 2020, encompasses more than 1,500 distinct five-digit codes organised hierarchically across 21 sections, representing the complete spectrum of economic activities conducted within the Indonesian economy.

The significance of accurate business classification extends far beyond mere administrative convenience. From a regulatory perspective, the correct assignment of KBLI codes determines which government regulations, licensing requirements, and tax obligations apply to a particular business entity. For economic analysis, aggregated classification data enables policymakers to understand sectoral composition, track industry trends, and formulate evidence-based policies. Furthermore, in the context of international trade and investment, accurate classification facilitates comparability with international standards such as the International Standard Industrial Classification (ISIC) maintained by the United Nations.

The Indonesian economy presents a particularly complex classification challenge due to its diversity and scale. As one of the largest economies in Southeast Asia, Indonesia hosts millions of registered business entities spanning agriculture, manufacturing, services, and extractive industries. The geographic distribution of economic activity

across the archipelago creates regional variations in industry composition that must be accurately captured through classification systems.

1.1.2. The Digital Transformation of Business Registration

The rapid digitalization of business registration processes in Indonesia has created both opportunities and challenges for classification accuracy. While digital systems enable faster processing of business registrations, the sheer volume of applications has made manual classification increasingly impractical. According to data from the Ministry of Law and Human Rights, more than 200,000 new business entities are registered annually through the Online Single Submission (OSS) system, each requiring accurate KBLI code assignment.

The Online Single Submission system, launched as part of Indonesia's economic reform agenda, represents a fundamental shift in how businesses interact with government regulatory processes. By consolidating multiple licensing requirements into a single digital platform, OSS has substantially reduced the administrative burden on businesses. However, this efficiency gain has concentrated classification decisions at a single point, increasing the importance of accurate and consistent KBLI assignment.

The digital transformation has also changed the nature of information available for classification decisions. Business registration documents submitted through OSS include structured data fields as well as free-text descriptions of business activities. These textual descriptions vary substantially in length, specificity, and language quality, presenting challenges for both human classifiers and automated systems.

1.1.3. Challenges of Manual Classification

Traditional approaches to business classification have relied heavily on manual review by trained personnel who examine business descriptions, articles of association, and other documentary evidence. However, this approach suffers from several limitations that become increasingly problematic at scale.

First, human classifiers may exhibit inconsistency due to fatigue, varying levels of expertise, or subjective interpretation of ambiguous business descriptions. Research in

cognitive psychology has demonstrated that human decision-making is susceptible to various biases and heuristics that can affect classification consistency (**Kahneman, 2011**). When different classifiers approach the same business description, they may reach different conclusions based on their individual interpretation and experience.

Second, the time required for manual classification creates bottlenecks in the registration process. Each classification decision requires careful reading and interpretation of business descriptions, cross-referencing with the KBLI taxonomy, and consideration of edge cases and ambiguities. As registration volumes increase, the accumulated time for manual classification becomes a significant constraint on processing capacity.

Third, the cost of maintaining a sufficient workforce to handle classification volumes represents a significant operational expense. Qualified classifiers require training in both the KBLI taxonomy and Indonesian business practices. Staff turnover necessitates ongoing training investment, and scaling capacity to handle peak periods requires maintaining excess staffing during normal periods.

Fourth, quality assurance for manual classification is challenging to implement systematically. While sample audits can identify individual errors, detecting systematic biases or category-level problems requires substantial analytical effort. The absence of comprehensive quality monitoring may allow classification errors to accumulate undetected.

1.1.4. The Promise of Automated Classification

The emergence of natural language processing (NLP) and machine learning technologies has opened new possibilities for automated text classification. These approaches can process textual descriptions of business activities and assign appropriate classification codes with increasing accuracy. Recent advances in transformer-based language models, particularly BERT (Bidirectional Encoder Representations from Transformers) and its variants, have demonstrated remarkable performance improvements over traditional feature-engineering approaches (Devlin et al., 2019).

Automated classification offers several potential advantages over manual approaches. Consistency is inherent to algorithmic systems: the same input will always produce the same output, eliminating the variability introduced by human factors. Scalability is also enhanced, as computational resources can be expanded more easily than human workforces. Speed improvements enable near-instantaneous classification decisions, removing bottlenecks from registration workflows.

Furthermore, automated systems can be systematically monitored and improved. Classification decisions can be logged and analysed to identify patterns of error. Model updates can be deployed to address identified problems. Performance metrics can be tracked over time to ensure sustained quality.

However, realising these potential advantages requires careful system design that addresses the specific challenges of Indonesian business classification. The Indonesian language, the hierarchical KBLI taxonomy, and the practical requirements of production deployment all introduce considerations that must be systematically addressed.

1.1.5. The Indonesian Language Context

The Indonesian language presents unique challenges for text classification systems. As an agglutinative language with complex morphological patterns, Indonesian requires specialised processing approaches that may not be adequately addressed by multilingual models trained primarily on Western languages (Koto et al., 2020).

Indonesian word formation involves the combination of root words with various affixes (prefixes, suffixes, infixes, and circumfixes) to create derived words with modified meanings. For example, the root word "kerja" (work) can become "bekerja" (to work), "pekerja" (worker), "pekerjaan" (job/work), or "mengerjakan" (to do/work on). This morphological richness creates a large vocabulary space and complicates text processing.

The Indonesian language also exhibits code-mixing with English, particularly in business and technology domains. Business descriptions often contain English terms

for products, services, and industry concepts that may not have established Indonesian equivalents. This bilingual mixing requires language models capable of handling multiple languages within single texts.

The development of Indonesian-specific pre-trained language models represents a significant recent advance. IndoBERT, introduced by (Koto et al., 2020) as part of the IndoNLU benchmark, was pre-trained on approximately 4GB of Indonesian text from various sources including Wikipedia, news articles, and web crawls. This Indonesian-optimised model demonstrates improved performance on Indonesian NLP tasks compared to multilingual alternatives.

1.2. Research Problem

Despite the promising capabilities of modern NLP systems, several challenges remain in applying these technologies to the Indonesian business classification context. These challenges motivate the current research and define the scope of investigation.

1.2.1. The Hierarchical Classification Challenge

The hierarchical nature of the KBLI system introduces complexity not present in flat classification tasks. A five-digit KBLI code exists within a nested structure where:

- The first digit indicates the section (21 sections, A-U)
- The first two digits indicate the division (88 divisions)
- The first three digits indicate the group (241 groups)
- The first four digits indicate the class (504 classes)
- All five digits indicate the subclass (1,497 subclasses)

Classification errors at different levels of this hierarchy have varying degrees of severity. Confusing two subclasses within the same class (a one-digit error) is less problematic than confusing subclasses from different sections (a five-digit error). Yet most classification systems treat all misclassifications equally, failing to distinguish between minor and major errors.

The hierarchical structure also creates relationships between categories that could potentially be exploited for classification. Parent categories contain semantic information that constrains their children. Errors that respect hierarchical relationships (staying within the correct division while selecting the wrong subclass) are more semantically appropriate than errors that cross major category boundaries.

1.2.2. The Calibration Challenge

The practical deployment of classification systems in business contexts requires not only accuracy but also calibrated confidence estimates. Decision-makers need to know when a system's predictions are reliable and when human review is warranted. This requirement introduces the concept of selective prediction, where the system abstains from making predictions when uncertainty is high.

The problem of neural network overconfidence, wherein models express high confidence even for incorrect predictions, has been well documented in the literature (Guo et al., 2017). Modern neural networks trained with cross-entropy loss tend to produce probability estimates that are systematically too confident. This miscalibration undermines the utility of confidence scores for decision-making.

Methods for probability calibration, such as temperature scaling and Platt scaling, have been proposed to address this problem. Temperature scaling introduces a single parameter that adjusts the sharpness of probability distributions. Platt scaling fits a logistic regression model to transform classifier outputs into calibrated probabilities. However, the effectiveness of these methods in the Indonesian business classification context remains unexplored.

1.2.3. The Agentic Systems Challenge

The integration of large language models (LLMs) into agentic workflows, where AI systems can iteratively refine their outputs through self-correction, represents an emerging paradigm with potential applications in classification tasks. Agentic systems can recognise uncertainty in their initial predictions and take actions to gather additional information or refine their analysis.

For classification tasks, an agentic approach might involve:

- Making an initial classification prediction with associated confidence
- Recognising when confidence is below an acceptable threshold
- Gathering additional information through targeted queries
- Combining multiple information sources to improve the final prediction

However, the design principles and effectiveness of such agentic classification systems have not been systematically studied for industry classification purposes. Questions remain about when agentic approaches provide benefits over single-pass classification, how to design effective re-query strategies, and how to balance accuracy gains against increased computational costs.

1.3. Research Questions

Based on the identified challenges and gaps in the existing literature, this research addresses two primary research questions:

RQ1: How effectively can a transformer-based classifier, specifically leveraging Indonesian language model embeddings, assign five-digit KBLI 2020 codes to Indonesian company descriptions compared to traditional machine learning baselines?

This question addresses the fundamental classification performance comparison between neural and traditional approaches. The effectiveness is operationalised through multiple metrics including:

- Balanced Accuracy (BA): The arithmetic means of recall values across all classes, providing equal weight to each class regardless of frequency
- Macro-F1 Score: The arithmetic means of F1 scores computed separately for each class
- Top-k Accuracy: The proportion of samples for which the correct label appears within the top k predictions
- Hierarchy-aware Error: A custom metric that measures the average distance between predicted and true codes in the KBLI hierarchy

The comparison with traditional baselines (TF-IDF with SVM) provides context for interpreting neural model performance and informing technology selection decisions.

RQ2: To what extent can probability calibration and agentic re-query mechanisms improve the reliability and practical utility of automated KBLI classification for production deployment scenarios?

This question examines enhancement strategies beyond basic classification accuracy. The investigation encompasses:

- **Calibration Assessment:** Measurement through Expected Calibration Error (ECE) and Brier score
- **Coverage Analysis:** Quantification of predictions meeting confidence thresholds for automated acceptance
- **Agentic Mechanism Evaluation:** Assessment of re-query strategies for uncertain predictions
- **Ablation Studies:** Isolation of individual component contributions

1.4. Research Objectives

The overarching objective of this research is to develop, implement, and rigorously evaluate an automated KBLI classification system that achieves both high accuracy and reliable confidence estimation suitable for production deployment.

1. To design and implement a semantic classification pipeline utilising transformer-based sentence embedding for matching company descriptions against the complete KBLI 2020 taxonomy of 1,497 five-digit codes.
2. To establish robust baseline performance metrics using the traditional TF-IDF feature extraction approach combined with linear SVM classification, enabling quantitative comparison that isolates the contribution of neural semantic representations.
3. To implement and evaluate temperature-based probability calibration methods to improve the reliability of confidence estimates, enabling more effective confidence-based routing of predictions.

4. To design, implement, and evaluate an agentic re-query mechanism that leverages calibrated confidence scores to identify uncertain predictions and gather additional contextual information.
5. To comprehensively evaluate the complete classification system against pre-registered performance targets, documenting both successes and shortcomings to inform future development directions.

1.5. Significance of the Study

By combining an interpretable baseline with a transformer model and by enforcing calibrated decision rules, dependable KBLI codes can be delivered from public data at scale. The contribution is expected to raise the precision and trust of market-intelligence datasets at PT DLN, reduce manual curation cost, and establish reusable indicators—Coverage, calibration error, and hierarchy-aware accuracy—for routine monitoring and continuous improvement (Silla and Freitas, 2011). This research contributes to both academic knowledge and practical application across multiple dimensions.

- **Academic Contributions**

From a theoretical perspective, this study provides empirical evidence on the effectiveness of transformer-based models for hierarchical industry classification in a low-resource language context. While substantial research has examined BERT-family models for English text classification, applications to Indonesian industrial taxonomy classification remain limited.

The investigation of probability calibration methods and their interaction with agentic re-query mechanisms extends the literature on reliability enhancement for neural classification systems. The hierarchy-aware evaluation framework developed in this research offers a more semantically meaningful assessment approach than standard flat-accuracy metrics.

- **Practical Contributions**

From a practical perspective, the developed system addresses real operational needs in business intelligence contexts. The classification pipeline provides immediate utility for company profiling, regulatory compliance verification, and market segmentation tasks.

The calibration and routing framework enable deployment strategies that balance automation efficiency with quality assurance. The empirical findings inform expectations regarding achievable performance levels and guide resource allocation decisions between automated processing and human review.

1.6. Scope, Assumptions, and Delimitations

This research is bounded by several deliberate scoping decisions that define its focus and limitations.

- **Classification Scope**

The classification task is limited to the assignment of five-digit KBLI 2020 codes, representing the most granular level of the taxonomy. Higher-level aggregations (two-digit divisions, one-letter sections) are derivable from five-digit predictions through prefix extraction and are not modelled as separate classification targets.

The research does not address multi-label classification scenarios where a company might legitimately belong to multiple KBLI categories. Each company is assumed to have a single primary classification.

- **Data Scope**

Only publicly accessible content (corporate websites, B2B directories) is utilized; authenticated sources and sensitive personal data are excluded. The research utilises Indonesian company descriptions obtained from the organisation's existing database. The dataset reflects the operational focus on manufacturing, trading, and technology sectors, with limited representation of

agriculture, mining, and certain service categories. While efforts have been made to ensure data quality and remove obvious errors, the findings may not generalise to all Indonesian business contexts, particularly those with minimal digital presence or non-standard business descriptions.

The domain is focused on PT DLN's product space (electric motors, gearboxes/reducers, VFDs, and related automation). Indonesian and English texts are included; code-mixing is expected. Organizational security governance frameworks are not analyzed in depth; nevertheless, robots' directives and polite crawling are respected as good practice. Evaluation is performed on a labelled corpus curated from public evidence and official references; external generalizability is discussed later.

- **Technical Scope**

The research focuses on offline batch classification rather than real-time streaming scenarios. System deployment considerations such as API architecture, containerisation, horizontal scaling, and continuous integration are considered outside the research scope.

The research does not address the fine-tuning of base language models on domain-specific corpora, instead utilising off-the-shelf pre-trained models. This decision enables faster prototyping while representing a potential avenue for future improvement.

1.7. Definitions of Key Terms

- **KBLI 2020:** the official Indonesian classification of economic activities at five digits, aligned to **ISIC Rev.4** at four digits (Badan Pusat Statistik, 2020; United Nations, 2008).
- **Agentic AI:** an orchestration pattern in which additional retrieval/analysis is triggered automatically when model confidence is low.
- **Coverage@ τ :** the proportion of predictions automatically accepted because calibrated confidence $\geq \tau$.

- **Hierarchy-aware error:** an evaluation that recognises near-misses within a taxonomy (Silla and Freitas, 2011).
- **Calibration (ECE/Brier):** the closeness of predicted probabilities to empirical frequencies (Guo et al., 2017).

1.8. Problem Statement

The principal drivers of KBLI misclassification from public web data are summarised in Figure 1.1, and the focal failure mode investigated in this thesis is highlighted in red (Adriani et al., 2007).

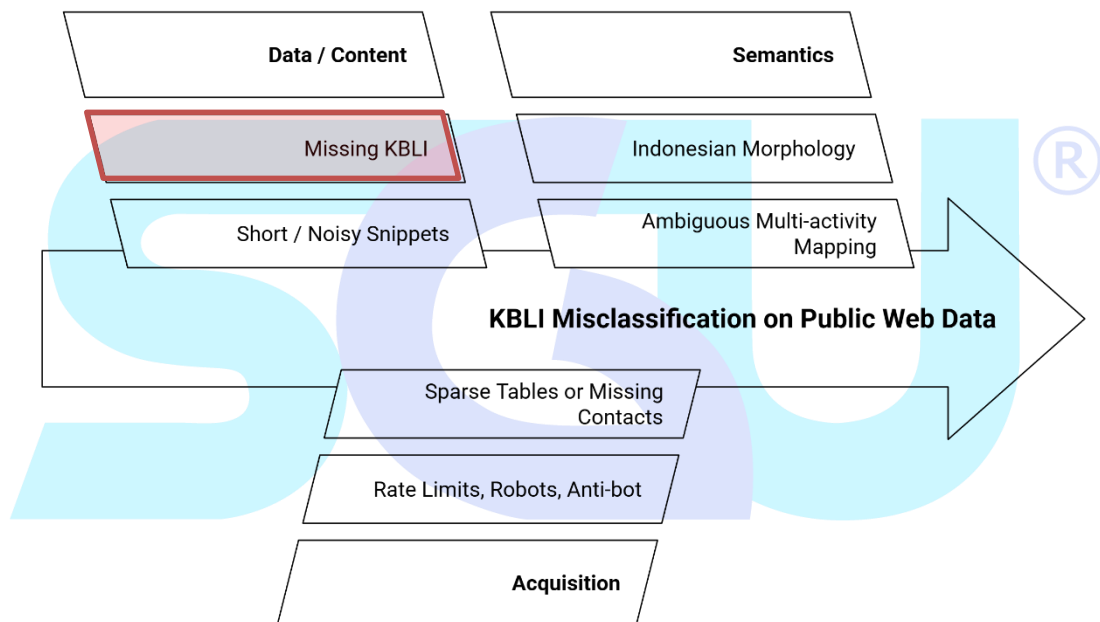


Figure 1. 1 Fishbone of KBLI misclassification drivers in public web data, with the focal node “Missing KBLI” highlighted in red.

This fishbone diagram Figure 1.1 Above organises root causes along three branches: **Data/Content** (short or noisy snippets, code-mixing, duplicate pages, and the highlighted **Missing KBLI** label), **Acquisition** (rate-limits, robots directives, anti-automation defences, and sparse tabular fields), and **Semantics** (Indonesian morphological variation, multi-activity ambiguity, and near-misses within the KBLI taxonomy at the five-digit level). The visual emphasis on **Missing KBLI** indicates the primary failure mode addressed by the proposed agentic pipeline; the remaining

branches inform supporting interventions (normalisation, calibrated classification, thresholded auto-acceptance, and human-in-the-loop review). The diagram should be read left-to-right: upstream causes accumulate on the spines and converge on the right-hand effect node, “KBLI Misclassification on Public Web Data” (Adriani et al., 2007).

1.9. Thesis Structure

The thesis is organised according to SGU’s Thesis Work Report template and Thesis Guidelines:

- **Chapter 1** presents the background, problem, aim, objectives, research questions with hypotheses, scope, significance, definitions, and the fishbone figure;
- **Chapter 2** presents a comprehensive literature review covering text classification methodologies, transformer-based language models, probability calibration techniques, and agentic AI system design principles. The chapter concludes with identification of research gaps.
- **Chapter 3** details the research methodology, presenting the design science framework, data collection procedures, model architectures, evaluation metrics, and experimental protocols.
- **Chapter 4** reports experimental results including performance comparisons, calibration analyses, ablation studies, and statistical significance tests. Discussion of findings is integrated throughout.
- **Chapter 5** concludes the thesis by summarizing key findings, articulating contributions, acknowledging limitations, and proposing directions for future research