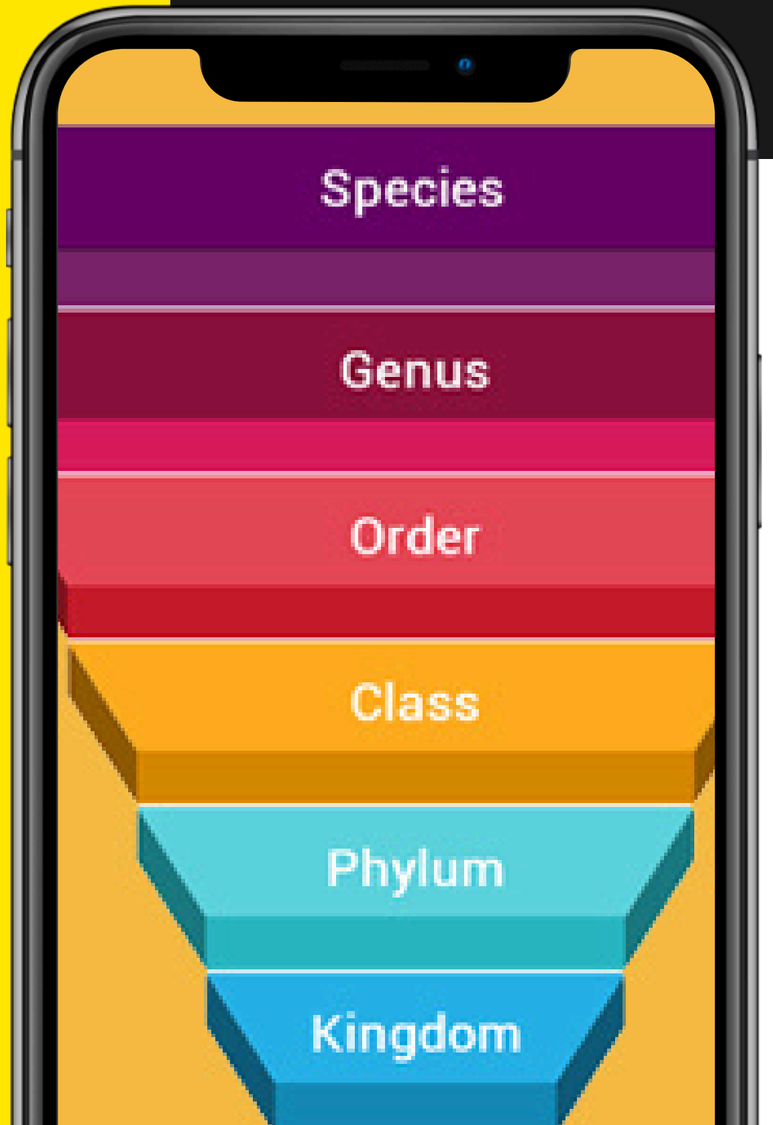
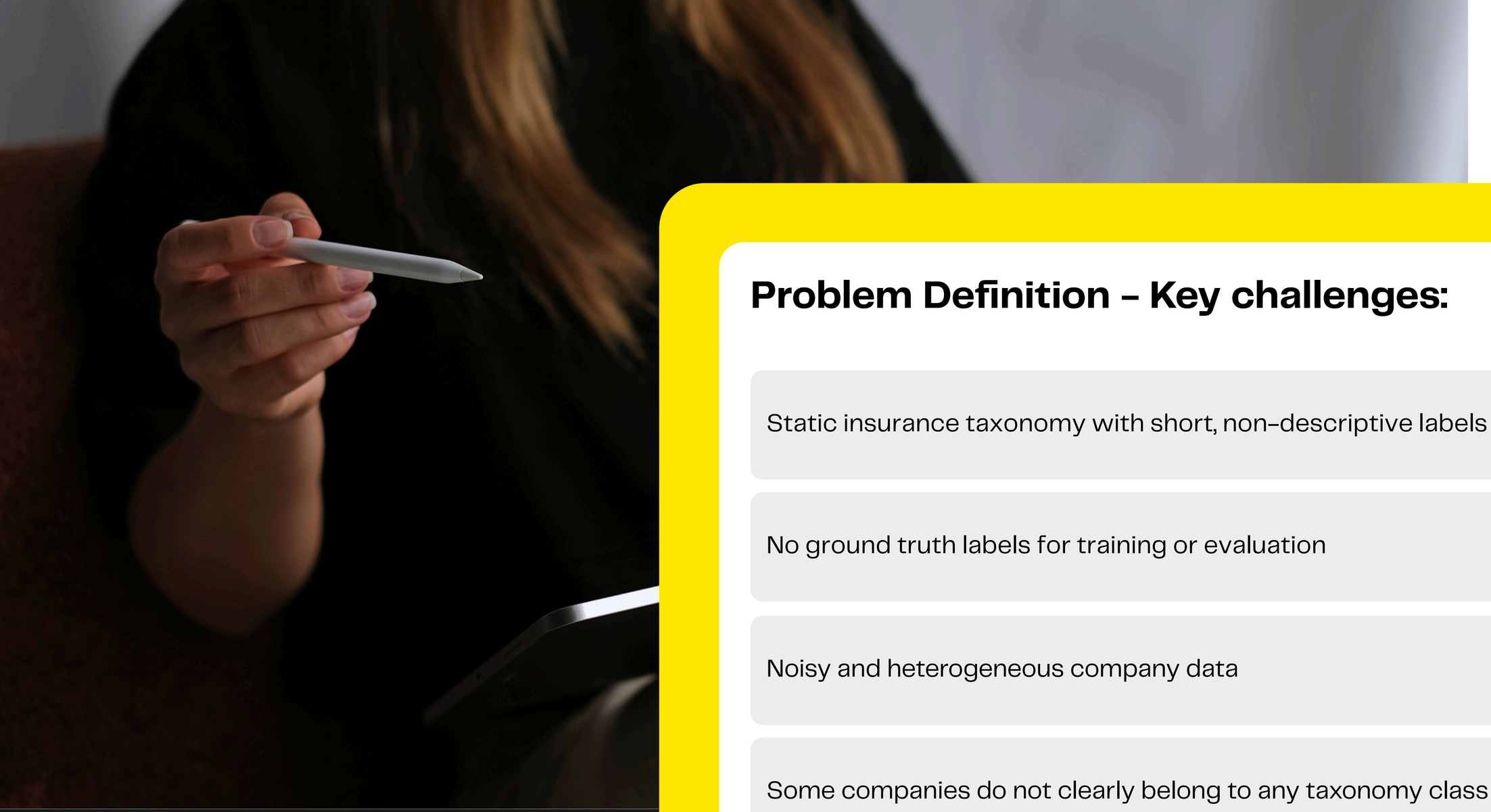
Company Classifier for a Static Insurance Taxonomy

Candidate: Ignat Ramona





Problem Definition – Key challenges:

- Static insurance taxonomy with short, non-descriptive labels
- No ground truth labels for training or evaluation
- Noisy and heterogeneous company data
- Some companies do not clearly belong to any taxonomy class

Objectives

Use all available company data effectively →

Assign one or more relevant insurance labels →

Be explainable and auditable →

Scale beyond the provided dataset →

Handle taxonomy coverage gaps gracefully

Understanding the Input Signal

High-signal fields

Category and niche: strongest indicators
Business tags: keyword-rich
Description: contextual, but noisy

Low-signal / risky patterns

Generic terms: services, manufacturing, solutions
Marketing-heavy descriptions

Technical Flow

The solution follows a simple and robust pipeline using 4 steps:

Risch Company Representation

We concatenate all informative fields (Company Description + Business Tags + Sector + Category + Niche) to build a complete and expressive company profile.

Hybrid Vectorization

We represent companies using:
Word N-gram – capture business context and multi-word expressions
Character N-grams – provide robustness to typos, abbreviations, and morphological variations

Weak Supervision (Key Innovation)

- We leverage structured fields to identify high-confidence ‘anchor companies` for each taxonomy label.
- We compute a centroid vector from these anchors. As a result, a label like “Plumbing Services” is no longer just the word “plumbing”, but a semantic vector enriched with terms such as pipes, water, leaks, installation, learned from real companies.

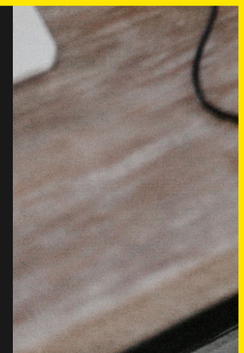
Classification

Each company is assigned the label whose centroid has the highest cosine similarity to the company vector

Why TF-IDF + Character N-grams

Why this works

- Fast, stable, and explainable
- Character n-grams handle: typos and abbreviations, brand names, multilingual noise
- Efficient for large-scale sparse data



Why not embeddings / zero-shot models

- Hard to validate without ground truth
- Risk of plausible but incorrect matches
- Costly and less controllable at scale

Weak Supervision for Label Representation

Problem

Taxonomy labels are short and semantically poor

Solution

- Automatically identify high-precision “seed companies” per label
- Use structured fields (category, niche, tags) with generic-term filtering
- Build label centroids from real company examples
- Fall back to label text only if seeds are insufficient

Impact

- Labels gain domain-specific meaning
- Reduces superficial keyword matching

Multi-Label Decision Strategy

Assignment logic

- Always assign the best-matching label
- Add additional labels only if: similarity exceeds a minimum threshold, score is close to the best label score

Confidence interpretation

- Similarity score, not a probability
- Used for audit prioritization and error analysis

Validation Without Ground Truth

Evaluation methodology

01 Semantic validation on stratified samples

02 Consistency checks with sector/category/niche

03 Stability analysis using alternative candidates

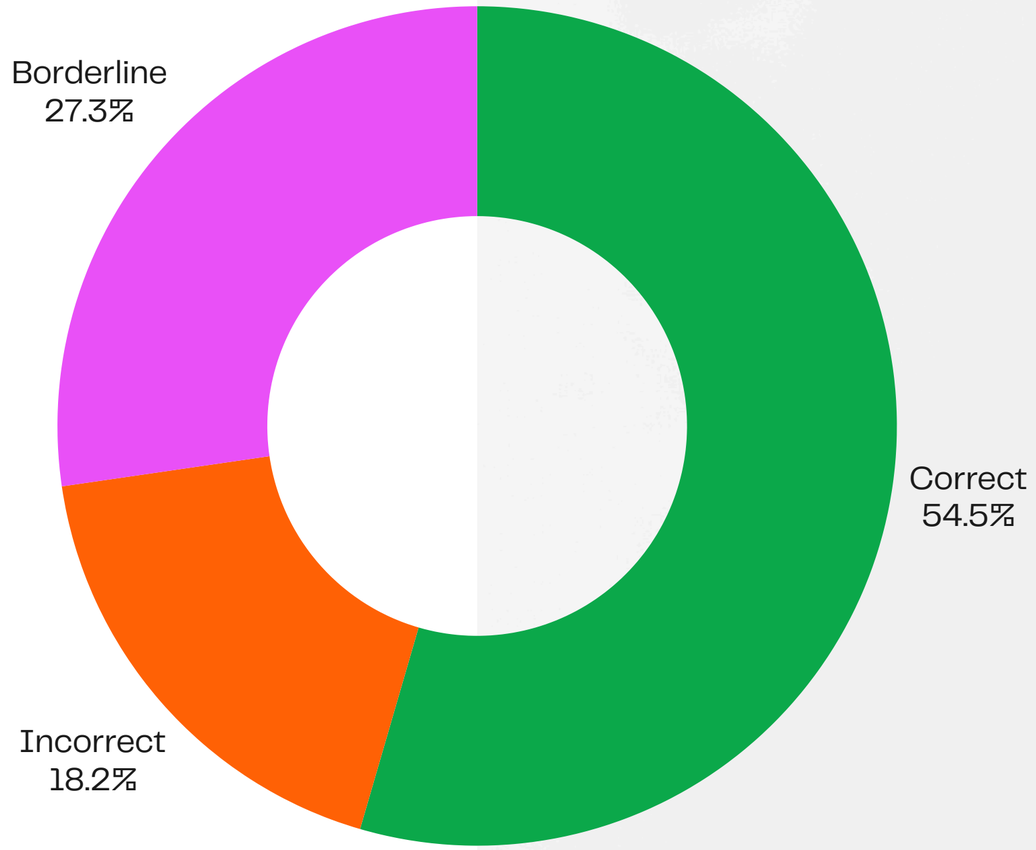
04 Error categorization:

- model limitations
- taxonomy gaps
- data quality issues

Qualitative Evaluation of Classification Results

Evaluation methodology

- Each company was manually compared against its description, category, and niche to determine whether the assigned insurance label accurately reflects its primary business risk.
- Most errors occur when the taxonomy does not contain a sufficiently specific label, forcing the model to choose the closest available alternative. This highlights taxonomy limitations rather than model failure.



Examples of Qualitative Results

Correct Classification

Food producer → Frozen Food Processing

Borderline classification

Utility Construction Company → Project Management Services

Incorrect classification

Auto body shop → Boiler Repair Services

Strengths and Weaknesses

Strengths

- Explainable and auditable decisions
- Robust to noisy and heterogeneous text
- Uses structured data effectively
- Efficient and scalable

Weaknesses

- Forced classification when taxonomy lacks coverage
- No explicit ‘Unmapped’ class
- Performance depends on label granularity

Scalability Considerations

Current performance

- Efficient sparse matrix operations
- Fast similarity computation for hundreds of labels

Scaling strategy

- Batch processing and sharding
- Approximate nearest neighbor search for large label sets
- Periodic or incremental re-training

Future Improvements

Planned enhancements

01 Introduce an explicit Unmapped / Other label

02 Enrich taxonomy labels with descriptions and synonyms

03 Active learning on low-confidence cases

04 Hybrid fallback using embeddings for ambiguous cases

Thank you for your time and attention

GitHub Link