VPBank Technology Hackathon 2025

General Brief

Please fill up this table and use this document as a template to write your proposal.

Challenge Statement	Auto Business Data Dictionary
Team Name	

Team Members

Full Name	Role	Email Address	School Name (if applicable)	Faculty / Area of Study	LinkedIn Profile URL

Content Outline

Solutions Introduction

Definition of Business Terminology

To ensure a shared understanding, we first define business terminology as follows:

Business terminology refers to the set of standardized words, phrases and expressions that carry specific and agreed upon meanings within an organization or business domain.

These terms describe business concepts, entities, metrics, processes and rules and are used to ensure consistent understanding and communication across teams, systems and documentation.

Key characteristics

- Represents domain specific vocabulary (e.g., Customer, Order, Revenue, Loan Application)
- Has a precise definition understood and accepted by business and technical stakeholders.
- May vary by by context or department (e.g., “Customer” in Retail vs Banking)
- Serves as the foundation for data dictionaries, glossaries and data governance efforts

Examples

Term	Definition	Context
COGS (Cost of Goods Sold)	The direct costs attributable to the production of goods sold by a company	Finance
Customer	An individual or organization that purchases or uses a company’s products or services	Sales
Premium Customer	A customer who is spending over $1000 a month	Sales
Active Account	An account which has at least one transaction in the past 90 days	Sales
Customer Onboarding	The process of welcoming and setting up a new customer in the system after account creation	Operations

In other words, business terminology represents the common language of an organization; defining what things mean, how they are used and where they apply, so that everyone interprets business information consistently.

End-to-End Process Overview

Once the terminology foundation is established, the process involves extracting business terms from documents and structuring them into a governed business glossary through the following steps:

1. Read business documents (e.g., BRDs, SRSs, policies)
2. Identify and extract business terms using AI/NPL
3. Model and enrich extracted terms to build a structured data dictionary.
4. Organize and publish these terms as a business glossary for enterprise-wide search and sharing.

Essentially, this solution transforms unstructured business language into a structured data dictionary and a governed business glossary.

Expected Benefits

This foundation enables:

- Consistent terminology across systems and departments
- Improved data lineage and governance.
- Easier onboarding and cross-team alignment through shared understanding.

Multilingual Support Consideration

An important aspect not explicitly mentioned in the challenge is multilingual capability. Since VPBank is a local Vietnamese bank, documents may exist in both English and Vietnamese. Therefore, the system should support Vietnamese language detection in addition to English.

However, since many AI language models are primarily optimized for English, the recommended approach is:

1. Translate non-English documents (Vietnamese → English)
2. Process and extract terms in English (as the base language)
3. Translate the final glossary entries back to Vietnamese for local usage

English will serve as the base reference language, while other supported languages will be translated versions of the standardized glossary.

Summary

The solution converts unstructured business documentation into a structured, multilingual business glossary; providing a single, searchable source of truth for business concepts, and laying the foundation for consistent understanding, data governance, and cross-department collaboration.

Impact of Solution

Implementing this solution will have significant and measurable impact across multiple dimensions of the organization; from business understanding to data governance and collaboration.

Improved Business Clarity and Alignment

- Creates a single source of truth for all business terminology across departments.
- Ensures consistent interpretation of key concepts such as “Customer”, “Revenue” or “Active Account”
- Reduces confusion and miscommunication between business, data and IT teams
- Strengthens alignment between business processes, system requirements and data models

Accelerated Knowledge Discovery

- Enables fast, centralized search for definitions and context of business terms
- Makes previously fragmented information (hidden in BRDs, SRSs and policy documents) easily accessible.
- Supports new employees or teams with a clear reference for domain knowledge and terminology.

Enhanced Data Governance and Quality

- Provides the foundation for enterprise data governance, linking business terms to data assets and metadata.
- Improves data lineage visibility; users can trace where terms appear in systems, reports and processes.
- Promotes standardization across reports, KPIs and system documentation, reducing inconsistencies.

Strengthened Decision Making and Compliance

- Ensure that reports and analytics are based on consistent definitions, improving decision accuracy.
- Supports audit and regulatory compliance by providing traceable, governed business definitions.
- Enhances transparency in how metrics and data are defined across business units.

Multilingual Inclusivity

- Support both English and Vietnamese, enabling teams to access and contribute to the glossary in their preferred language.
- Promotes a more inclusive environment for cross functional collaboration within a bilingual organization like VPBank.

Long-Term Strategic Benefits

- Lays the groundwork for future AI and data initiatives (e.g., semantic search, knowledge graphs, or intelligent documentation).
- Becomes a core asset for enterprise knowledge management and data democratization efforts.
- Positions the organization as a data-driven, knowledge centric enterprise with improved agility and communication.

The solution transforms scattered, static business knowledge into a structured, accessible and multilingual knowledge base.

It improves understanding, reduces ambiguity, enhances governance and empowers every team; from business users to data engineers to speak the same language when interpreting business information.

Deep Dive into Solution

The most critical and technically challenging part of this solution is the automatic identification and extraction of business terms from unstructured documents. Business terms often carry different meanings depending on their context.

For example, the word “Customer” can refer to:

- an end-customer purchasing a product
- a corporate customer holding a business account
- a customer record stored in CRM, or
- a customer support process within operations

This variation highlights the need for a context-aware extraction approach to ensure that each term is meaningful and correctly interpreted within its business domain.

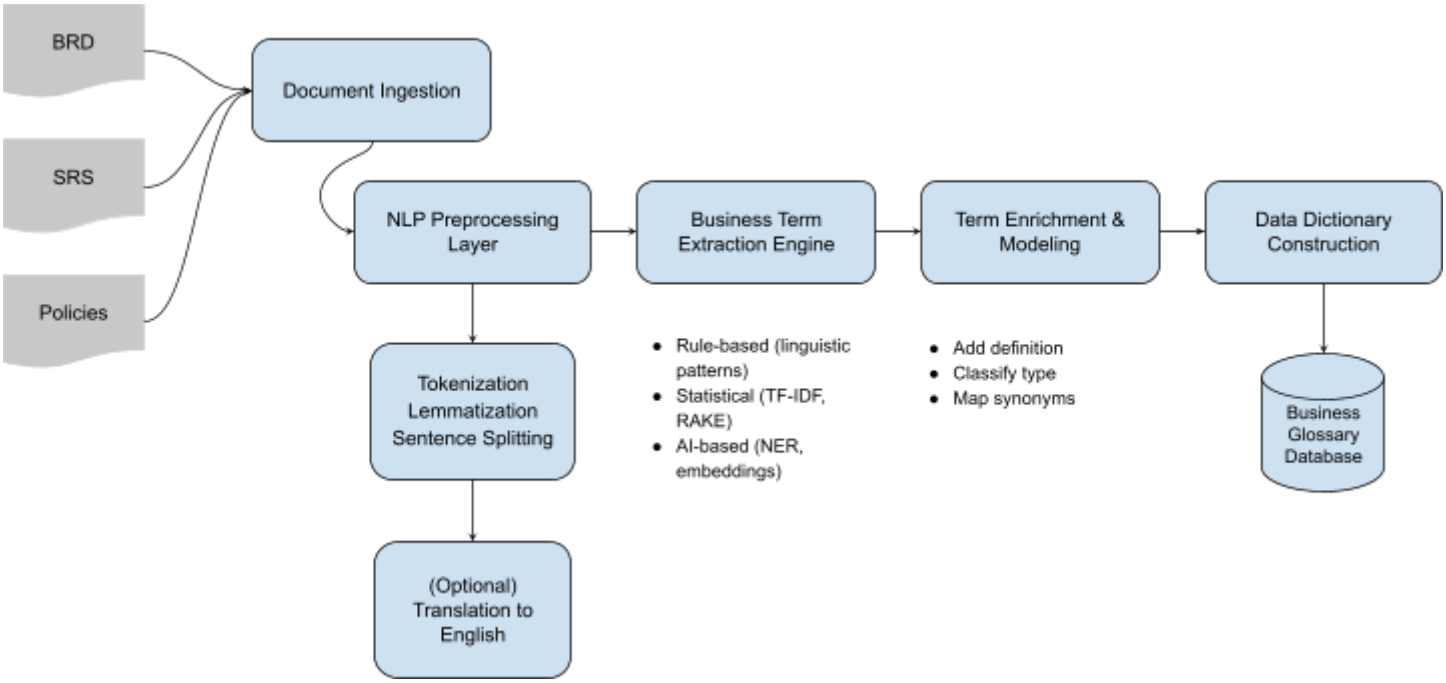
Strategies for Business Term Extraction

Several strategies can be combined to accurately identify business terms from documents:

1. Pattern-based extraction: identify specific linguistic structures (e.g., “<term> means/defines/refers to…”)
2. Frequency-based extraction: leverage statistical metrics like TF-IDF or N-gram frequency to identify candidate terms.
3. Contextual/AI-based extraction: use embedding models and transformers to capture semantic meaning and relationships.

An effective solution typically requires a hybrid approach that combines linguistic, statistical and semantic insights.

System Components and Process Flow



1. **Document Ingestion**
    - Input sources: BRDs, SRSs, policies, etc.
    - Read from multiple formats (PDF, DOCX, TXT)
    - Detect document language and perform translation (e.g., Vietnamese -> English) for consistent processing
2. **NLP Preprocessing Layer**

    Prepare raw text for structured analysis through:
    - Tokenization and lemmatization
    - Sentence segmentation
    - Part-of-Speech (POS) tagging and phrase detection

3. Business Term Extraction Engine

Implements multi-strategy extraction:

- **Linguistic/Rule-based**: identify noun phrases and definition patterns
- **Statistical**: apply TF-IDF (Term Frequency - Invert Document Frequency), RAKE (Rapid Automatic Keyword Extraction) or N-gram frequency scoring to detect high-value terms.
- **Semantic/AI based:**
    - Use language models or embeddings to capture contextual meaning.

- Apply pre-trained transformer models to tag and classify business entities
- Perform sentence similarity analysis to detect definition-like patterns
- Use zero-shot classification to categorize terms (e.g., Entity, Metric, Process, Rule) without labeled training data.

4. Hybrid Filtering & Validation

Combines multiple signals to:

- Remove noise or duplicates
- Merge similar terms
- Ensure consistency and quality in extracted term set

5. Term Enrichment & Modeling

Enhance each term with:

- Definition derived from context
- Classification (Entity, Metric, Process, Rule)
- Synonym and relationship mapping
- Domain and ownership metadata

6. Data Dictionary Construction

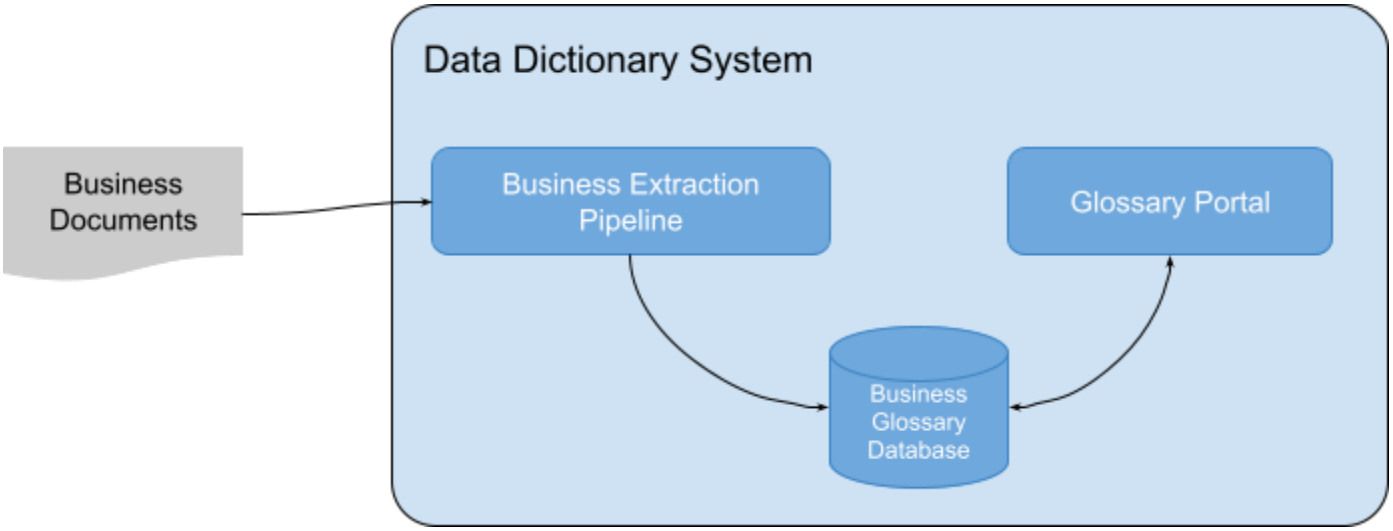
Creates a structured, standardized record for each term, including:

- Term name
- Definition
- Context
- Relationships
- Owner
- Synonyms
- Lineage information

These are stored in a centralized Business Glossary Database, providing a unified and searchable repository.

Overall System Architecture

While the previous diagram illustrated the detailed extraction flow, the following diagram represents the end-to-end view of the Data Dictionary System, showing how extracted terms are stored and accessed by end users.



This high-level view demonstrates how the Business Extraction Pipeline processes input documents and stores the results in a centralized Business Glossary Database which can be accessed via a Glossary Portal for search, review and collaboration.

Extended Considerations (Out of Scope)

While the main focus is on term extraction and modeling, the broader system could include:

- Document ingestion from multiple sources (S3, shared drives, local repositories)
- Process comparison and compliance checks against standard business process libraries
- Business owner notification workflow for reviewing or approving new or changed terms

These capabilities form part of an enterprise-wide metadata management ecosystem but are beyond the current solution's scope.

Summary

This solution establishes an automated, intelligent pipeline for discovering, enriching and managing business terms from documents. It bridges the gap between unstructured knowledge (text documents) and structured metadata (data dictionaries and glossaries), enabling:

- Consistent business language across systems
- Improved data governance and traceability
- Enhanced collaboration between business and technical stakeholders.

Conclusion

The proposed solution provides a systematic and scalable approach to extract, structure and govern business terminology across the organization. By leveraging Natural Language Processing and AI-based context understanding, the system transforms fragmented and unstructured information from business documents into a centralized, searchable business glossary.

This approach not only improves the consistency and clarity of business language but also lays the foundation for better data governance, data lineage tracking and cross-departmental collaboration. Teams will have a unified reference point to understand business definitions, reducing ambiguity in requirements, reporting and process design.

Moreover, the inclusion of multilingual support ensures adaptability to VPBank's bilingual environment, enabling accurate term extraction and interpretation across English and Vietnamese sources. As a result, the solution contributes directly to higher operational efficiency, easier onboarding and improved business-IT alignment.

In summary, this solution establishes a foundation for a data-driven knowledge ecosystem, where business terms, definitions and processes are consistently understood, easily accessible, and continuously improved.