Linguistic Patterns in European Public Organization Names

Álvaro del Ser [a,*] and Carlos Badenes-Olmedo [b]

[a] *Ontology Engineering Group, Universidad Politécnica de Madrid, Spain*
E-mail: alvaro.fontecha@upm.es
[b] *Ontology Engineering Group, Universidad Politécnica de Madrid, Spain*
E-mail: carlos.badenes@upm.es

Abstract. This work addresses the challenge of classifying public sector organizations across multiple European languages using only their official names, a critical step for entity disambiguation in knowledge graph population. We employ ontology-based knowledge extraction to evaluate three Natural Language Processing approaches: rule-based keyword extraction, zero-shot Natural Language Inference, and embedding-based semantic similarity —under low-context, low-resource assumptions. Large Language Models are integrated accross all three techniques. Our methodology systematically evaluates multilingual preprocessing, various state-of-the-art models, different supervision regimes, classification structures, and parameter optimization. We conduct a detailed evaluation across three specific domains (healthcare, administration, education) spanning multiple European countries, analyzing performance in relation to lexical structure and class balance.

Results demonstrate that lightweight rule-based methods, particularly TF-IDF keyword selection, are effective in multilingual scenarios with minimal supervision. Natural Language Inference models offer competitive zero-shot performance but show deficiencies with unbalanced class distribution. Embedding-based methods provide the most consistent generalization across languages, with evidence of class coherence in vector space. We apply these techniques to a real-world use case — classifying contracting authorities in the EU Contract Hub platform - and outline additional applications and extensions for governance objectives and ontology refinement. This work highlights the feasibility of ontology-guided multilingual classification from short texts and its contribution to entity disambiguation challenges in formal knowledge representation systems, particularly when integrating diverse European organizational entities into structured knowledge bases.

Keywords: Ontology Extraction, Natural Language Processing, Multilingual, Low-Context, Low-Resource

*Corresponding author. E-mail: alvaro.fontecha@upm.es.

1. Introduction

Proper names can, in certain domains, carry descriptive information about an entity's function, scope, or affiliation. In the case of organizations, names frequently reflect institutional roles, legal status, or sectoral domains. This quality renders them particularly valuable in low-context settings, where such names may constitute the only accessible representation of an entity and, when systematically extracted, can support ontology-driven classification and entity disambiguation in knowledge graphs.

Inferring information about entities based solely on their names is a foundational task in natural language processing and knowledge engineering. It supports semantic interpretation, entity linking, and structural categorization when richer metadata is unavailable. Robust name-based classification methods are therefore essential, not only for extracting key attributes, but also for enabling accurate entity resolution, knowledge graph population, and data integration processes. In the context of European public institutions, a finer entity categorization provides deeper insights, enabling knowledge engineers to not just identify who organizations are, but but to properly classify them according to their functional domains and operational roles, enhancing the semantic richness of knowledge representations. Improved access to structured organizational data supports multiple governance goals:

- **Efficient Public Procurement:** Improved data may allow institutions to better identify target groups for funding or policy interventions through precise entity linking.
- **Transparency and Market Integrity:** Detailed classification of organizations improves transparency and facilitates cross-dataset entity resolution, reducing ambiguity when populating knowledge graphs.
- **Regulatory Harmonization:** Standardized categorization enables cross-national entity alignment, ontology mapping, and consistent knowledge representation across European information systems.
- **Private Sector Oversight:** Visibility into organizational structures aids in detecting conflicts of interest, tracing beneficial ownership, and identifying systemic vulnerabilities during crises.

These challenges are directly reflected in the author's involvement in an EU-funded study on healthcare public procurement. ProCure, formally titled "Public Procurement Assessment in the Healthcare Sector," is an EU-funded initiative involving multiple European countries aimed at assessing and enhancing public healthcare procurement practices, particularly in the aftermath of health crises such as the COVID-19 pandemic. Its central goal is to identify best practices and strengthen procurement resilience across Europe, requiring precise categorization of healthcare entities—particularly hospitals—to inform policy decisions effectively. One specification of the data analysis tool named developed for this study, named EU Contract Hub [29], is the accurate identification of hospitals and university hospitals among public procurers across Europe. This task exemplifies the broader difficulties discussed next: operating in a multilingual, data-sparse environment where organization names serve as the primary source of semantic information. The methods developed in this work aim to support such classification tasks, while also contributing to the larger goals of knowledge graph population and semantic interoperability across European information systems [19] [8].

1.1. Problem Definition

Semantic classification of organization names in a cross-national European context presents several non-trivial challenges for entity disambiguation and knowledge graph population. **First**, the multilingual landscape comprising 24 official EU languages and several co-oficial or regional variants introduces substantial linguistic variation. Particularly public sector naming conventions tend to integrate regional languages more commonly. This study spans organization names in 29 European languages, covering Germanic, Romance, Slavic, Uralic, Baltic, and Celtic families, as well as typological isolates such as Basque and Maltese. These languages differ in morphological complexity, compounding, and orthographic norms. Highly inflected languages (e.g., Finnish, Hungarian) and those with extensive lexical compounding (e.g., German, Swedish) pose particular challenges for tokenization, rule-based methods, and semantic segmentation. From a practical standpoint, multilingual classification requires distinguishing between an organization's official name and its available translations—an often ambiguous task when sourcing entity labels from Knowledge Bases (e.g. Wikidata). **Second**, data harmonization remains a major obstacle for entity alignment. National databases often function in isolation with limited interoperability. This fragmentation restricts cross-border

data integration and complicates efforts to construct a unified, multilingual corpus for entity disambiguation tasks. **Third**, the lack of unified language processing tools for multilingual classification complicates implementation. Deploying a set of comparable language-specific models and processing techniques introduces design constraints and computational overhead. Furthermore, substantial variability in model performance across languages derived from these disparities in the underlying language models' coverage and task difficulty, complicates direct comparison of classification scores. **Fourth**, the inherent brevity and semantic ambiguity of proper names significantly hinder entity disambiguation accuracy. Traditional bag-of-words approaches, including TF-IDF, rely on term frequency patterns across longer texts. Limited lexical diversity and lack of statistical signal render many standard methods poorly suited for short-text classification tasks. **Finally**, the task requires a pragmatic balance between model complexity and operational efficiency. Given its narrow scope, organization name classification does not justify the computational overhead of large-scale models. Instead, lightweight methods—appropriately tuned—often yield competitive results with greater interpretability and lower resource requirements, which is essential for practical knowledge graph population workflows.

To address these challenges, this study proposes leveraging structured knowledge bases (i.e. Wikidata) to generate the ground truth labels for entities. By querying property hierarchies via SPARQL, this method enables adaptable ground truth generation, which can be repurposed for any classification tasks within the knowledge graph's semantic scope. This study implements multiple classification strategies. Rule-based systems serve as an interpretable baseline, though they require some training data. Our baseline method includes expert-informed rules derived from domain knowledge in healthcare. A structured questionnaire was used to elicit classification heuristics from practitioners, enabling the design of transparent and domain-specific rule sets. Natural Language Inference models further extend this by allowing flexible application without task-specific annotated data, offering a robust solution in low-resource settings. Embedding-based methods were introduced to capture the semantic information of textual names. These leverage sentence-level vector representations, enabling classification initially through semantic similarity, then trained classifiers that are able to derive more complex decision boundaries for entity classification.

The evaluation involved a systematic comparison of state-of-the-art model performances in entity disambiguation tasks. Multiple preprocessing pipelines were tested on rule systems to assess their impact on classification outcomes, and extensive hyperparameter tuning was conducted to optimize model behavior. The experimental framework was designed for modularity and reproducibility, enabling robust identification of the most effective strategies across varying data conditions and domain requirements.

1.2. Research Questions

Building on the methodological foundation outlined above, this study formulates a set of research questions to guide the empirical evaluation of classification approaches. These questions are designed to assess not only the performance of different models and techniques but also their ability to capture meaningful semantic patterns and domain distinctions for entity disambiguation in multilingual organizational data.

- **RQ1:** How effectively can general-purpose Natural Language Processing (NLP) techniques paired with knowledge graph data distinguish between medical, government, and educational organizations in multilingual, low-resource environments to support entity linking processes?
- **RQ2:** In which aspects do naming conventions of public sector organizations exhibit significant semantic variation across different European Union members, and how effectively can these variations be captured, interpreted, and mitigated using multilingual NLP techniques enhanced by knowledge graph resources for consistent entity disambiguation

2. Background

The classification of organization names has evolved significantly in recent years through the integration of structured semantic resources, hybrid information-extraction techniques, and enriched vector-space representations. First, knowledge graphs and ontologies have proven essential for providing supervision and explainability to classification systems. Resources such as Wikidata allow structured queries to retrieve key attributes of an organization—its

class, official name, and country of origin—which enhances classifier consistency and interpretability [26]. Complementarily, domain ontologies supply class hierarchies that serve as label sources, embedding symbolic knowledge into NLP pipelines and filling gaps left by purely data-driven methods [18, 27].

Nevertheless, the inherent ambiguity and variability of organization names demand dedicated disambiguation and entity-resolution strategies. Hybrid approaches combining string similarity, lexical rules, and minimal human oversight have proven highly effective at aligning textual mentions with their canonical forms [1]. Furthermore, recent generative systems like mGENRE have transformed multilingual entity linking by generating knowledge-base identifiers (e.g., Wikidata QIDs) for ambiguous mentions—thereby not only disambiguating them but also exposing structured properties that further support classification [9].

The challenge intensifies in low-context, multilingual scenarios, where large transformer models often underperform without external support. Shared tasks such as MultiCoNER have shown that multilingual transformers require additional mechanisms—like Wikipedia-based retrieval, translation-based augmentation, and weak supervision—to improve recognition of ambiguous proper names in low-resource languages, underscoring the importance of enriching context beyond model scale [21].

Finally, embedding-based methods have advanced to incorporate structured semantics from knowledge graphs. Through contrastive learning, prototype-based classification, and graph neural networks, these approaches align name embeddings with class-level semantics, enabling effective zero-shot and multilingual classification of organization entities [18, 27]. By marrying the flexibility of vector spaces with the depth of symbolic information, they achieve a balance that enhances both precision and generalization.

Building on these hybrid advances, in this work we apply three complementary strategies—(i) interpretable rule-based heuristics, (ii) zero-shot classification via multilingual NLI, and (iii) embedding-based classifiers—to multilingual, low-context public-sector organization names drawn from three domains (medical, administration, education). We further leverage Wikidata's class hierarchy to ground our labels in an ontology, and we conduct a comparative evaluation across 24 countries, nested versus flat class structures, and an external EU Contract Hub benchmark. This lets us assess not only raw performance but also interpretability, cross-lingual generalization, and real-world applicability.

3. Methodology

This study presents a broad methodological framework designed to evaluate and compare classification strategies for identifying and disambiguating organizational entities in the public sector. The classification problem involves substantial language variations and limited contextual information, which challenges conventional NLP methods. To address this, we integrate knowledge graph technologies to develop interpretable, generalizable, and semantically grounded models for entity disambiguation. The methodological approach supports the study's overarching goal of assessing how effectively different rule-based, zero-shot LLM-based, and embedding-based techniques can classify organizations across domains and languages, facilitating accurate entity linking when populating knowledge graphs with organizational instances using data enriched through ontological structures from Wikidata.

The research questions are evaluated through a structured series of experiments, each designed to test the specific capabilities of different classification approaches under controlled conditions. **RQ1**, which investigates the effectiveness of classification techniques across languages, we focus on scenarios with highly unbalanced classes. This condition is intrinsic to the problem of extracting specific organizational types from a broader institutional dataset when performing entity linking tasks. To address this class imbalance and task specificity, we adopt F1-score as our primary evaluation metric. Recall will also provide insight into the model's ability to identify relevant instances. For each classification strategy (rule-based, zero-shot, and embedding-based) we implement multiple model variants and preprocessing schemes, and conduct hyperparameter optimization where applicable. Comparisons across methods are systematically presented, including a rule-based classifier benchmark grounded in domain knowledge from experts via structured questionnaires. **RQ2** focuses on a comparative analysis of models performance across countries, centered specially on the more interpretable rule-systems, their optimal parameters and keywords, and aims to identify standardized guidelines for processing multilingual names in a consistent and scalable manner to support entity disambiguation in cross-lingual knowledge graphs.

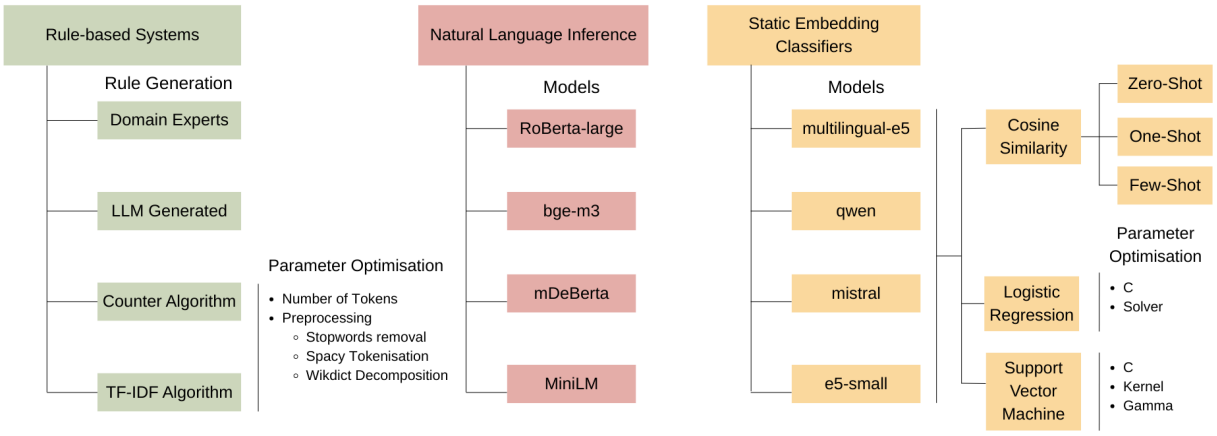


Fig. 1. Evaluated methods, models and parameter configurations.

3.1. Ontology-Guided Data Extraction

In the context of European public institutions, organization classifications facilitate accurate entity linking, knowledge graph population, and semantic interoperability. To this end, we focus on three pivotal domains: medical, administrative and educational. We regard the three domains as essential for public service delivery defined as a service of public utility, instead of public ownership. Accurately addressing classification challenges in the medical, governmental, and educational domains is crucial, as each domain directly impacts critical aspects of public welfare and governance. A cross-sectoral focus was adopted in order to draw generalizable conclusions and obtain flexible methodologies valid for inference in a variety of use-cases.

Wikidata is a large-scale, community-driven knowledge graph that we leverage to build semantically rich datasets for entity disambiguation tasks. By extracting entities from Wikidata instances, we obtain both the class from the hierarchical structure induced by the ontology and the name of organizations in all 29 official and co-official languages, creating a robust ground truth for entity linking experiments.

For class definition we refer to WikiData´s definitions of each as it will be our primary data source. We also include the WikiData identifier of each class.

- **Medical Domain**: This problem involves categorizing healthcare institutions into a nested structure, specifically distinguishing general hospitals from specialized institutions such as university hospitals.
  - Hospital (Q16917): Defined as a health care facility. However, it can be observed that in practice the hospital category only includes those facilities offering comprehensive and continuous care. In particular clinics are not typically categorized as hospitals, the key semantic and practical distinction is that hospitals provide inpatient care—patients can be admitted for extended periods of observation, surgery, or specialized care.
  - University Hospital (Q1059324): Defined as a hospital which is part of a university.
  - Specifically excluded categories comprise elder-care centers, rehab centers, research centers and medical clinics, for example.
- **Administration Domain**: This classification task addresses the identification of local administrations across EU member states.
  - Local Government (Q6501447): Defined as the lowest tier of administration within a sovereign state.
- **Education Domain**: The educational classification problem involves assigning institutions multiple relevant labels, such as primary and secondary education, due to the overlapping roles many educational facilities fulfill.
  - Primary School (Q9842): Defined as a school in which children receive primary or elementary education from the age of about five to twelve

* Secondary School (Q159334): Defined as an organization where secondary education is provided

WikiData's descriptions do not offer much insight into precise definitions for organizational subclasses. However, we will limit information on the classification task to the label in order to maintain alignment with the ontology. A potential direction for future research is to explore how definition-based approaches, where subclass descriptions are explicitly incorporated, might influence model performance in comparison to label-only methods for entity disambiguation tasks.

The use of properties `wdt:P31` (instance of) and `wdt:P279` (subclass of) provides a robust means to traverse the class hierarchy in the knowledge graph. This ontology-based approach ensures that we capture a sufficient scope and variety of organizations for entity linking purposes. Although our current work restricts these classes to three domains, there is an open possibility for exploring to what extent our category assignments mirror or deviate from the ontology's hierarchy, and whether trained classification models could enhance Wikidata's coverage by identifying missing relations or refining organizational taxonomies.

Dataset creation involved targeted SPARQL queries (as shown in listings 1 and 2) against the Wikidata endpoint to retrieve relevant instances per country and organizational domain. Due to endpoint limitations (query size and timeouts), queries were performed iteratively in smaller batches. An initial query extracts the instance URI of all subclasses of a certain category, followed by a second query that retrieves names and class information.

The process included:

1. Retrieving entity instances per domain-country combination using transitive class properties `wdt:P31` and `wdt:P279*`.
2. Downloading JSON raw output in manageable batches and performing basic error handling to recover from incomplete data responses.
3. Extracting multilingual labels for entities in each country's official and co-official languages. Wikidata does not offer a preferred language for labels, so we select the first label of the official languages of each country. We have decided against data augmentation by keeping multiple names per instance. Proper names contain very instance-specific tokens which makes them particularly prone to overfitting. Were this to be done, special care would also have to be taken to not pollute test sets with instances present in the train datasets.
4. Consolidating data into a structured format, including instance identifiers, country labels, class IDs, and multilingual textual labels.
5. To ensure balanced representation, we apply random sampling with a cap of 5,000 items per subclass-country combination, and maintain organization variety by also sampling from the national pool of organizations - double the number of instances of the most frequent class label in that country.
6. Annotating instances with binary class labels.

This methodology produces an annotated, varied dataset suitable for entity classification and disambiguation tasks. Future improvements could explore extraction from full Wikidata dumps to overcome endpoint constraints. The results are published on [28]. The final dataset exhibits an unbalanced distribution that we assume is representative of the reality of this problem. It is represented in Figure 2.

Our experiments are structured to reflect heterogeneous classification schemas, encompassing binary, nested, and multi-label tasks. In the medical domain, we compare both flat and nested classification to distinguish general hospitals from specialized entities such as university hospitals. For the governmental domain, a standard binary classification is applied to identify whether an entity functions as a local government authority. The educational domain requires multi-label classification to account for institutions providing multiple educational levels.

Given that our dataset in some countries lacks sufficient instances across all classes of interest, we necessarily restrict certain training-based methods to those countries with robust, representative coverage. This ensures that our classifier metrics are not skewed by sparse data and remain reliably interpretable. Meanwhile, techniques that do not rely on supervised training or language-independent are still applied. Otherwise, we selectively include for evaluation only those countries meeting a minimum data threshold to ensure that reported performance metrics reflect generalizable results.

In this work, EU Contract Hub's [7] not only acts as motivation, but also as a practical benchmark to assess the applicability of our classification methods in real-world scenarios. We utilize processed European Contract Award

data, specifically focusing on contracting authorities associated with medical product procurement (Common Procurement Vocabulary Division 33), in line with our domain requirements. From this subset, we manually annotated a balanced dataset of 150 contracting authorities across multiple countries and contract years. Annotations were informed by official organizational websites and geolocation data from widely used mapping platforms. Notably, a significant number of local governments were also identified, allowing us to extend our evaluation to the governmental domain using this same gold-standard dataset.

3.2. Classification Experiments Design

In this study, we consider several architectures suitable for classifying organization names, selected in accordance with the multilingual and low-context nature of the entity disambiguation task. Our approaches are organized into three families: **(i) rule-based heuristics**, leveraging interpretable country and domain-specific keyword patterns; **(ii) zero-shot classification using Natural Language Inference (NLI)**, which enables label prediction using large language models (LLM) without training; and **(iii) lightweight embedding-based classifiers**, including logistic regression and support vector machines (SVMs) trained on static semantic representations.

While various low-supervision classification methods were considered, including SetFit, Prototypical Networks, and fine-tuned language models, all were excluded for the same reason: they involve adaptation of the embedding space. Our objective is to evaluate the semantic structure encoded by general-purpose pretrained models without introducing task-specific optimization or fine-tuning. To preserve the integrity of our evaluation and ensure that results reflect the semantic meaning in organization names, we restrict our methodology to fixed-representation and inference-only approaches.

We adopt a low-resource computational restriction by prioritizing rule-systems, efficient NLP models and locally-run LLMs, motivated by accessibility, sustainability, and task suitability. Deployment in public sector settings often goes hand-in-hand with a limited access to high-end computational infrastructure, necessitating cost-effective and broadly accessible solutions. Environmentally, we strive to implement more sustainable AI practices, and for few tokens tasks like organization name classification, smaller models can offer sufficient performance. Methods such as model distillation, quantization, and pruning further support the use of compact architectures without substantial performance loss. Nonetheless, this approach entails trade-offs: under-performance, generalization issues, and their use on local machines imposes temporal and computational constraints. Despite these limitations, the low-resource paradigm proves essential and practical in resource-constrained environments. In terms of model selection we also restrict our choices to open-access models to ensure transparency, reproducibility, and suitability for public sector deployment.

Our methodology also compares different classification structures across domains, recognizing that task complexity influences model performance. In particular, we evaluate a nested classification scheme in the medical domain and contrast it with a flat three-class approach. As the number and structure of classes increase, either across or within domains, so does task complexity. It is therefore important to account for these differences when interpreting performance comparisons.

To ground our analysis in the current state of the art, we draw on recent advances in zero-shot classification and semantic representation. To substitute some training processes, we draw on the generative capabilities of `deepseek-R1` [20] to synthesize rules for classification. DeepSeek-R1 was selected due to open-source accessibility and best performance as of March 2025. Benchmark evaluations indicated that DeepSeek-R1 matched or exceeded the performance of leading proprietary models. The model's architecture, utilizing a Mixture-of-Experts (MoE) framework with 671 billion parameters and 37 billion active per query, allowed for efficient computation and local deployment of distilled versions. Natural Language Inference (NLI) models trained on the XNLI corpus [4], including `xlm-roberta-large-xnli`, `mDeBERTa-v3-xnli`, and `multilingual-MiniLMv2-mnli-xnli`, have demonstrated strong cross-lingual generalization and are widely used for multilingual zero-shot classification tasks. Complementary to this, embedding-based models such as `bge-m3-zeroshot-v2.0`, `e5-mistral-7b-instruct` [15] and `gte-Qwen2-7B-instruct` [3] provide dense vector representations optimized for classification, clustering, and retrieval.

We adopt the Massive Text Embedding Benchmark (MTEB) [16] and its multilingual extension MMTEB [17] as methodological foundation for our selection of models. These frameworks cover over 100 languages and a broad

range of tasks, including classification, clustering, and semantic search, offering a reproducible basis for comparing multilingual embedding models.

3.2.1. Rule-Based Systems

Our rule-based approach consists of an initial extraction of domain-specific keywords, to then apply regular expression (regex) matching against organization names using inclusion list logic. An entity is assigned to a class if its name contains one or more predefined keywords associated with that class. This method produces immediately interpretable results directly traced to a set of lexical features. Rule-based systems are particularly well-suited for domains with standardized naming conventions and provide transparent entity disambiguation decisions when populating knowledge graphs. However, this approach is region-dependent. As a result, effective keyword selection must be tailored to each language and regional context. This makes possible to study regional variation in naming practices. For example, healthcare institutions in France frequently include terms like *Hôpital* or acronyms like *CHU*, while Spanish entities may use the term *Hospital*. Our keyword extractor will identify these terms and classify as positive any name containing these substrings.

The interpretability of these systems can be exploited to integrate manual input. To establish a benchmark for medical classification, we leveraged domain expertise available within the ProCure project consortium. A questionnaire was distributed to healthcare procurement professionals from several participating countries, asking them to identify linguistic patterns commonly found in hospital names. The survey included the following sections:

- **Hospital Inclusion List:** Keywords that typically appear in hospital names (e.g., *Hôpital* in France, *Krankenhaus* in Germany).
- **Hospital Exclusion List:** Terms that may be healthcare-related but should not be associated with hospitals (e.g., *Clinics*, *Nursing Homes*, *Daycare Centers*).
- **University Hospital Inclusion List:** Terms exclusive to university hospitals (e.g., *Universitätsmedizin* in Germany).
- **Additional Rules & Comments:** Open-ended field for respondents to share regulatory constraints or naming conventions specific to their countries.

We received responses from experts in Austria, Italy, and France. While the inclusion lists entries were clear and actionable, the exclusion list proved ambiguous for several respondents and was ultimately excluded to avoid inconsistencies. Notably, the open-ended comments revealed that in some countries, national regulations require specific terms or acronyms to appear in official hospital names, further reinforcing the suitability of rule-based systems for the task.

To evaluate whether automated keyword generation could replicate expert-provided patterns, we prompted a generative language model for the keyword generation sub-task. Specifically, we used `DeepSeek-R1-Distill-Qwen-7B`, a 7B-parameter instruction-tuned model, deployed locally via the Ollama framework to ensure control over inference, reproducibility, and adherence to data constraints. DeepSeek was selected for its open nature, strong multilingual capabilities, competitive performance, and efficient local deployment compared to other LLMs of similar scale. The model was queried with the following standardized prompt mirroring the survey distributed to experts, obtaining 5 keywords per class and country:

```
Return a structured text response containing a dictionary. Your output **must strictly** be a valid JSON
    ↪ dictionary without any additional text, explanation, or formatting.

### Structure:
The dictionary will strictly have the 27 following keys, each corresponds to a EU Member State in this
    ↪ equivalence:
Q29: Spain, Q45: Portugal, Q142: France, Q233: Malta, Q41: Greece, Q38: Italy, Q183: Germany, Q31: Belgium,
    ↪ Q55: Netherlands, Q34: Sweden, Q33: Finland, Q211: Latvia, Q191: Estonia, Q37: Lithuania, Q36: Poland
    ↪ , Q28: Hungary, Q218: Romania, Q214: Slovakia, Q213: Czech Republic, Q215: Slovenia, Q40: Austria,
    ↪ Q219: Bulgaria, Q224: Croatia, Q229: Cyprus, Q35: Denmark, Q27: Ireland, Q32: Luxembourg

Each country code must be mapped to a **list of exactly five keywords** that commonly appear in hospital
    ↪ names in that country.

### **Hospital Inclusion List**
- These keywords will be used for **regular expression-based name matching** to identify hospitals from **
    ↪ organization** names.
```

```
- Keywords must specifically appear in hospital names within the given country. It is important they do not include other healthcare facilities.
- Consider all official and widely used languages of the country.
- What keywords typically appear in **hospital** names in each country?
- Select terms that uniquely distinguish **hospitals** from other **organizations**.

### **Final Output Rules:**
1. **Must be a valid JSON dictionary** no extra text, explanations, or formatting.
2. **Each country must have exactly five keywords** in its respective languages.
3. **No additional commentary or metadata.** Return only the JSON object.

### **Example Output (Format Only, Not Real Data):**
For example your output will begin: {'Q29':["keyword1", "keyword2", "keyword3", "keyword4","keyword5"], 'Q45':[...

Ensure your response **strictly follows** this structure. No additional text or formatting is allowed.
```

Outputs were generated in two batches per class to prevent prompt overflow, each covering half of the countries. All templates and responses are available in our GitHub repository.

Model	DeepSeek-R1-Distill-Qwen-7B
Architecture	Distilled, Decoder-only transformer, instruction-tuned variant of Qwen
Parameters	7 billion
Training Objective	Instruction-following and task generalization
Languages	20+ languages
Deployment	Locally run via Ollama
Purpose	Automatic keyword generation

Table 1
Model Card: DeepSeek-R1-Distill-Qwen-7B

Finally, we implemented two lightweight supervised keyword extraction methods tailored to the short-text nature of organization names. The first is a frequency-based heuristic that identifies the most frequent tokens associated with positively labeled examples for each class. The second leverages TF-IDF vectorization combined with chi-squared statistical testing to select the top-k discriminative tokens. In this setting, the term frequency (TF) component offers no variance as token repetiton is rare and non-relevant in this case. Despite this, IDF remains effective at highlighting class-distinctive terms that appear infrequently across the overall corpus but consistently within particular classes. While grounded in classical bag-of-words representations, this approach adapts well to low-token scenarios. In future extensions involving a broader label space, this method could be scaled to treat each class instances as an aggregated document, allowing global TF and IDF to work together for classification. In both metods, we optimized the number of keyword tokens by selecting the configuration that achieved the highest F1 score on the evaluation set with token numbers ranging from 3 to 10.

To support keyword extraction, we implement a multilingual preprocessing pipeline tailored to the challenges of organization name classification. The first step is stopword removal, applied based on the most prevalent official language of each instance's country. This approach, though it may not capture the language of every name, ensures a common standardization while avoiding the often unfeasible task of obtaining comprehensive stopword lists across regional and under-resourced languages.

The inherent difficulty of building multilingual NLP pipelines is especially pronounced in tokenization, where language-specific morphological patterns require tailored solutions. A key example is the prevalence of compound words, which are common in Germanic languages and pose a challenge to general-purpose tokenizers. Ideally, such names should be segmented into their lexical components to expose semantically meaningful units for downstream tasks. In early experiments, we applied spaCy-based tokenization and lemmatization using language-specific models, but found these methods inadequate for reliably decomposing such compounds. For instance, the Austrian school name *Bundesgymnasium und Bundesrealgymnasium Gleisdorf* is tokenized by spaCy and WikDict as shown in Table 2.

Name	Bundesgymnasium und Bundesrealgymnasium Gleisdorf
SpaCy tokenization	Bundesgymnasium, Bundesrealgymnasium, Gleisdorf
WikDict decomposition	Bund, Gymnasium, und, Bund, Realgymnasium, Gleis, Dorf[1]

Table 2
Example Tokenization and decomposition.

To address this, we adopted a dictionary-based compound decomposition strategy using the Wikdict project. We downloaded SQLite lexicon databases per language and applied per-token decomposition. This method allowed us to segment compound words into known lexemes, improving the interpretability and discriminative power of tokens across languages. While not universal and far from the performance of language-specific decompounders, this approach proved more robust than tokenizers.

To evaluate the effectiveness of the extracted keywords, we extract the results from the regex-based classification approach using the selected tokens and compute performance metrics both globally and per country. This allows us to assess how well the method generalizes across different linguistic and geographical contexts.

Given the method's sensitivity to class imbalance and data sparsity, reliable evaluation is challenging for countries or labels with few instances. To ensure robustness, we limit quantitative analysis to countries where the least frequent label has at least 30 examples, based on manual inspection of prediction quality. While meaningful patterns can emerge with as few as 5 instances, this threshold avoids overinterpreting noise and ensures reported metrics generalize well.

3.2.2. Natural Language Inference

Natural Language Inference (NLI) enables zero-shot classification by evaluating the relationship between the organization name and a hypothesis representing each potential class label. This approach is particularly suited for low-resource contexts, as it does not require task-specific training data. We formulate the classification task as an entailment problem: for each organization name, we evaluate whether it entails the following hypothesis: *"This organization is a* `{class}`*"*. We use pretrained multilingual transformer models capable of producing entailment scores across languages.

The Zero-shot classification pipeline outputs scores for each hypothesis tested, which can provide flexibility in handling imbalanced datasets. However, for this a threshold selection, which requires some knowledge of the class distribution in a labeled training set, is needed. Since our objective for this method is to remain fully training-free, for each instance, we select the most probable label from the candidate classes and a generic 'other' class.

We followed a principled approach to model selection grounded in multilingual support, architectural diversity, and community validation. All models were sourced from Hugging Face hub filtering by zero-shot classification NLP task.

Models were then prioritized based on **(i) explicit multilingual fine-tuning**, **iii) endorsement from the research and practitioner community as indicated by popularity metrics** (e.g., number of "likes" on the platform), and **(iii) architecture variety**, including lighter models. This ensured the inclusion of models that are both methodologically sound and practically validated by wide adoption. The final selection includes four models that span different transformer families, parameter sizes, and multilingual training coverage:

- **`joeddav/xlm-roberta-large-xnli`**[2] is based on the XLM-RoBERTa architecture[6], a multilingual extension of RoBERTa. This model is fine-tuned on the XNLI (Cross-lingual Natural Language Inference) corpus[5], which comprises approximately 550,000 sentence pairs in 14 languages, including 10 European languages: French, Spanish, German, Greek, Bulgarian, Russian, Polish, Portuguese, and Romanian. It serves as a strong zero-shot multilingual classification baseline due to its broad language coverage and stable performance across tasks.

[1] Translations: *Bund* = federal, *Gymnasium* = high school, *Realgymnasium* = science-oriented high school, *Gleis* = track, *Dorf* = village.
[2] https://huggingface.co/joeddav/xlm-roberta-large-xnli

- **`MoritzLaurer/bge-m3-zeroshot-v2.0`**[3] multilingual variant fine-tuned from BAAI/bge-m3 for zeroshot classification in 100+ languages and with a context window of 8192 tokens. It is based on the M3-Embedding architecture [2]. This model integrates dense, sparse, and multi-vector retrieval capabilities and leverages self-knowledge distillation for enhanced multilingual and multi-function performance. The NLI-based zero-shot classification training approach is described in [12].
- **`MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7`**[4] is a multilingual adaptation of the DeBERTa-v3 architecture[10]. The model is fine-tuned on a multilingual NLI corpus combining XNLI, MultiNLI[25], and WANLI[14], along with high-quality translations.
- **`MoritzLaurer/multilingual-MiniLMv2-L6-mnli-xnli`**[5] is a compact multilingual NLI model based on the MiniLMv2 architecture[24], which distills knowledge from larger transformer models by reducing depth and width while maintaining effective attention mechanisms. It is fine-tuned on both MNLI and XNLI datasets and supports a wide range of languages.

Consistent with our prior methodology, in the medical domain we evaluate both a nested classification approach and a flat classification strategy. Additionally, we experimented with augmenting the hypothesis prompts using one-shot and few-shot examples. However, preliminary results indicated low performance, likely due to confusion introduced by the added context in the entailment hypotheses.

Model	**xlm-roberta**	**bge-m3**	**mDeBERTa-v3**	**multilingual-MiniLMv2**
Architecture	XLM-R	XLM-R (RetroMAE)	DeBERTa	MiniLM
Parameters	561M	568M	279M	107M
Objective	XNLI fine-tune	Self-distilled contrastive + NLI	XNLI, MultiNLI, WANLI	MNLI + XNLI distilled
Languages	100 (pretrain.) + 15 XNLI (6 EU)	100+ (EU Coverage)	100 (15+ EU)	100 (pretrain.) + 15 XNLI (6 EU)
Deployment	Python, Hugging Face Transformers library.			
Purpose	Zero-shot classification via Natural Language Inference (NLI).			

Table 3
Selected NLI models.

3.2.3. Embedding-based Methods.

Embedding-based methods are instrumental to address the central research questions posed in this study. These methods rely on general-purpose multilingual sentence embeddings derived from pretrained transformer models, which incorporate the semantic content of organization names by extracting the contextual information embedded within their tokens. By converting organization names into dense vector representations, this approach yields semantically enriched features in a structured representation space. This strategy provides a mechanism for testing the ability of pretrained language models to generalize across linguistic and domain boundaries, supporting our investigation into multilingual classification performance and methodological robustness.

To assess the effectiveness of the embeddings, we construct a classifier using cosine similarity to class prototypes. An organization will be classified if its sufficiently close to (one of) the prototype(s). Three strategies are tested: a zero-shot setting using only the embedding of the class label, a one-shot configuration that with a single, randomly chosen, organization per class, and a few-shot experiment that introduces one example per class per country (ie. 24-shot) to better capture regional variation.

Positive classification is determined by comparing similarity against a class-uniform threshold. While this approach does not account for different semantic "widths" of each class, it provides an approachable starting point. By avoiding learning complex decision boundaries in this first iteration we also obtain a somewhat informative measure of class separability and semantic cohesion.

[3] https://huggingface.co/MoritzLaurer/bge-m3-zeroshot-v2.0
[4] https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7
[5] https://huggingface.co/MoritzLaurer/multilingual-MiniLMv2-L6-mnli-xnli

By comparing the performance of the similarity-based approach with that of the later experiments, we can evaluate the extent of conceptual dispersion the capacity of each method to accommodate hierarchical classes. Moreover, the inclusion of few-shot, country-specific examples allows us to explore whether regional linguistic and institutional variations are adequately captured by the class label or the multilingual encoding of a single example, or rather, there is a higher cross-country heterogeneity in organizational naming.

Given the limitations of fixed-threshold similarity measures, we transitioned to trained classifiers on Static Embeddings capable of learning optimized decision boundaries. The encoded organization names vectors serve as input features for logistic regression and Support Vector Machine (SVM) classifiers. We instantiated both logistic regression and SVM experiments with a range of hyperparameter configurations. Specifically, logistic regression models were tuned over combinations of regularization strengths $C \in \{0.01, 0.1, 1, 10\}$ and solvers $\{\text{liblinear}, \text{lbfgs}\}$, with penalty terms dynamically selected as L1 or L2 depending on solver compatibility. Similarly, SVM classifiers were tuned across $C \in \{0.1, 1, 10\}$, kernels $\{\text{linear}, \text{rbf}\}$, and applicable $\gamma \in \{\text{scale}, \text{auto}\}$ values for the radial basis function kernel.

To accommodate both multiclass and multilabel scenarios, all classifiers were wrapped using a One-vs-Rest (OvR) strategy, enabling the decomposition of complex tasks into independent binary subproblems. We also supports the training of nested classifier hierarchies, wherein distinct classifiers are instantiated for parent-child class relations or grouped categories. For binary classification tasks involving class imbalance, we employed balanced class weights, dynamically adjusted to mitigate bias toward majority classes.

In selecting models for the embedding generation task, we utilized the Massive Multilingual Text Embedding Benchmark (MMTEB), a comprehensive evaluation framework encompassing over 500 quality-controlled tasks across more than 250 languages, in particular the European languages benchmark.

GritLM, a high-performing model, was excluded from our evaluation due to their absence from the Sentence Transformers library, which we initially deemed necessary for compatibility with specific techniques. Additionally, while the broader MTEB leaderboard includes a wide range of models, not all are present in the MMTEB paper. To ensure coverage of recent and competitive architectures, we included `Qwen`, the best performing open model from MTEB with strong multilingual capabilities. Finally, we deliberately selected a lightweight model with fewer than 100 million parameters to benchmark performance under resource constraints.

- **`intfloat/multilingual-e5-large-instruct`**[6]: This model is based on the XLM-RoBERTa architecture with 24 transformer layers and produces 1024-dimensional embeddings. It is trained using a weakly-supervised contrastive objective on multilingual datasets and supports 100 languages, including all major European languages. It consistently achieves high accuracy on the Multilingual Text Embedding Benchmark (MTEB) tasks [16].
- **`Alibaba-NLP/gte-Qwen2-7B-instruct`**[7]: Built upon the Qwen2-7B architecture with 32 layers and 3584-dimensional embeddings, this model is instruction-tuned for embedding tasks using a multilingual training corpus. It achieved high performance on the Massive Multilingual Text Embedding Benchmark (MMTEB) [17].
- **`intfloat/e5-mistral-7b-instruct`**[8]: This model utilizes the Mistral-7B architecture with 32 transformer layers and 4096-dimensional embeddings. It is instruction-tuned on a large-scale multilingual corpus, resulting in extensive language coverage. As of March 2025, it ranks first on the MTEB leaderboard for both English and Chinese tasks, marking it as a leading multilingual embedding model [15].
- **`intfloat/e5-small-v2`**[9]: A compact alternative designed for efficiency, this model features 12 layers and 384-dimensional embeddings. It offers a balanced trade-off between performance and computational cost.

[6] https://huggingface.co/intfloat/multilingual-e5-large-instruct
[7] https://huggingface.co/Alibaba-NLP/gte-Qwen2-7B-instruct
[8] https://huggingface.co/intfloat/e5-mistral-7b-instruct
[9] https://huggingface.co/intfloat/e5-small-v2

Model	**Multilingual-E5-Large**	**E5-Mistral-7B-Instruct**	**GTE-Qwen2-7B-Instruct**	**E5-Small-V2**
Architecture	XLM-RoBERTa (24 lay.)	Mistral-7B (32 layers)	Qwen2-7B decoder-only	Custom (12 layers)
Parameters	560M	7B	7.61B	110M
Objective	Contrastive	Instruction-tuned	Instruction-tuned	Contrastive
Languages	100 (incl. EU languages)	Primarily English	Multilingual support	Primarily English
Embedding Dim.	512	4096	8192	512
Deployment	Python, Hugging Face Transformers library.			
Purpose	Embedding generation for classification.			

Table 4
Overview of selected multilingual sentence embedding models.

3.3. Evaluation

The primary metric used is the F1-score. This is particularly important given the imbalanced nature of our dataset, inherent to the real-world distribution. Therefore we must avoid relying on metrics sensitive to class frequency, such as accuracy. In addition to the F1-score, we complement with recall. Our classification objectives center on detecting very specific organizational types, and recall directly quantifies how many instances are successfully retrieved.

To accommodate varying levels of data availability across domains, we adopt domain-specific strategies for splitting the dataset into training, validation, and test sets. For domains with limited class coverage—particularly in the medical and governmental sectors—we apply a 50/25/25 split. Conversely, in the educational domain, where data availability is higher and class balance is more stable, we employ a more conventional 80/10/10 split.

For rule-based generation algorithms, rule sets are learned independently per country, partitioning training data into 24 subsets with uneven coverage. Due to scarcity (in some cases 0 instances per class and country), we restrict performance evaluation to countries for which sufficient labeled data is available. In contrast, for embedding-based zero-shot methods—such as cosine similarity—missing few-shot examples in specific countries are less critical, as these models are expected to generalize concepts across languages through their shared semantic space. Similarly, while training classifiers also requires labeled data, we assume that sufficient generalization can be achieved from the aggregate training set. Thus, our evaluation strategy adapts to the dependencies of each technique.

To benchmark the performance of our classifiers, we reference expert-curated rules as a comparative baseline. While such systems offer valuable insights grounded in domain knowledge, it is unrealistic to expect experts to cover the linguistic diversity and naming conventions across 29 languages and 24 countries. Evidently in the case of expert-input we evaluate our method only on the answered regions.

To evaluate the practical applicability and generalization of our models beyond the Wikidata-derived dataset, we additionally assess performance on a gold-standard validation set curated from the EU Contract Hub. By testing on this independent source, we aim to examine whether the performance observed on Wikidata generalizes to other data environments and institutional contexts. This secondary evaluation also allows us to explore the feasibility of deploying models trained on semantically structured data in operational settings involving heterogeneous, domain-specific datasets.

4. Discussion

As discussed throughout the evaluation, our three target domains exhibit differences in classification performance (Table 8), reflecting varying levels of structural complexity and lexical regularity in organization names.

All experiments were conducted on a local machine equipped with 16 GB RAM and 512 GB SSD, running on Apple's M4 architecture with Metal Performance Shaders (MPS) acceleration. We encountered memory constraints when loading language-dependant language models and when computing embeddings, underscoring the non-trivial resource demands of multilingual embedding tasks even in low-token scenarios. The complete code can be found on the Github [30]. Datasets and generated embeddings can be found on [28].

Domain	Structure	Max F1	Mean F1
medical	Nested (3 classes)	0.682510	0.354737
administrative	2 classes	0.892814	0.720648
education	Multilabel (3 classes)	0.775537	0.482317

Table 5
Mean and Maximum Scores per Domain.

4.1. Rule-based Approach

Our rule-based classification section comprises a total of 7 experiments each for the educational and administrative domains (not including parameter optimization), and 14 for the healthcare domain. Processing time was negligible in all cases, with all experiments completing within a matter of minutes. Complete results and charts are contained in Table 9 and Figures 4 and 5.

Performance in this approach could be interpreted in terms of how effectively domain-specific naming conventions can be captured through defined lexical patterns or token sets. The results suggest that, in many cases, the functional class or purpose of an organization can be inferred with reasonable accuracy from lexical information alone. This is particularly evident in domains such as healthcare, where terms like *hospital*, *clinic*, and *university hospital* are both semantically distinctive and cross-linguistically consistent.

Our benchmark experiment based on a manual rule set demonstrated limited generalization across the dataset, despite being informed by regulatory frameworks and institutional naming norms, the expert rules. Our findings suggest that even expert-informed linguistic rules often fall short in the face of real-world variability. In comparison with algorithmically obtained keywords. A contributing factor could be the ambiguity of Wikidata labels, which frequently do not reflect the standardized official names. As a result, expert rules may either be too technical or overly narrow, missing broader naming patterns—particularly in the healthcare domain. These challenges highlight the importance of flexible, data-informed approaches for robust multilingual classification.

Experiments with rule sets generated by LLM, aiming to automate the extraction of plausible lexical patterns for classification, performed reasonably well overall. Achieving scores that approached those of later, trained methods, they remained slightly inferior. Notably, they did not exhibit clear signs of hallucination or semantic implausibility; instead, their limitations were primarily due to insufficient lexical coverage across the diversity of national naming conventions. A particularly strong outlier experiment in the Finnish LLM-generated rules, significantly outperforms the other methods. To further contextualize this result, we performed a per-country evaluation of rule-based performance, focusing in particular on the healthcare domain, the most affected by limited data. This analysis revealed that classification effectiveness varies considerably by national context. Trained methods tend to under-perform in countries with little or no labeled data, while outperforming untrained approaches only in cases where the minority class has at least 30 labeled instances. While lexical pattern extraction can obtain relevant keywords even with very low data, they can fail to approach the complete distribution of names, especially if they do not have sufficient diversity. Although LLM-generated rules do not match the overall performance of well-trained models in high-resource countries, their low dependency on labeled supervision makes them a viable alternative in multilingual and data-scarce environments. A possible alternative and bridge between both approaches could be the generation of synthetic organization names instead of the rules. In any case, as mentioned in the methodology, we restrict our evaluation metrics to countries that ensure a minimum of 30 instances per class. This affects mostly the healthcare domain, though final scores do not differ much.

Among all rule-based methods, the best-performing approaches were the keyword selection algorithms, particularly the refined TF-IDF best-k-words method. This technique consistently outperformed the counter algorithm across domains. Notably, the optimal number of selected tokens differed substantially between both algorithms: the TF-IDF variant achieved comparable or superior performance using approximately half as many keywords. Token count also varied by domain, reflecting differences in lexical regularity. The administrative domain is the simplest in this sense, needing only 3 keywords to identify that class reliably.

In our comparison of preprocessing strategies, we observed only marginal improvements in classification performance. This is notable given the implementation challenges to deploy these techniques, particularly due to the

Country	No Preprocessing	Spacy tokenization	Decomposition	Difference (Spacy - None)
Poland	0.66	0.74	0.66	0.08
Croatia	0.66	0.71	0.66	0.05
Spain	0.68	0.73	0.63	0.04
Portugal	0.62	0.65	0.62	0.03
Country	No Preprocessing	Spacy tokenization	Decomposition	Difference (Decomposition - None)
Germany	0.67	0.68	0.70	0.03
France	0.54	0.54	0.56	0.02
Netherlands	0.58	0.59	0.60	0.02
Finland	0.57	0.58	0.58	0.01

Table 6
Counter and TF-IDF Experiments average F1 Score per Country and Preprocessing technique. Top 4 countries ordered by differences.

integration of multiple tokenization models and language-specific dictionaries. Among the tested methods, SpaCy-based tokenization slightly outperformed no decomposition and compound word decomposition. However, these are the aggregated findings. On a per-country level preprocessing does affect the final results mostly decompostion. Calculating the mean performance of counter and TF-IDF methods, we observe the most change in Poland, Croatia, Spain and Portugal using Spacy and in Germany, France and the Netherlands using Decomposition.

In terms of classification structure, a nested approach proved more effective strategy. This involves first extracting keywords to predict broad domain classes (e.g., hospitals), followed by finer classifiers trained on the corresponding subsets. By decomposing the task into stages, the nested architecture not only improves precision but also enhances interpretability.

Table 13 presents selected keywords that inform our classifiers. For illustrative purposes, we display a small subset of the results, focusing on two countries—France and Germany—while the full dataset spans 28 rows and 24 columns. Expert-informed keyword selection selects acronyms while wikidata-trained systems tend to rely on words. Furthermore, terms like *clinique* are included, which should not be categorized as hospitals. This supports our earlier point that Wikidata classes are not always fully aligned with the target concepts of our task, suggesting opportunities for future improvements in entity linking. Decomposition effects can clearly be seen in the French educational domain, where lexical cues such as the prefix *éle-* help distinguish terms like *élementaire*, enabling more accurate classification of school types.

Overall, our experiments demonstrate that lightweight rule-based methods can achieve remarkable classification performance, comparable to more sophisticated embedding-based approaches. These methods offer significant advantages in terms of computational efficiency and interpretability. Their success is further reinforced by the ontological grounding of the training data, which provides annotated instances. Variability and coverage within each class are critical factors and in some cases we suspect that results may be adversely affected by wikidata's composition, particularly when most instances within a class are overly homogeneous.

The results suggest that the functional role of a public sector organization is often strongly grounded in the lexical presence of specific keywords. Our approach constructs a country-specific lexical dictionary, capturing these discriminative features which reinforces the value of lexical pattern mining over even expert-informed knowledge, for tasks where domain semantics are linked to naming conventions.

However, the generalization of these approaches to more complex classification tasks remains an open question. Potential limitations include diminished effectiveness when applied to non-class properties, or a high number of classes. Performance also deteriorates with sparse labeled data, as LLM-generated rules have shown limited reliability in our experiments. Furthermore, the scalability of rule-based methods to longer input sequences and broader context windows has yet to be validated. While the current findings underscore the strength of ontology-guided lexical modeling in varied, multilingual environments, further research is needed to evaluate its robustness and adaptability in higher-complexity domains.

Model	Max F1
roberta-large	0.767278
bge-m3	0.772669
mDeBerta	0.674289
MiniLM	0.527614

Table 7
Maximum Scores per Model.

4.2. Natural Language Inference

For the Natural Language Inference (NLI) experiments, we evaluated four multilingual models using a consistent zero-shot classification pipeline. This setup requires minimal configuration but is computationally intensive: the largest models process approximately 11 instances per second, while the compact `MiniLMv2` model achieves around 40 instances per second. Across the full dataset, inference took approximately 24 hours to complete. Complete results and charts are contained in Table 10 and Figures 4 and 5.

Our results indicate that larger models deliver superior classification performance. In particular, `BGE-M3 ZeroShot v2.0`, which incorporates additional multilingual fine-tuning beyond `XLM-RoBERTa`, outperforms the `XLM-R`-based baseline across all domains. By contrast, `mDeBERTa-v3`, although highly effective in the medical domain, performs noticeably lower in the administrative and educational domains. While `MiniLMv2` offers significantly faster inference and reduced computational demands, it fails to match the accuracy of larger models, highlighting the trade-off between efficiency and performance in zero-shot NLI classification.

Interestingly, in contrast to our rule-based experiments, the nested classification structure did not lead to improved performance in the NLI setting. This outcome could stem from error propagation. Additionally, the NLI model may struggle to resolve fine-grained distinctions when presented with hypothesis classes that are semantically close but hierarchically nested. This is particularly problematic in our threshold-agnostic setup, where always selecting the most probable label can lead to misclassification . For instance, if the system consistently selects "university hospital" over "hospital" due to marginally higher entailment scores, it may skew the evaluation metrics, which penalize equally all misclassifications. These findings suggest that either threshold calibration or error design are critical when adapting nested structures to NLI-based pipelines.

We additionally experimented with augmenting the NLI prompts using one-shot and few-shot example. However, this strategy yielded very degraded performance. NLI models are optimized to assess the entailment relationship between a single hypothesis and premise pair. Introducing prior examples into the hypothesis may distract the model. As such, we do not include these experiments in our study.

These findings highlight the potential of rule-based systems as lightweight, interpretable alternatives in low-resource and multilingual classification settings. However, their effectiveness diminishes as the number of target classes increases. We suspect that, NLI-based models may suffer from increased confusion, and misinterpretation of the "other" class. One possibility for improvement involves implementing threshold-based classification. Preliminary binary classification experiments using thresholding showed promising results, suggesting that this strategy could mitigate over-prediction. Nevertheless, a critical challenge remains: obtaining the thresholds without over reliance on labeled training data or class distributions. Moreover, the stability of threshold-based approaches under different distributions must be assessed. Specifically, it remains an open question whether thresholds calibrated on Wikidata-derived samples would transfer reliably to new datasets with different class balances.

An additional possibility involves a definition-based classification approach. Rather than relying solely on label matching, this method would integrate class definitions into a structured set of inference questions about each instance. For example, in the case of university hospitals, classification could proceed by sequentially verifying whether the organization (i) functions as a healthcare facility, (ii) provides comprehensive general care distinguishing it from specialized clinics, and (iii) engages in the education of medical students. Each of these properties could be individually tested by suitably modifying the hypothesis formulation. Although the low-context setting limits the immediate applicability of definition-based classification, access to richer textual sources, such as descriptions obtained via web scraping, could substantially broaden its viability. A key advantage of employing NLI models lies

Model	Max F1 (similarity)	Max F1 (logreg)	Max F1 (svm)
multilingual-e5	0.654192	0.862211	0.885668
qwen	0.741955	0.868487	0.892814
mistral	0.793699	0.883531	0.682510
e5-small	0.574878	0.758679	0.862910

Table 8
Maximum Scores per Model.

in this flexibility: by appropriately modifying the hypothesis prompt, it becomes feasible to infer different ontological properties beyond simple class membership. This enables multi-property classification, ontology enrichment, and dynamic taxonomy construction, provided that sufficient and reliable contextual information can be gathered. Such an approach would align well with knowledge graph augmentation efforts and holds potential for expanding automated understanding of public sector organizational structures.

4.3. Embedding-based Classification

Embedding-based methods proved to be the most computationally intensive component of our experiments. The generation of embeddings for the full dataset required approximately 100 hours for the `Mistral-7B` and `GTE-Qwen2-7B` models, while even the smaller `Multilingual-E5-Large` model needed close to 20 hours. The embedding output files for `GTE-Qwen2-7B` and `Mistral-7B` individually exceeded 20 GB. Evaluation time of Cosine Similarity and Classifier experiments is negligible. Training of classifiers takes variable time depending on the dimension of embeddigs. Logistic Regression training can take from one hour to five while Support Vector Machines take between 8 and 60 hours (8 models are trained per experiment to tune the hyperparameters). These observations highlight that, while embedding-based classification can offer high performance, its deployment poses logistical challenges. Complete results and charts are contained in Tables 12 and 12 and Figures 4 and 5. In terms of performance, the initial zero-shot cosine similarity approach—matching embeddings directly to class labels—proved to be quite competitive relative to other techniques, especially considering that it operates without any labeled training data. Nevertheless, this remains a restricted form of classification and, overall, performs worse than NLI models. Furthermore, one-shot experiments, which use a single randomly selected example per class for comparison, consistently yielded lower scores than zero-shot label matching. This suggests that there exists semantic variability in the vector space within class clusters, and that instance distributions are not perfectly compact with respect to semantic similarity. In fact some one-shot examples tend to optimize distance lower than the label prototypes. In fact, this distance varies from domain to domain, with higher thresholds suggesting separability of the class. For the binary administrative domain, some distances are at 0.7, while on the others range from 0.1 to 0.3. Few-shot provides more granular boundary and tends to reduce the optimal distance to around 0.5, suggesting that the provided examples cover the initial zero-shot classification boundary well. They showed performance improvements in some cases but were highly dependent on the choice of encoder and the domain. Improved performance in few-shot setups could point to regional or linguistic dispersion instead of stronger class-specific cohesion. This phenomenon was slightly notable in the administrative domain and most evident with the `GTE-Qwen2-7B` embeddings. These observations imply that entity clustering is not very influenced on these multilingual embeddings by regional naming conventions rather than by functional class.

The use of per-class optimized thresholds—rather than a global decision boundary—could potentially have enhanced performance, especially in domains exhibiting higher semantic variability, but we have opted to relegate that study for the classifier experiments.

Across all evaluated embedding models, performance was relatively consistent, with comparable results achieved across domains. The base `Multilingual-E5-Large` model generalized particularly well in the zero-shot label matching task, delivering robust performance with less resource requirements. While the larger models—`GTE-Qwen2-7B-Instruct` and `E5-Mistral-7B-Instruct` offered slightly higher overall scores, the performance gains were marginal relative to their computational cost. In particular, `Mistral` outperformed

`Qwen2` across domains by a narrow margin. Given the minimal context provided by organization names, these results suggest that deploying large-scale generative embedding models may not be justified in such a constrained setting.

The similarity-based embedding approach performance in the zero-shot configuration, demonstrates effective generalization from class labels to unseen instances and highlighting the model's capacity to capture latent semantic structure. Despite the constrained nature of our experimental design, these results suggest that class clusters exhibit meaningful internal coherence, potentially skewed due to semantic overlap, but without strong alignment to regional variables.

Our final set of experiments focuses on improving decision boundaries using logistic regression and support vector machines (SVMs). In all domains, both classifiers outperform semantic similarity methods by a significant margin, making them the most effective approaches overall. However, the degree of improvement varies notably by domain. In the medical domain, the gain is most pronounced, as semantic similarity fails to capture the nested structure inherent in the data. Instead, we now train two classifiers, which are applied sequentially.

Logistic regression is generally less accurate than SVMs, although the computational cost of training SVMs is substantially higher[10].

Model size does not substantially alter performance across configurations. While larger models yield modest improvements, the gains do not appear to justify their additional computational cost. However, the `e5-small` model does show a clear performance drop with respect to the others.

Integrating embedding representations with supervised classification techniques consistently improved overall performance, making this the best-performing method across our evaluation. These findings offer partial validation of our hypothesis: that the embedding space captures latent semantic meaning sufficient for distinguishing public sector organization types. While embedding-based inference may not be suitable to map complex relational reasoning tasks, the potential of these embedding spaces for unsupervised analysis, such as clustering, warrants further investigation. We conjecture that such embeddings may not cluster primarily by language or country, but instead organize instances according to institutional function. This positions embedding-based techniques as a powerful and generalizable tool for multilingual classification and ontology discovery.

4.4. Evaluation on EU Contract Hub Use Case

We now turn to the evaluation of the previously trained Wikidata-based models using our gold-standard dataset from the EU Contract Hub. Earlier observations suggested that the ontology-derived data might lack the diversity and quality needed to generalize well. This is reflected in the evaluation results, where we observe a drop in performance across trained methods. Rule-based approaches perform worse than in previous experiments. Semantic drift plays an important part here, where we have seen clinics classified as hospitals in the Wikidata ground truth. Notably, expert-informed strategies fall short in reliably identifying all relevant instances. Natural Language Inference (NLI) models achieve therefore the best results, though their performance remains close to that of embedding-based models. Cosine similarity fails to classify correctly for the nested medical domain experiments. This is caused as mentioned by the nested structure and data sparsity of the new dataset (150 instances), prototype mismatch or semantic drift can exacerbate this issue.

5. Results and Conclusions

This study addressed the multilingual classification of public sector organization names across healthcare, administrative, and educational domains. We showed that lightweight rule-based methods—particularly TF-IDF keyword extraction—can achieve strong performance in low-context, multilingual settings. However, preprocessing remains a challenge for multilingual standardized compound handling and tokenization, with only marginal gains in classification scores. Regional differences continue to play a significant role in language processing, and no single method

[10] Due to unforeseen computational constraints, including a national power outage in Spain, SVM results for the education and administrative domains could not be completed in time. These will be reported in future work.

proves universally effective across all contexts. Our results suggest that the lexical presence of domain-relevant terms is a key factor, with meaningful keywords emerging consistently across different regions and languages.

Natural Language Inference (NLI) models demonstrated competitive zero-shot performance, yet they require careful calibration and are sensitive to fine-grained distinctions between semantically close classes. Their effectiveness underscores the potential of ontology-guided classification based on lexical cues, particularly in settings with minimal supervision.

Embedding-based approaches demonstrate consistent generalization across domains, enabling effective clustering of organization types based on latent semantic features. These methods revealed that class structure within the vector encodings is driven more by functional semantics than by linguistic variation. Also, larger models did not offer substantial performance gains over the multilingual-e5 model. Instead, performance was primarily dependent on the quality of the downstream classifier, suggesting that embedding-based pipelines can scale efficiently if well optimized.

Upon manual inspection, we observed a highly unequal distribution of instances across countries, which may point to systemic coverage biases or incomplete curation within certain regions. This uneven representation can negatively affect model generalization. Semantic drift has also been observed when inspecting rule-classifiers keywords, primarily due to the mislabeling of instances. This reflects a broader issue within Wikidata, where many classes are ambiguous or overlapping, making it unsuitable to serve as a ground truth without further refinement. The lack of clear class boundaries often leads to noisy supervision. However, it is worth noting that overall performance on the new dataset remains consistent with our previous findings. The conclusions drawn regarding effective methodologies, preprocessing strategies, and best practices continue to hold with some caveats, indicating that the training dataset retains practical utility. Nevertheless, the results also suggest that a larger validation set may be necessary to further support the robustness of these conclusions.

Practical applications of these findings are immediate and diverse. Within the EU Contract Hub project, the classification models are already being deployed to classify procurement contracts by contracting authority type, such as hospitals and university hospitals. Future work could extend this analysis to assess joint procurement participation by government entities, an area of strategic interest for fostering collaboration in digital infrastructure projects. Similarly, the classification of public schools could support geographic coverage studies and policy planning by analyzing distribution patterns. Additional organizational properties, such as public versus private ownership, could also be inferred with limited extensions to the methodology.

5.1. Future Work

This work opens several promising directions for future research. First, NLI pipeline would benefit from threshold calibration, particularly to address class imbalance and improve performance on hierarchical or semantically overlapping classes. In the embedding spaces, evaluating the distribution of class instances and conducting clustering analysis, using metrics such as silhouette scores, v-measure, or intra/inter-cluster distances, could support more interpretable, cluster-based classification strategies.

Second, the current study operates under a strict low-context assumption. Incorporating contextual information, such as organizational descriptions or web-scraped summaries, could significantly enhance classification accuracy, particularly for fine-grained or ambiguous cases. This additional information would also enable more sophisticated NLI strategies, such as decomposing class definitions into logical subcomponents and verifying them through targeted entailment queries.

Another future direction involves expanding the classification task beyond the current class property to include other organizational attributes—such as legal status (public vs. private, SMEs...) or relationships (eg. modeling of organization relations). By extending the classification task to other ontological properties, the methodology developed remains relevant and can help automated validation of the semantic structure. Furthermore, applying the proposed methods to private sector entities could test the generalizability of the approach outside of public administration.

Finally, a particularly impactful application lies in the integration of these models with semantic resources like Wikidata. Our methods could support knowledge graph validation by detecting inconsistencies, suggesting new class assignments, or recommending changes to existing taxonomies through structured inference. This would contribute

to automated ontology refinement and enhanced alignment between textual data and formal semantic representations.

5.2. Conclusions

Our main contributions include the development of a multilingual, low-supervision pipeline for organization name classification; a detailed comparative analysis of rule-based, NLI-based, and embedding-based approaches; and the demonstration of scalable methods that adapt to low-context, short-text scenarios. We also introduced preprocessing strategies for compound word decomposition and language-specific tokenization in resource-limited multilingual settings, and conducted a fine-grained evaluation across domains and countries, highlighting the relationship between lexical regularity and classification success.

The methodologies employed revealed distinct trade-offs. Rule-based methods offered interpretability and efficiency but struggled with dependency in training coverage. NLI-based approaches performed well with minimal labeled data, yet multi-class classification remained challenging without threshold tuning. Embedding-based methods delivered the highest overall performance when coupled with supervised classifiers, confirming that semantic spaces can effectively capture functional organizational attributes even in low-context inputs.

References

[1] F. Ancona, V. Beretta, A. De Marco and A. Graziosi, A novel methodology to disambiguate organization names: An application to EU Framework Programmes data, *Scientometrics* **128** (2023), 1111–1134.

[2] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian and Z. Liu, M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation, *arXiv preprint arXiv:2402.03216* (2024). https://arxiv.org/abs/2402.03216.

[3] Y. Chen, W. Zhao, H. Zhang and et al., GTE-Qwen2: A General Text Embedding Model for Retrieval and Classification, *arXiv preprint arXiv:2404.06881* (2024).

[4] A. Conneau, R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk and V. Stoyanov, XNLI: Evaluating Cross-lingual Sentence Representations, in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2475–2485.

[5] A. Conneau, R. Rinott, G. Lample, A. Williams, S.R. Bowman, H. Schwenk and V. Stoyanov, Xnli: Evaluating cross-lingual sentence representations, in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2475–2485.

[6] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer and V. Stoyanov, Unsupervised cross-lingual representation learning at scale, *arXiv preprint arXiv:1911.02116* (2019).

[7] del Ser, Alvaro and Ramon, Virginia and Olmedo, Carlos, EU Contract Hub Dashboard, 2024, Accessed: 2025-04-30.

[8] European Commission, Public Procurement: A data space to improve public spending, boost data-driven policy-making and improve access to tenders for SMEs, *Official Journal of the European Union* **C 98 I**(01) (2023), 2023/C 98 I/01.

[9] M. Garcia, V. Labatut and L. Padró, Context-Enriched Multilingual Named Entity Recognition Using External Knowledge, in: *Proceedings of SemEval 2022*, Association for Computational Linguistics, Seattle, Washington, 2022, pp. 1047–1055.

[10] P. He, X. Liu, J. Gao and W. Chen, DeBERTa: Decoding-enhanced BERT with disentangled attention, in: *International Conference on Learning Representations*, 2021.

[11] G. Lample and A. Conneau, Cross-lingual Language Model Pretraining, *Advances in Neural Information Processing Systems* **32** (2019).

[12] M. Laurer, W. van Atteveldt, A. Casas and K. Welbers, Building Efficient Universal Classifiers with Natural Language Inference, 2023, arXiv:2312.17543 [cs.CL]. http://arxiv.org/abs/2312.17543.

[13] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie and M. Zhang, Towards General Text Embeddings with Multi-Stage Contrastive Learning, *arXiv preprint arXiv:2308.03281* (2023). https://arxiv.org/abs/2308.03281.

[14] Z. Liu, X. Wang, H. Sun and X. Ren, WANLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation, *arXiv preprint arXiv:2201.12091* (2022).

[15] N. Muennighoff, M. Laurer and I. Gupta, E5-Mistral: Efficient Instruction-Finetuning for Multilingual Embeddings, *arXiv preprint arXiv:2403.01910* (2024).

[16] N. Muennighoff, N. Tazi, L. Magne, T.V.V. Le, C. de Masson d'Autume, S. Ruder, R. Tang, K. Cho, S. Swayamdipta and I. Gupta, MTEB: Massive Text Embedding Benchmark, *arXiv preprint arXiv:2302.04867* (2023).

[17] N. Muennighoff, T.V.V. Le, M. Laurer, A. Eisele, J. Bingel and I. Gupta, MMTEB: Massive Multilingual Text Embedding Benchmark, *arXiv preprint arXiv:2402.13595* (2024).

[18] M. Rizinski, M. Alvermann, K. Lochner and M. Stede, Comparative Analysis of NLP-Based Models for Company Classification, *Information* **15**(2) (2024), 534–553.

[19] A. Soylu, O. Corcho, B. Elvesæter, C. Badenes-Olmedo, T. Blount, F.Y. Martínez, M. Kovacic, M. Posinkovic, I. Makgill, C. Taggart, E. Simperl, T.C. Lech and D. Roman, TheyBuyForYou Platform and Knowledge Graph: Expanding Horizons in Public Procurement with Open Linked Data (2024), Accessed: 2025-04-30.

[20] D. Team, DeepSeek Embeddings: High-Performance Embeddings from Large Language Models, *arXiv preprint arXiv:2404.06651* (2024).
[21] B. Wang, R. Zhang and W. Sun, DAMO-NLP at SemEval-2022 Task 11: A Knowledge-Based System for Multilingual NER, in: *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, Association for Computational Linguistics, Seattle, Washington, 2022, pp. 1146–1154.
[22] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder and F. Wei, Improving Text Embeddings with Large Language Models, *arXiv preprint arXiv:2401.00368* (2024). https://arxiv.org/abs/2401.00368.
[23] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder and F. Wei, Multilingual E5 Text Embeddings: A Technical Report, *arXiv preprint arXiv:2402.05672* (2024). https://arxiv.org/abs/2402.05672.
[24] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang and M. Zhou, MiniLMv2: Multi-head self-attention relation distillation for compact language models, in: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 2140–2151.
[25] A. Williams, N. Nangia and S.R. Bowman, A broad-coverage challenge corpus for sentence understanding through inference, in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, 2018, pp. 1112–1122.
[26] W.X. Zhao, K. Zhou, M. Chang and J.-R. Wen, Connecting embeddings for knowledge graph entity typing, *Transactions of the Association for Computational Linguistics* **8** (2020), 181–196.
[27] K. Zhu, P. Xie and T. Zeng, BERT-KG: A Short Text Classification Model Based on Knowledge Graph and Deep Semantics, in: *NLPCC 2021*, Springer, Shanghai, China, 2021, pp. 273–285.
[28] Álvaro Fontecha del Ser, Dataset for Organization Names NLP: Multilingual Public Sector Entity Classification, 2024, Zenodo repository. doi:10.5281/zenodo.15312407.
[29] Álvaro Fontecha del Ser, EU Contract Hub, an Integrated Search and Analysis Tool for Public Procurement Contracts within the European Union, Master's thesis, Universidad Politécnica de Madrid, Escuela Técnica Superior de Ingenieros Informáticos, 2024, Master in Data Science.
[30] Álvaro Fontecha del Ser, OrgNamesNLP: Linguistic Patterns in European Public Organization Names, 2024, Comparative study of rule-based, NLI, and embedding-based NLP methods for subclass inference under low-context, low-resource conditions..

6. Annex I: Tables and Figures

Listing 1: Query to extract all instances of a class and its subclasses per country.

```
SELECT DISTINCT * WHERE {{
     ?item wdt:P31/wdt:P279* wd:{class_id};
           wdt:P17 wd:{country_id};
}}
LIMIT 100000
```

Listing 2: Query to link Instance ID to name and class.

```
SELECT ?item (GROUP_CONCAT(DISTINCT ?name; SEPARATOR=", ") AS ?names)
             (GROUP_CONCAT(DISTINCT ?class; SEPARATOR=", ") AS ?class_ids)
             (GROUP_CONCAT(DISTINCT ?classLabel; SEPARATOR=", ") AS ?classes)
WHERE {{
     VALUES ?item {{ {instances_str} }}
     ?item wdt:P31 ?class;
           rdfs:label ?name.
     FILTER(LANG(?name) IN ({languages_str}))
     ?class rdfs:label ?classLabel.
     FILTER(LANG(?classLabel) = "en")
}}
GROUP BY ?item
```

ID	Domain	Technique	Method	Accuracy	Structure	Preprocessing	Tokens	F1	Recall
med-r-expert-0	medical	rules	expert	0.956268	nested-class	None	N/A	0.148029	0.109266
med-r-llm-0	medical	rules	llm_generated	0.964596	nested-class	None	5	0.372017	0.534446
adm-r-llm-0	administrative	rules	llm_generated	0.907144	2-class	None	5	0.713588	0.689637
edu-r-llm-0	education	rules	llm_generated	0.825499	3-multiclass	None	5	0.338022	0.216542
med-r-counter-0	medical	rules	counter_algorithm	0.975849	3-class	None	7	0.437205	0.581476
med-r-counter-1	medical	rules	counter_algorithm	0.977055	3-class	Spacy tokenization	6	0.441113	0.567093
med-r-counter-2	medical	rules	counter_algorithm	0.982720	3-class	Decomposition	3	0.443624	0.427147
med-r-counter-3	medical	rules	counter_algorithm	0.982522	nested-class	None	7	0.472809	0.554880
med-r-counter-4	medical	rules	counter_algorithm	0.984144	nested-class	Spacy tokenization	6	0.486180	0.545816
med-r-counter-5	medical	rules	counter_algorithm	0.980594	nested-class	Decomposition	7	0.481839	0.554304
med-r-idf-0	medical	rules	idf_best	0.976386	3-class	None	3	0.444856	0.616430
med-r-idf-1	medical	rules	idf_best	0.978216	3-class	Spacy tokenization	3	0.458555	0.622497
med-r-idf-2	medical	rules	idf_best	0.975663	3-class	Decomposition	3	0.440440	0.616503
med-r-idf-3	medical	rules	idf_best	0.981481	nested-class	None	3	0.598167	0.605792
med-r-idf-4	medical	rules	idf_best	0.982435	nested-class	Spacy tokenization	3	0.612678	0.627816
med-r-idf-5	medical	rules	idf_best	0.983859	nested-class	Decomposition	3	0.585724	0.600546
adm-r-counter-0	administrative	rules	counter_algorithm	0.939678	2-class	None	3	0.822869	0.799975
adm-r-counter-1	administrative	rules	counter_algorithm	0.943688	2-class	Spacy tokenization	3	0.832702	0.805629
adm-r-counter-2	administrative	rules	counter_algorithm	0.936971	2-class	Decomposition	3	0.820697	0.807304
adm-r-idf-0	administrative	rules	idf_best	0.877142	2-class	None	3	0.739451	0.814084
adm-r-idf-1	administrative	rules	idf_best	0.918343	2-class	Spacy tokenization	3	0.800033	0.834796
adm-r-idf-2	administrative	rules	idf_best	0.874907	2-class	Decomposition	3	0.736436	0.812602
edu-r-counter-0	education	rules	counter_algorithm	0.823033	3-multiclass	None	5	0.649656	0.688904
edu-r-counter-1	education	rules	counter_algorithm	0.784982	3-multiclass	Spacy tokenization	10	0.615246	0.724199
edu-r-counter-2	education	rules	counter_algorithm	0.796159	3-multiclass	Decomposition	5	0.631252	0.705601
edu-r-idf-0	education	rules	idf_best	0.849578	3-multiclass	None	6	0.667354	0.709943
edu-r-idf-1	education	rules	idf_best	0.875466	3-multiclass	Spacy tokenization	5	0.687306	0.666769
edu-r-idf-2	education	rules	idf_best	0.821499	3-multiclass	Decomposition	6	0.650861	0.727680

Table 9: Rule-based classifiers Results.

ID	Domain	Technique	Method	Accuracy	Model	Structure	F1	Recall
med-n-0-0	medical	nli	0_shot	0.948466	roberta-large	3-class	0.405432	0.747089
med-n-0-1	medical	nli	0_shot	0.963248	bge-m3	3-class	0.413264	0.736307
med-n-0-2	medical	nli	0_shot	0.980167	mDeBerta	3-class	0.476352	0.593545
med-n-0-3	medical	nli	0_shot	0.522452	MiniLM	3-class	0.123156	0.625190
med-n-0-4	medical	nli	0_shot	0.253989	roberta-large	nested-class	0.054539	0.809255
med-n-0-5	medical	nli	0_shot	0.045716	bge-m3	nested-class	0.041871	0.565684
med-n-0-6	medical	nli	0_shot	0.149934	mDeBerta	nested-class	0.053806	0.777962
med-n-0-7	medical	nli	0_shot	0.698137	MiniLM	nested-class	0.142762	0.739025
adm-n-0-0	administrative	nli	0_shot	0.896647	roberta-large	2-class	0.767278	0.826847
adm-n-0-1	administrative	nli	0_shot	0.909281	bge-m3	2-class	0.772669	0.797062
adm-n-0-2	administrative	nli	0_shot	0.800449	mDeBerta	2-class	0.674289	0.831855
adm-n-0-3	administrative	nli	0_shot	0.635634	MiniLM	2-class	0.527614	0.695312
edu-n-0-0	education	nli	0_shot	0.833607	roberta-large	3-multiclass	0.473872	0.416544
edu-n-0-1	education	nli	0_shot	0.831991	bge-m3	3-multiclass	0.497835	0.537146
edu-n-0-2	education	nli	0_shot	0.827772	mDeBerta	3-multiclass	0.447585	0.419011
edu-n-0-3	education	nli	0_shot	0.711922	MiniLM	3-multiclass	0.241515	0.366130

Table 10: Natural Language Inference Results.

ID	Domain	Technique	Method	Accuracy	Model	Structure	Examples	Distance	F1	Recall
med-e-similarity-0	medical	embedding	similarity	0.984168	multilingual-e5	3-class	0_shot	0.100000	0.348230	0.304990
med-e-similarity-1	medical	embedding	similarity	0.921418	multilingual-e5	3-class	1_shot	0.100000	0.105151	0.408667
med-e-similarity-2	medical	embedding	similarity	0.978404	multilingual-e5	3-class	few_shot	0.050000	0.220443	0.407954
adm-e-similarity-0	administrative	embedding	similarity	0.770938	multilingual-e5	2-class	0_shot	0.150000	0.639797	0.804084
adm-e-similarity-1	administrative	embedding	similarity	0.771179	multilingual-e5	2-class	1_shot	0.100000	0.543606	0.574040
adm-e-similarity-2	administrative	embedding	similarity	0.919950	multilingual-e5	2-class	few_shot	0.050000	0.654192	0.607796
edu-e-similarity-0	education	embedding	similarity	0.871706	multilingual-e5	3-multiclass	0_shot	0.150000	0.505434	0.614236
edu-e-similarity-1	education	embedding	similarity	0.752082	multilingual-e5	3-multiclass	1_shot	0.100000	0.248134	0.434147
edu-e-similarity-2	education	embedding	similarity	0.575447	multilingual-e5	3-multiclass	few_shot	0.100000	0.262027	0.679848
med-e-similarity-3	medical	embedding	similarity	0.971403	qwen	3-class	0_shot	0.500000	0.149797	0.530709
med-e-similarity-4	medical	embedding	similarity	0.954201	qwen	3-class	1_shot	0.500000	0.185169	0.328241
med-e-similarity-5	medical	embedding	similarity	0.977911	qwen	3-class	few_shot	0.300000	0.256171	0.509904
adm-e-similarity-3	administrative	embedding	similarity	0.794364	qwen	2-class	0_shot	0.700000	0.644378	0.766306
adm-e-similarity-4	administrative	embedding	similarity	0.891397	qwen	2-class	1_shot	0.500000	0.486496	0.502398
adm-e-similarity-5	administrative	embedding	similarity	0.925626	qwen	2-class	few_shot	0.300000	0.741955	0.697743
edu-e-similarity-3	education	embedding	similarity	0.870782	qwen	3-multiclass	0_shot	0.700000	0.485113	0.581780
edu-e-similarity-4	education	embedding	similarity	0.869994	qwen	3-multiclass	1_shot	0.500000	0.262263	0.315616
edu-e-similarity-5	education	embedding	similarity	0.906132	qwen	3-multiclass	few_shot	0.300000	0.340588	0.228346
med-e-similarity-6	medical	embedding	similarity	0.980090	mistral	3-class	0_shot	0.200000	0.194220	0.241478
med-e-similarity-7	medical	embedding	similarity	0.975926	mistral	3-class	1_shot	0.200000	0.136733	0.113083
med-e-similarity-8	medical	embedding	similarity	0.975729	mistral	3-class	few_shot	0.150000	0.253893	0.486291
adm-e-similarity-6	administrative	embedding	similarity	0.919572	mistral	2-class	0_shot	0.300000	0.737436	0.705013
adm-e-similarity-7	administrative	embedding	similarity	0.772348	mistral	2-class	1_shot	0.300000	0.609657	0.713248
adm-e-similarity-8	administrative	embedding	similarity	0.935417	mistral	2-class	few_shot	0.150000	0.793699	0.758484
edu-e-similarity-6	education	embedding	similarity	0.908266	mistral	3-multiclass	0_shot	0.300000	0.451101	0.350068
edu-e-similarity-7	education	embedding	similarity	0.611930	mistral	3-multiclass	1_shot	0.300000	0.316307	0.565826
edu-e-similarity-8	education	embedding	similarity	0.915368	mistral	3-multiclass	few_shot	0.150000	0.491348	0.402072
med-e-similarity-9	medical	embedding	similarity	0.981571	e5-small	3-class	0_shot	0.100000	0.173329	0.117828
med-e-similarity-10	medical	embedding	similarity	0.953576	e5-small	3-class	1_shot	0.100000	0.050910	0.084934
med-e-similarity-11	medical	embedding	similarity	0.816069	e5-small	3-class	few_shot	0.100000	0.073753	0.522673
adm-e-similarity-9	administrative	embedding	similarity	0.900600	e5-small	2-class	0_shot	0.100000	0.474728	0.500223
adm-e-similarity-10	administrative	embedding	similarity	0.898979	e5-small	2-class	1_shot	0.100000	0.496698	0.510050
adm-e-similarity-11	administrative	embedding	similarity	0.911513	e5-small	2-class	few_shot	0.050000	0.574878	0.554500
edu-e-similarity-9	education	embedding	similarity	0.671910	e5-small	3-multiclass	0_shot	0.150000	0.213756	0.409485
edu-e-similarity-10	education	embedding	similarity	0.524017	e5-small	3-multiclass	1_shot	0.200000	0.189732	0.533857
edu-e-similarity-11	education	embedding	similarity	0.782214	e5-small	3-multiclass	few_shot	0.100000	0.291436	0.407158
med-e-classifier-0	medical	embedding	classifier	0.979204	multilingual-e5	nested-class	NaN	NaN	0.535724	0.674330
med-e-classifier-1	medical	embedding	classifier	0.987224	multilingual-e5	nested-class	NaN	NaN	0.583769	0.618593

Table 11: Embedding Cosine Similarity Results.

ID	Domain	Technique	Method	Model	Classifier	Structure	C	solver	penalty	F1	Recall
med-e-classifier-0	medical	embedding	classifier	multilingual-e5	logreg	nested-class	10	liblinear	l1	0.562499	0.723484
adm-e-classifier-0	administrative	embedding	classifier	multilingual-e5	logreg	2-class	10	liblinear	l1	0.862211	0.941253
edu-e-classifier-0	education	embedding	classifier	multilingual-e5	logreg	3-multiclass	10	liblinear	l1	0.714403	0.912029
med-e-classifier-2	medical	embedding	classifier	qwen	logreg	nested-class	10	liblinear	l1	0.629873	0.783560
adm-e-classifier-2	administrative	embedding	classifier	qwen	logreg	2-class	10	liblinear	l1	0.868487	0.925343
edu-e-classifier-2	education	embedding	classifier	qwen	logreg	3-multiclass	10	liblinear	l1	0.748070	0.891451
med-e-classifier-4	medical	embedding	classifier	mistral	logreg	nested-class	10	liblinear	l1	0.643318	0.826420
adm-e-classifier-4	administrative	embedding	classifier	mistral	logreg	2-class	10	liblinear	l1	0.883531	0.942878
med-e-classifier-6	medical	embedding	classifier	e5-small	logreg	nested-class	1	lbfgs	l2	0.270243	0.734472
adm-e-classifier-6	administrative	embedding	classifier	e5-small	logreg	2-class	10	liblinear	l1	0.758679	0.874330
edu-e-classifier-4	education	embedding	classifier	mistral	logreg	3-multiclass	10	liblinear	l1	0.775537	0.918018
edu-e-classifier-6	education	embedding	classifier	e5-small	logreg	3-multiclass	10	liblinear	l1	0.537592	0.821275
ID	Domain	Technique	Method	Model	Classifier	Structure	C	Kernel	Gamma	F1	Recall
med-e-classifier-1	medical	embedding	classifier	multilingual-e5	svm	nested-class	10	rbf	scale	0.624416	0.665972
adm-e-classifier-1	administrative	embedding	classifier	multilingual-e5	svm	2-class	10	rbf	scale	0.885668	0.952657
edu-e-classifier-1	education	embedding	classifier	multilingual-e5	svm	3-multiclass	10	rbf	scale	0.771529	0.936686
med-e-classifier-3	medical	embedding	classifier	qwen	svm	nested-class	1	rbf	scale	0.666347	0.765587
adm-e-classifier-3	administrative	embedding	classifier	qwen	svm	2-class	1	rbf	scale	0.892814	0.943593
med-e-classifier-5	medical	embedding	classifier	mistral	svm	nested-class	10	rbf	scale	0.682510	0.672002
med-e-classifier-7	medical	embedding	classifier	e5-small	svm	nested-class	10	rbf	scale	0.537543	0.620456
adm-e-classifier-7	administrative	embedding	classifier	e5-small	svm	2-class	10	rbf	scale	0.862910	0.918194

Table 12: Static-Embedding Classifier Results.

	Germany	France
med-r-expert-0		Hôpital, Hospitalier, CH, CHU, Médical, GHT, GHU, Clinique
med-r-llm-0	Krankenhaus, Medizinische Einrichtung, Gesundheitszentrum, Klinikum, Heilanstalt	Hospital, Hôpital, Services Hospitaliers, Unités Hospitalières, Maternités
med-r-counter-0	krankenhaus, ehemaliges, klinikum, spital, hospital, klinik, nervenheilanstalt	hôpital, hospitalier, dieu, hospice, local, clinique, ancien
med-r-counter-1	krankenhaus, klinikum, klinik, spital, hospital, nervenheilanstalt	hôpital, hospitalier, dieu, hospice, clinique, local
med-r-counter-2	krankenhaus, klinik, klinikum	hôpital, hospitalier, hospice
med-r-counter-3	krankenhaus, ehemaliges, klinikum, spital, hospital, klinik, nervenheilanstalt	hôpital, hospitalier, dieu, hospice, local, clinique, ancien
med-r-counter-4	krankenhaus, klinikum, klinik, spital, hospital, nervenheilanstalt	hôpital, hospitalier, dieu, hospice, clinique, local
med-r-counter-5	krankenhaus, klinik, klinikum, spital, sachgesamtheit, hospital, nervenheilanstalt	hôpital, hospitalier, hospice, dieu, clinique, local, ancien
med-r-idf-0	ehemaliges, krankenhaus, spital	dieu, hospitalier, hôpital
med-r-idf-1	klinikum, krankenhaus, spital	dieu, hospitalier, hôpital
med-r-idf-2	klinik, krankenhaus, spital	hospice, hospitalier, hôpital
med-r-idf-3	ehemaliges, krankenhaus, spital	dieu, hospitalier, hôpital
med-r-idf-4	klinikum, krankenhaus, spital	dieu, hospitalier, hôpital
med-r-idf-5	klinik, krankenhaus, spital	hospice, hospitalier, hôpital
adm-r-llm-0	Gemeinde, Stadtverband, Kommune, Verwaltungseinheit, Raad	Mairie, Commune, Conseil Municipal, EPCI, Région
adm-r-counter-0	verwaltungsgemeinschaft, landkreis, amt	aubin, mairie, déléguée
adm-r-counter-1	verwaltungsgemeinschaft, landkreis, amt	aubin, les, mairie
adm-r-counter-2	verwaltungsgemeinschaft, landkreis, amt	aubin, ign, mon
adm-r-idf-0	amt, landkreis, verwaltungsgemeinschaft	centre, commune, sociale
adm-r-idf-1	amt, landkreis, verwaltungsgemeinschaft	centre, commune, social
adm-r-idf-2	amt, landkreis, verwaltungsgemeinschaft	centre, commune, social
edu-r-llm-0	Volksschule, Grundschule, Unterstufe, Lehrstrafene, Basiskenntnisse	École Primaire, CP, Cours Primaire, Maternelle, Enfants Scolaires
edu-r-counter-0	grundschule, volksschule, hauptschule, grimmelshausenschule, renchen	ecole, elementaire, primaire, élémentaire, specialisee
edu-r-counter-1	grundschule, volksschule, hauptschule, katholisch, grimmelshausenschule, renchen, grimm, don, bosco, montessori	ecole, elementair, primaire, élémentaire, specialisee, bert, élementaire, hirsch, lucie, aubrac
edu-r-counter-2	grundschule, grimm, volksschule, hauptschule, weg	ele, primaire, mentir, élémentaire, eco
edu-r-idf-0	grimmelshausenschule, grundschule, hauptschule, renchen, schule, volksschule	ecole, elementaire, mazaire, primaire, école, élémentaire
edu-r-idf-1	grundschule, hauptschule, katholisch, schule, volksschule	ecole, elementair, primaire, école, élémentaire
edu-r-idf-2	grimm, grundschule, hauptschule, schule, volksschule, weg	eco, ele, mentir, primaire, école, élémentaire

Table 13: France and Germany's keywords for selected classes (hospital, local government and primary school). Complete results available in [30].

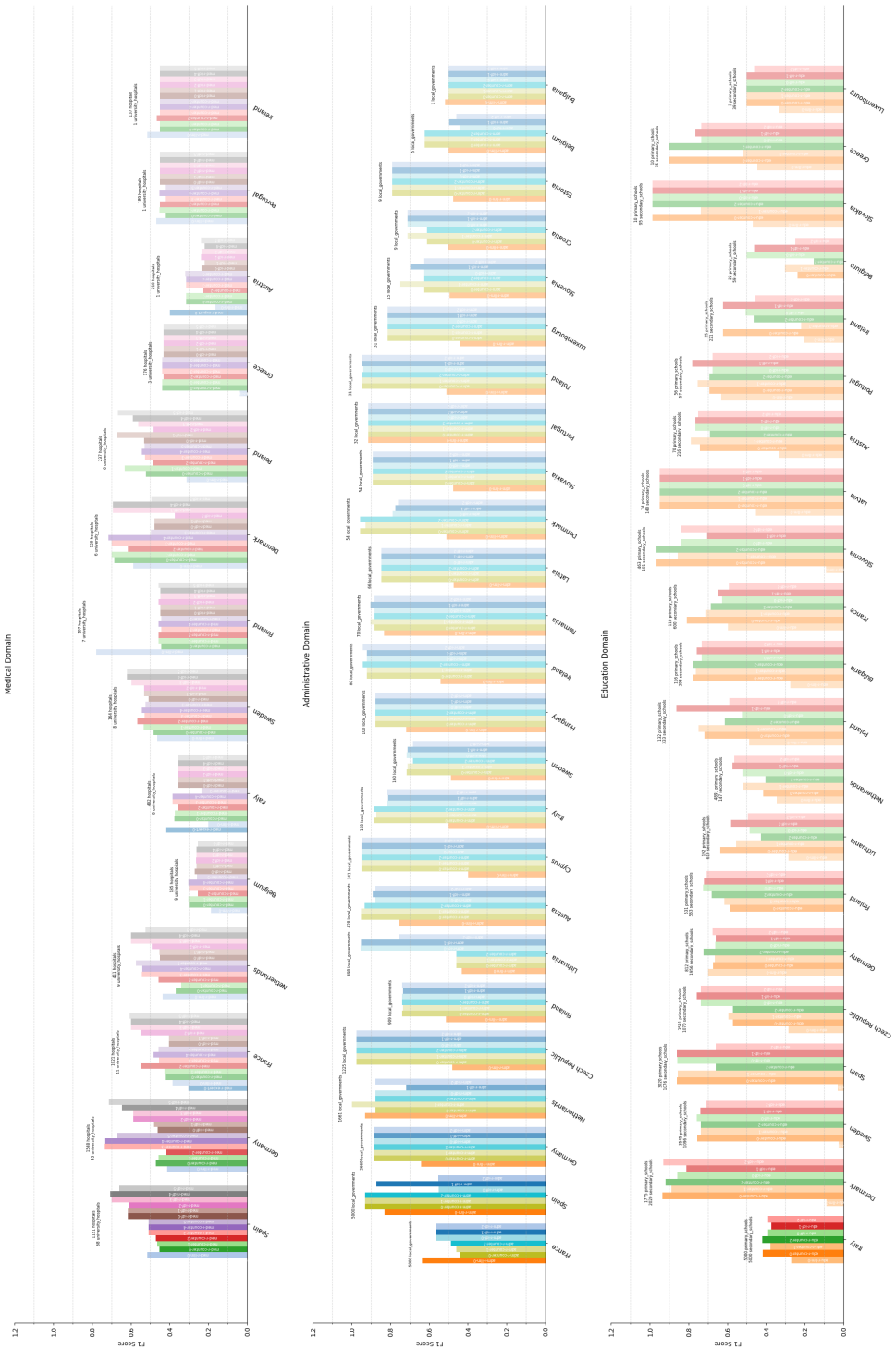
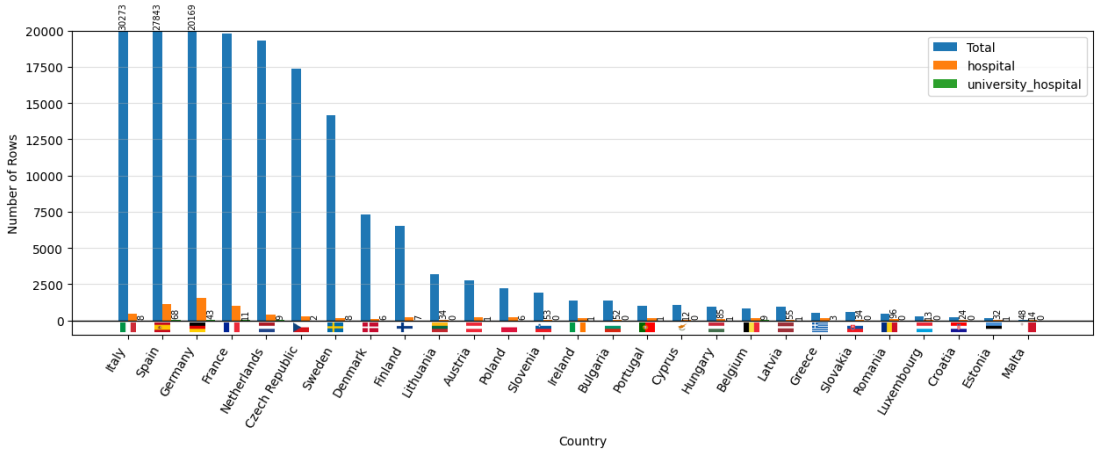
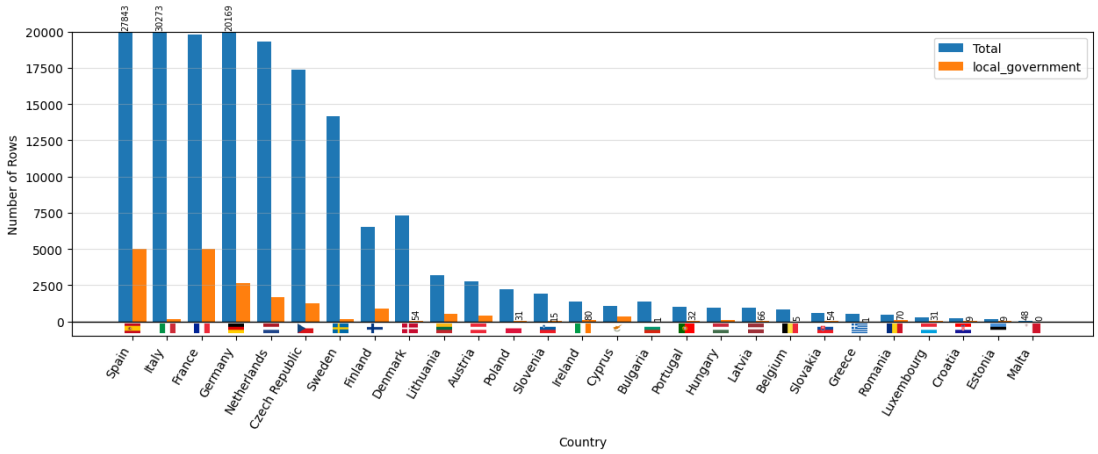


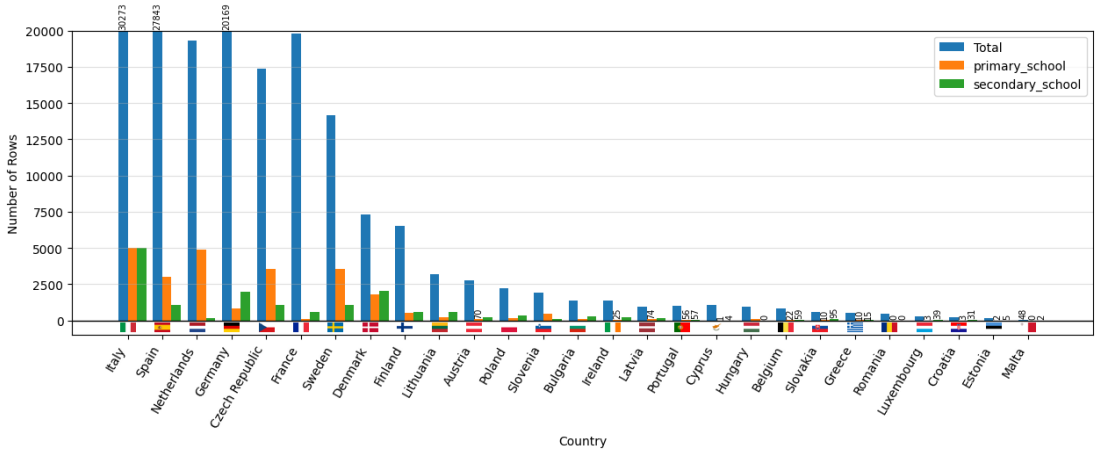
Fig. 3. F1 Scores of Rule-based experiments per Country. Opacity encodes data availability in the smallest class.



(a) Medical domain



(b) Administrative domain



(c) Educational domain

Fig. 2. Volume of entities per domain. Each subfigure shows the distribution of entities in the Medical, Administrative, and Educational classes respectively.

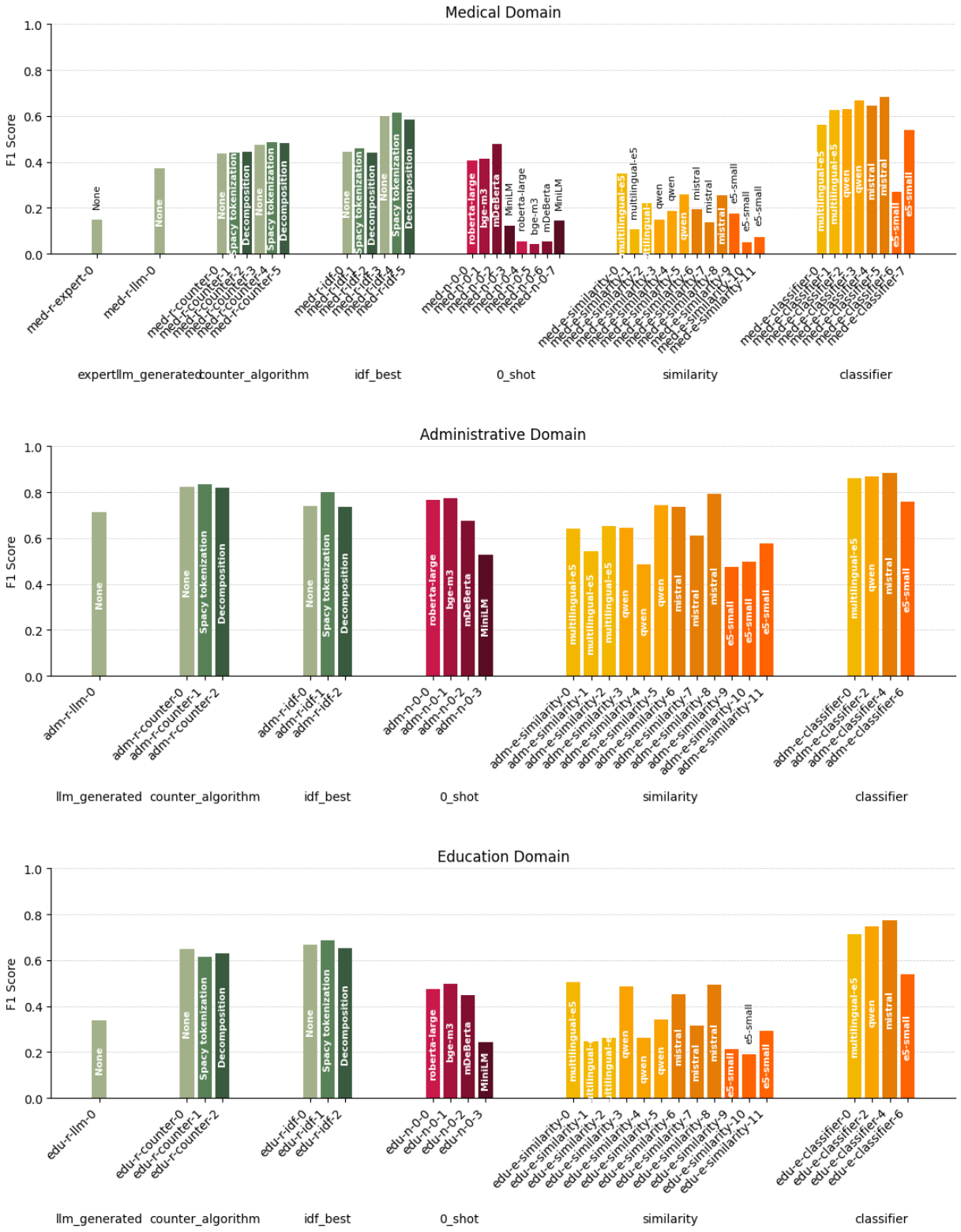


Fig. 4. Experiments F1 scores.

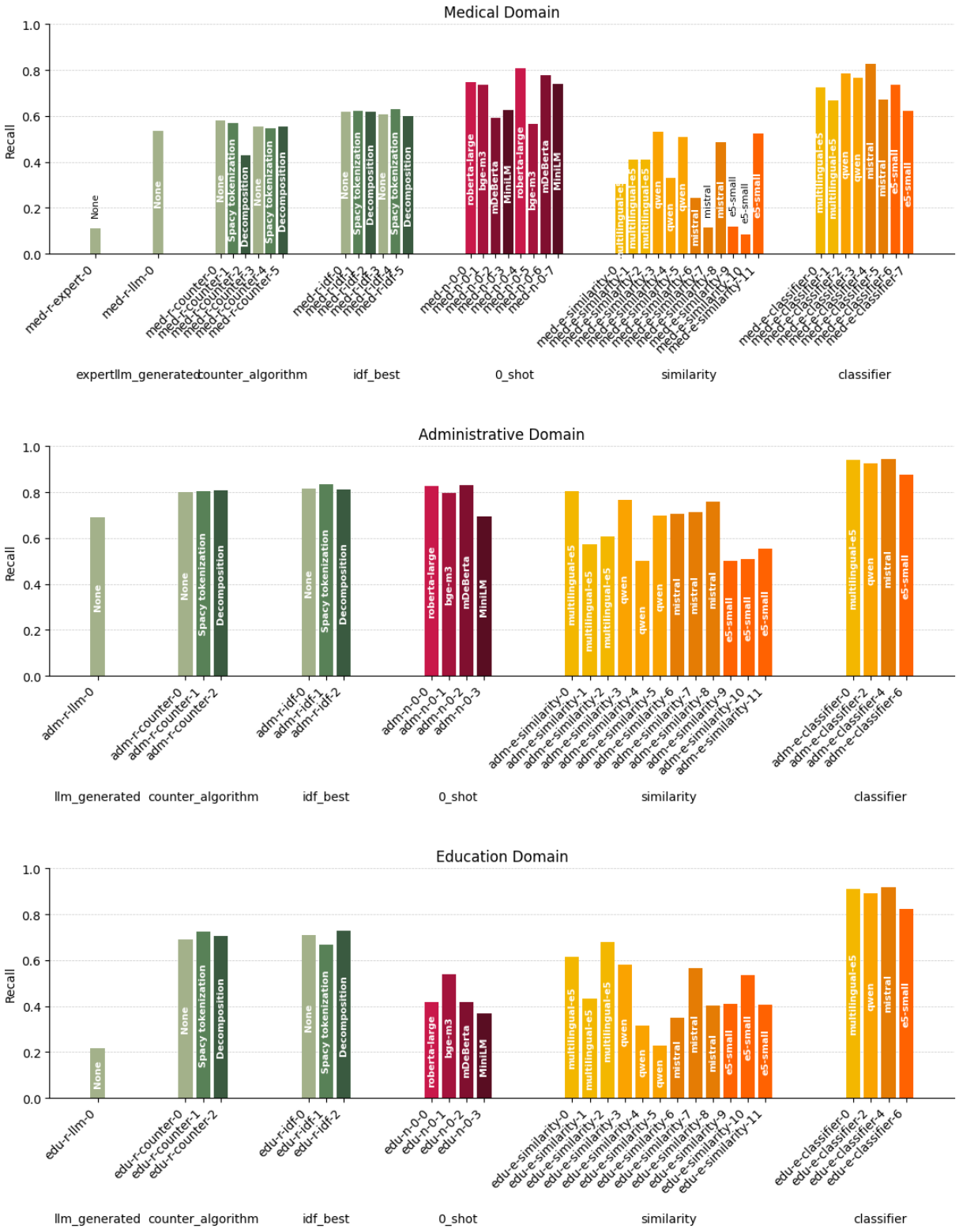


Fig. 5. Experiments Recall Scores.

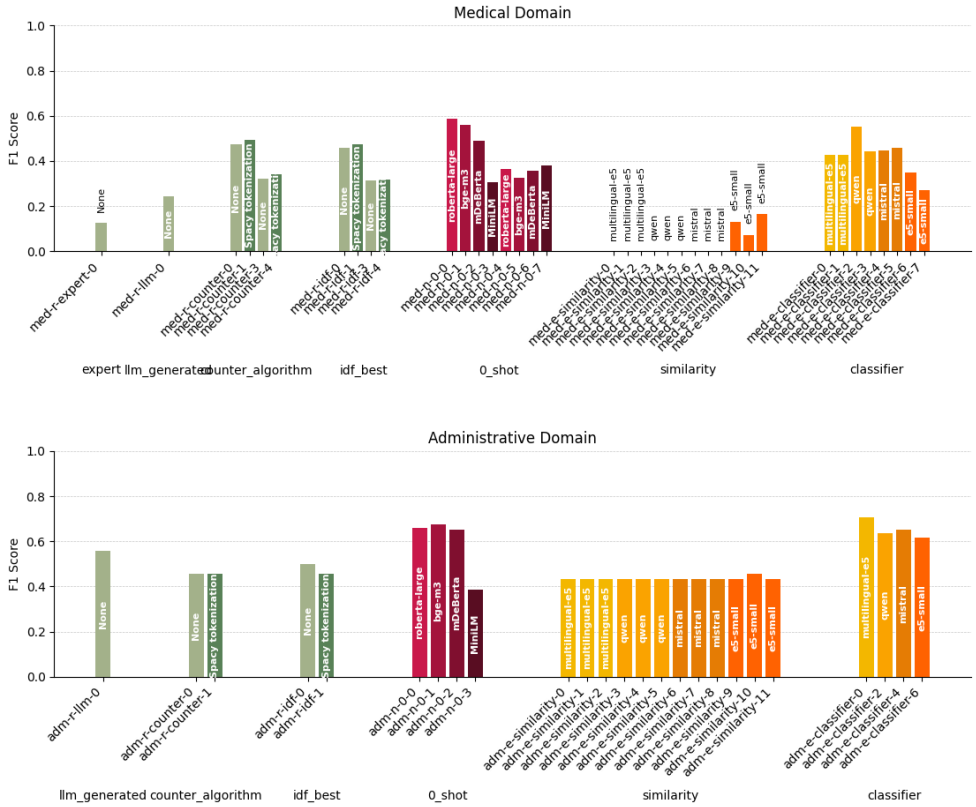


Fig. 6. F1 scores for EU Contract Hub experiments.