

Students

Dilan Sataloğlu, Doğa İzci, Arda Keleş
Computer Science and Engineering,
Sabancı University

Faculty Member

Assist. Prof. Polat Göktaş
FENS, Sabancı University

Company Advisor

Volkan Yurtseven
Akbank



Abstract

InfoGuide is a modular enterprise document-processing pipeline designed to transform raw corporate documents into structured outputs. The system integrates ingestion, preprocessing, anonymization, classification, taxonomy assignment, and evaluation within a single configurable workflow.

Problem & Motivation

Problem is that enterprise documents often:

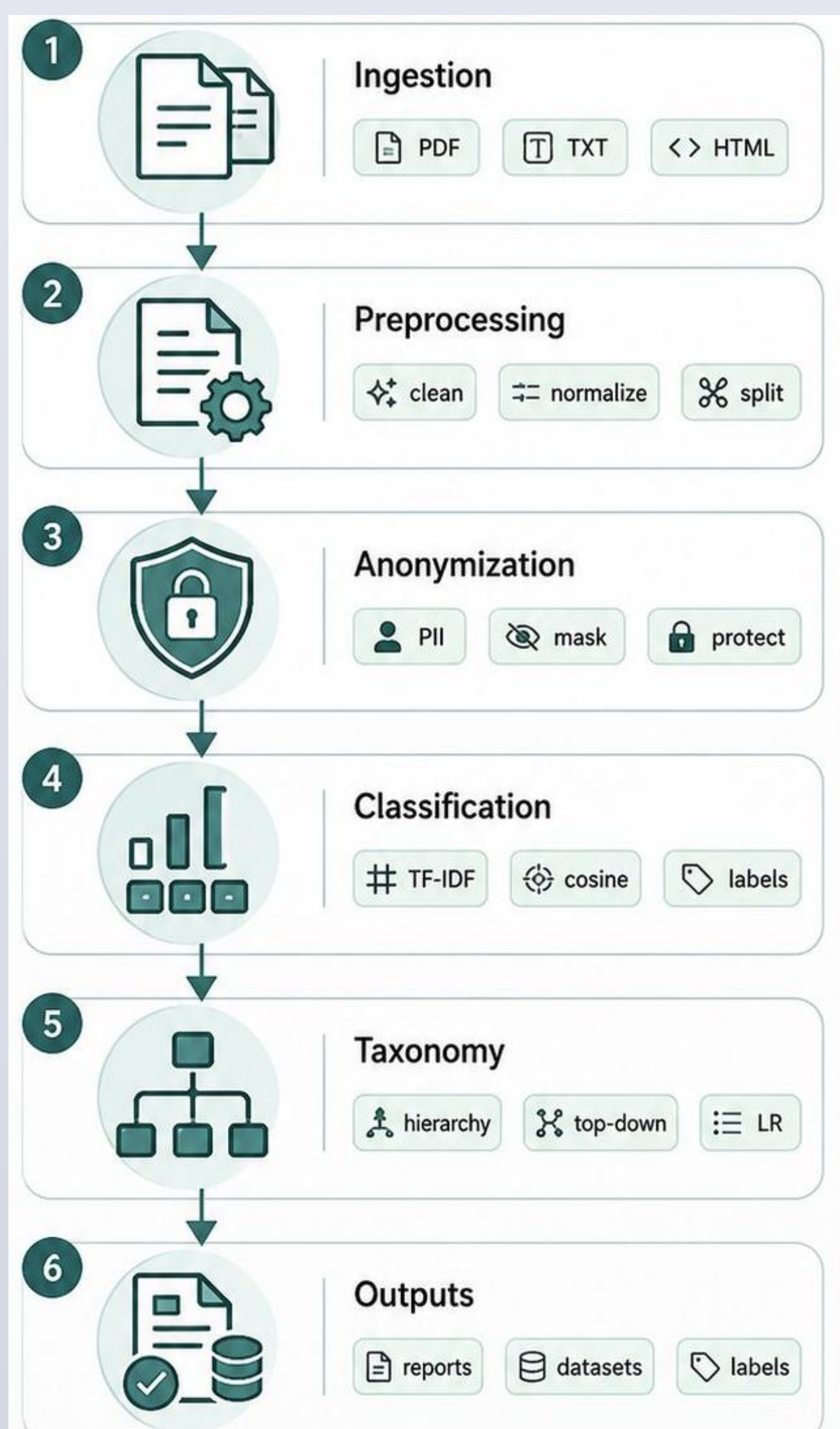
- Exist in multiple file formats,
- Include sensitive information,
- Lack consistent organization and labels,
- Difficult to search, classify, and group.

The pipeline produces cleaned document text, anonymized versions, document-type labels, and taxonomy assignments.

Research Questions

- **RQ1:** How can a modular document-processing pipeline transform raw enterprise documents into clean text, anonymized outputs, document-type labels, and taxonomy-organized datasets?
- **RQ2:** How can sensitive information be detected and anonymized while maintaining document utility?
- **RQ3:** How can enterprise documents be automatically classified and organized into meaningful hierarchical categories?

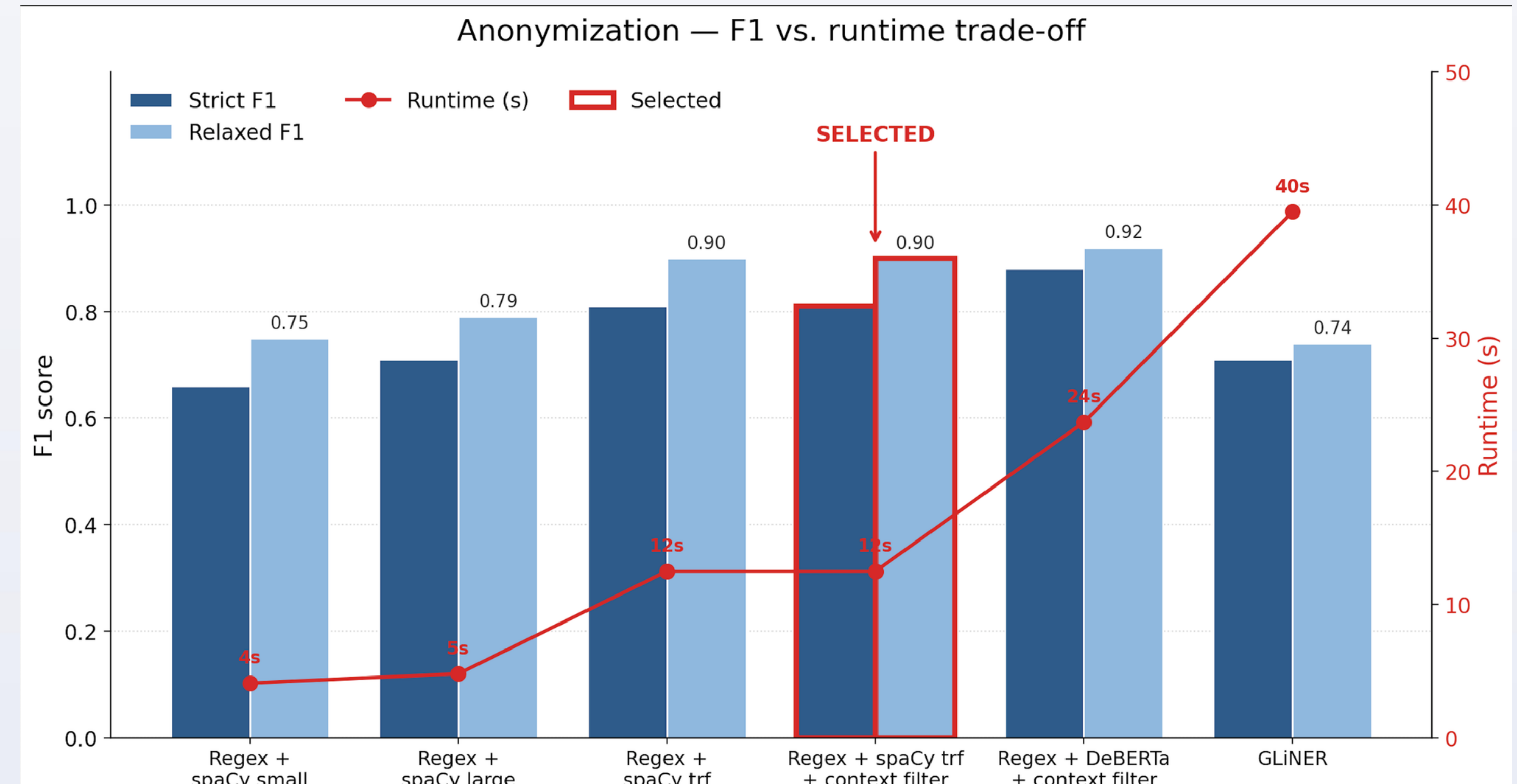
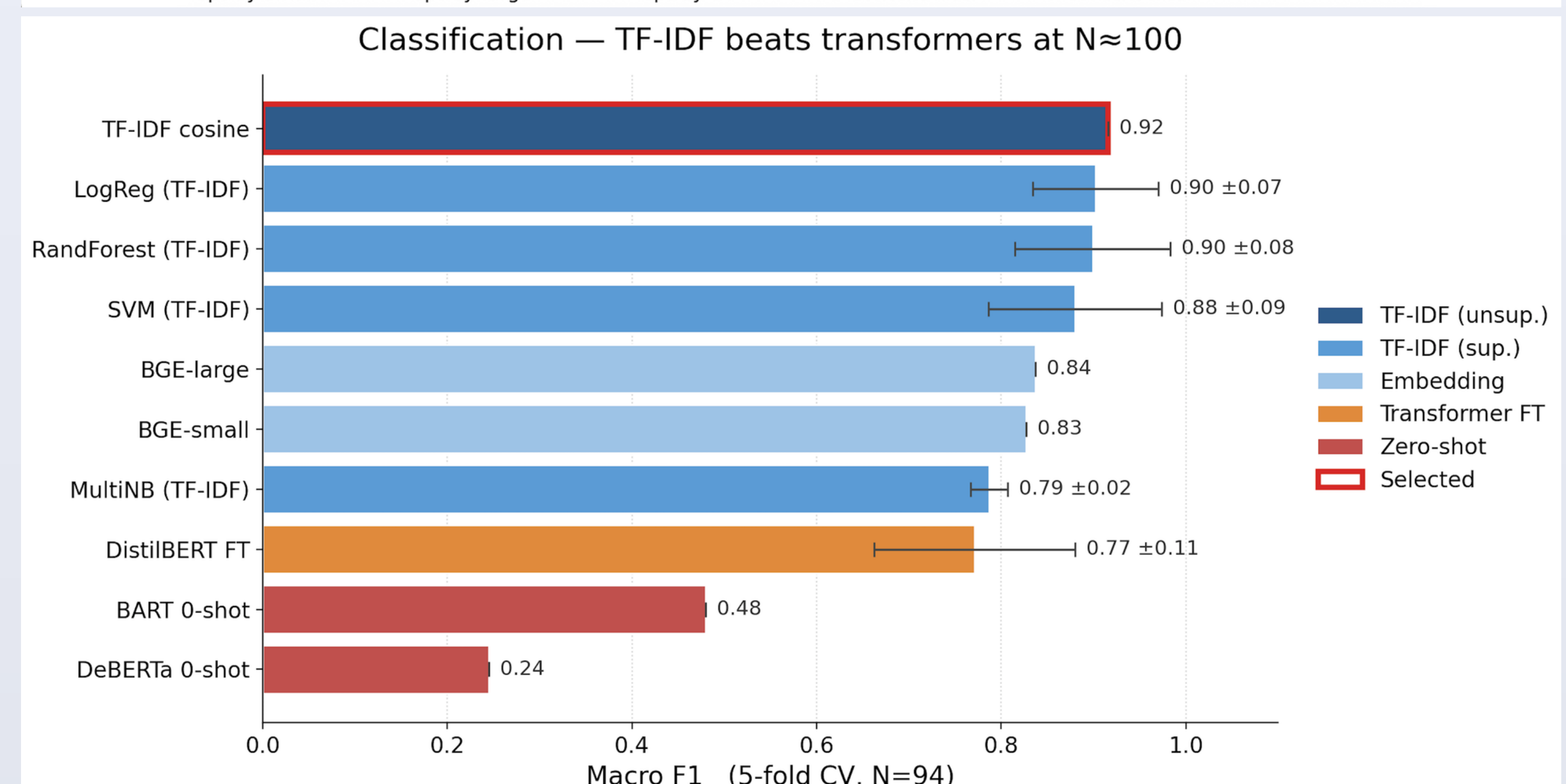
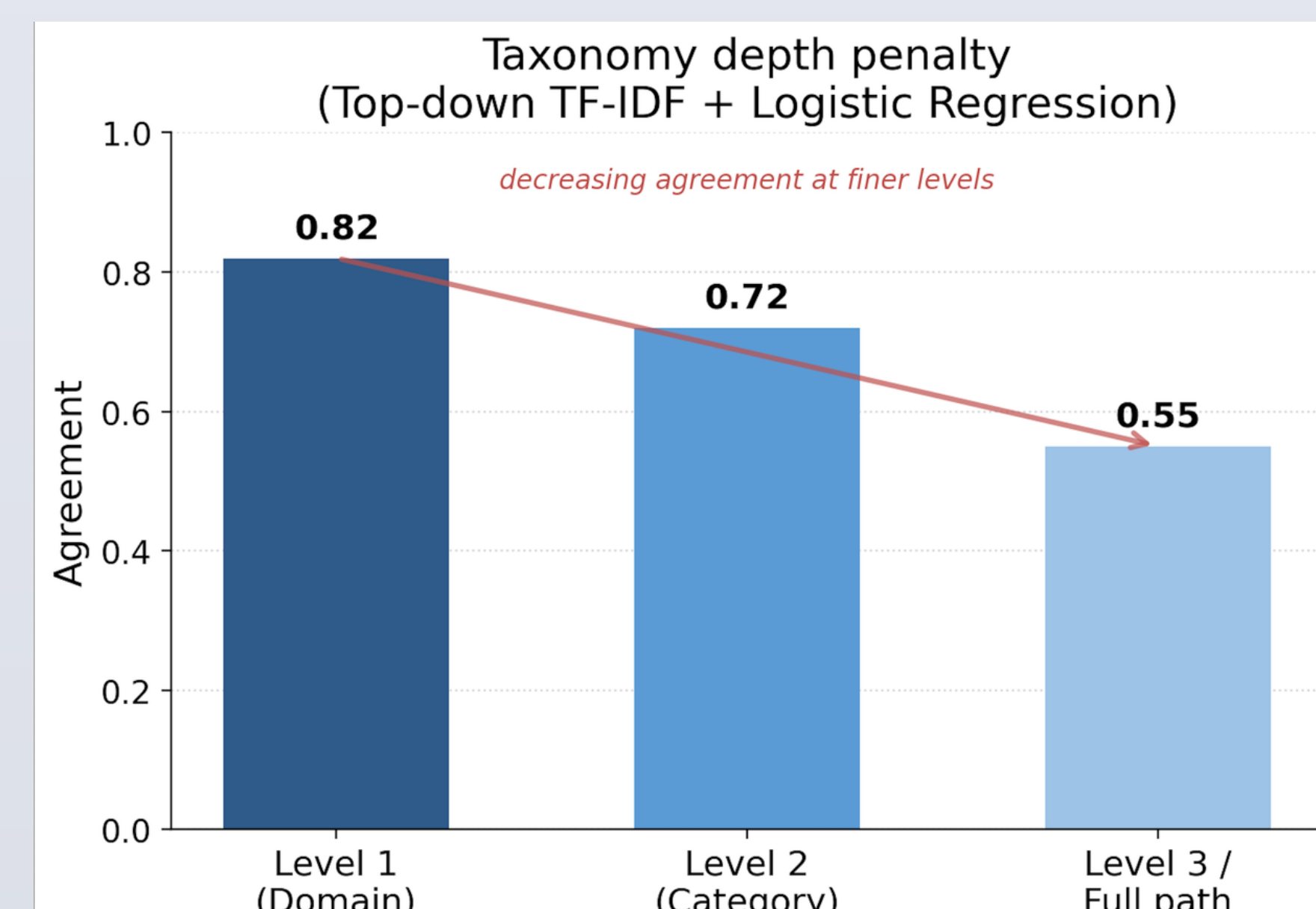
Architecture & Methodology



Experimental Setup

- InfoGuide was evaluated as a six-stage pipeline given above. Each stage was treated as a separate experiment: multiple methods were compared per stage and the strongest configuration was selected.
- The dataset combined 133 real enterprise documents (PDF / TXT / HTML) and 66 synthetic ones .
- **Metrics:** precision / recall / F1 (strict vs. relaxed) for anonymization, accuracy and macro-F1 for classification with a confidence threshold for Needs_Review.

Results & Discussion



- **Anonymization:** DeBERTa achieved the highest F1 score (0.92), while Regex + spaCy-trf + Context Filter was selected as the best accuracy-runtime trade-off.
- **Classification:** TF-IDF Cosine Similarity achieved the strongest overall performance (Macro F1 = 0.92), outperforming transformer-based approaches on the current dataset.
- **Taxonomy:** Taxonomy assignment was most accurate at higher levels of the hierarchy (82% Level 1 agreement), while fine-grained Level 3 prediction remained more challenging (55% full-path agreement).

Contributions & Impact

Our work enabled InfoGuide Pilot-2 to:

- Transform raw, mixed-format documents into clean, structured text
- Produce anonymized, privacy-safe document versions
- Assign document-type labels and organize documents under a banking taxonomy
- Provide a configurable framework for repeatable experimentation

Future Work

- **Evaluate on real enterprise documents** to validate performance beyond the synthetic benchmark.
- **Improve scalability and runtime efficiency** for large-scale document collections.
- **Support layout-aware document understanding** for forms, tables, and scanned documents.