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 all European Union 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 some training data available. 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. This work highlights the feasibility of ontology-guided model training from short texts and which can prove valid approaches 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.

1570-0844/$35.00 © 0 – IOS Press. All rights reserved.

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 when name 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. In particular, within the broader task of Entity Linking [55], one of the subtasks is Entity Type Classification which serves the important function of reducing the search space downstream. Similarly, in Named Entity Recognition, Entity Classification is one of the key steps with modern benchmarks such as FewNERD [15] highlighting the increasing necessity for classification in complex scenarios. While general-purpose named entity recognition methods are designed to identify and broadly type entities within rich textual contexts; most are ill-suited to the problem of fine-grained classification from a name string in isolation, without semantic context. 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.

The task of organization classification itself serves several critical functions across the public and private sectors. 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.

In large-scale administrative datasets, such as public procurement registries, organizational entries are frequently recorded with no accompanying metadata beyond a name string. Knowledge extraction must therefore operate on the name alone. Whereas in the domains of finance, economics, and business operations, organization classification enables investment and risk management through peer-group comparisons, it also supports performance benchmarking against similar entities. Finally, it allows for a variety of analysis such as competitor, market-share and sales studies.

The motivation for these methods is further reinforced by recent comparative analyses of NLP-based models, ranging from zero-shot learning to deep learning classifiers, which demonstrate the feasibility of automating organization classification across diverse industry tasks using existing semantic datasets. By shifting toward ontology-grounded classes, these systems can accommodate flexible and evolving classification targets within extensive hierarchies. This provides a structured alternative to the generative "Generative Information Extractors" [78] popularized through LLM-based methods. While generative models are undoubtedly competitive, grounding these extractors in verified knowledge bases like Wikidata ensures a less resource-intensive, less hallucination-prone approach, and by virtue of open access to the training corpus, more interpretable as well.

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 developed for this study, named EU Contract Hub [84], 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 [58] [18].

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, Dutch) pose particular challenges for tokenization, rule-based methods, and semantic segmentation. **Second**, 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. **Third**, the inherent brevity and semantic ambiguity of proper names significantly hinder entity disambiguation accuracy and most methods in the literature have not been tested under edge conditions with no further input. **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. In this paper, the term low-resource will refer to the characteristics of the execution environment: restricted in volatile memory, hard-drive volume and consumer-grade hardware, where all executions, datasets and models will be deployed.

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.

The primary challenge in classifying public sector organizations stems from the extreme brevity of a single name string which must be accurately mapped to complex, sometimes hierarchical classes, across 24 countries and 29 languages. To address these varying degrees of linguistic and structural complexity, our methodology is organized into three distinct families that progressively move from deterministic heuristics to latent semantic representations. In administrative settings, classification decisions must often be traceable for transparency or regulatory purposes. This requirement leads to our first family: (i) Rule-based systems. By utilizing regular expression (regex) matching against domain-specific keywords, we can ensure that the model's logic remains immediately interpretable by human experts. Rule-based systems serve as an interpretable baseline, though they require labelled data. To handle settings without the need for laborious manual annotation, we investigate (ii) Zero-shot classification via Natural Language Inference (NLI). By reformulating classification as an entailment problem we leverage the cross-lingual knowledge of pre-trained transformers to evaluate task performance in the complete absence of labeled training entities, enabling the categorization of inputs into classes the model has never explicitly encountered. Finally, to determine if the class of an organization is captured beyond lexical features, we must investigate the underlying semantic structure of the data. This motivates the use of (iii) Embedding-based methods. By converting organization names into dense, high-dimensional vector representations using models like RoBERTa or E5, we can assess whether these systems naturally capture the hierarchy defined in Wikidata's ontology.

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 trained on knowledge graph data distinguish between medical, government, and educational multilingual organizations by name alone?
- **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?

2. Background

The task of classifying organizations, whether they be companies, non-profit entities, or public service institutions such as hospitals and schools, has evolved into a core task of modern economic analysis and academic research. Traditionally, the categorization of entities relied on manual labor by domain experts using established, rigid standard taxonomies.

One common example is industry classification, which involves grouping into categories based on shared characteristics, such as business activities, market performance and reports. Standardized industry classification systems include the Global Industry Classification Standard (GICS) [48], the Standard Industrial Classification (SIC) or the North American Industry Classification System (NAICS) [26]. While GICS and NAICS dominate the private corporate sector, the classification of public organizations, civic entities, and non-profits relies instead on the National Taxonomy of Exempt Entities (NTEE)[50] with industry groups, among others, education and hospitals. In the European Union the categorization of public entities, the processing of procurement contracts, and the delineation of economic sectors relied on similar frameworks, most notably the Statistical Classification of Economic Activities in the European Community (NACE) [25]. The European Union fundamentally relies on highly standardized, hierarchical vocabularies to establish semantic interoperability across its diverse member states. Another key taxonomy, this time standardizing the object of procurement contracts is the Common Procurement Vocabulary (CPV) [19]. Public organizations have specific characteristics that enable alternative formulations. On one hand, they tend to contain more descriptive information on their name. While private corporations might include designations such as 'Ltd.', we can expect hospital names to contain 'hospital', schools to be named after a historical figure or local governments to include the name of the region. While this might seem trivial, it does complicate with the introduction of diverse regions. Public organizations can in many domains be identified with a physical building. While ontologically, organization and building are not the same entity, in terms of data, they share many properties. Some approaches link OpenStreetMap (https://www.openstreetmap.org) data to organizational records [60], to knowledge graphs [13, 62, 63, 82], and specifically to Wikidata [40].

In recent years, public institutions and academic researchers have initiated a shift toward Natural Language Processing (NLP), and the deployment of transformer-based Large Language Models (LLMs). These technologies map administrative data to standardized EU taxonomies. In e-procurement research, the automation of CPV code assignment is a classification task commonly approached through two distinct paradigms: discriminative text classification and generative modeling tasks [57]. However, due to class imbalance at more granular levels, training algorithms suffer from a lack of examples. That issue is approached by using pre-trained language models that rely on the semantic signal of the label [47]. While the CPV classifies the subject matter of contracts, NACE Revision 2 organizes economic activities into a four-level hierarchy. The challenge in modeling the NACE taxonomy itself via NLP is dimensionality reduction without collapsing the hierarchical topology of the economic framework, which has been attempted in [65] through iterative processing of descriptions. Simultaneously, some approaches rely on Support Vector Machines (SVM) to predict NACE codes utilizing text scraped from web pages [36]. NACE classification

of organizations has been utilized for diverse downstream tasks such as enhancing retrieval, grounding answers and further classification[3, 79]. To address structured label spaces, Hierarchical Text Classification methods are supervised learning algorithms explicitly designed to take advantage of the labels' taxonomic organization [73, 77]. Recent advancements in the field have introduced Zero-Shot Hierarchical Text Classification [76], where the semantic meaning of the category names themselves is embedded into dense vector prototypes. The inputs for these NLP classification models are vastly heterogeneous: from business descriptions to financial filings. Transformer-based language models, such as BERT and RoBERTa, are tasked with reading these dense financial documents and generating high-dimensional contextual embeddings. [53]. Beyond direct classification, unsupervised topic modeling can cluster startups based on the latent themes in their textual descriptions. This approach has proven vital in tracking macroeconomic trends, and could provide better distinguishable and more concrete classes than manual categorization. [54]. This latter possibility was studied for the case of CPV codes in [58].

The literature on automated classification is skewed toward private corporations (SIC, GICS, NAICS contexts). Public institutions, hospitals, schools, local governments, have received far less attention, despite having distinct naming conventions. Similarly they leverage rich textual or structured metadata. Classifying organizations solely from their name is underexplored and harder than it looks once linguistic and regional diversity enters the picture. However, many of the above mentioned methods have been adapted for our task, comparing among others, pre-trained models to understand labels and Support Vector Machines.

2.1. Entity Classification

The specific NLP task can be defined as follows: Let $\mathcal{X}$ denote the space of organization name strings and let $\mathcal{Y} = \{y_1, \ldots, y_K\}$ be a finite set of K organizational categories. The task consists of learning a function $f : \mathcal{X} \to \{0,1\}^K$ that maps an organization name $x \in \mathcal{X}$ to a binary label vector $\mathbf{y} \in \{0,1\}^K$, where $y_k = 1$ indicates membership in category k. The active label set of a sample is $\mathcal{Y}(x) = \{y_k \in \mathcal{Y} \mid y_k = 1\}$; samples for which $\mathcal{Y}(x) = \emptyset$ constitute a residual *other organizations* class. A taxonomy $\mathcal{T} = (\mathcal{Y}, E)$ is a directed acyclic graph whose whose edges E encode hyponymy, moving from general to increasingly specific concepts. Formally, $\forall x \in \mathcal{X}$ with associated ground-truth label vector $\mathbf{y} \in \{0,1\}^K$:

$$(y_k = 1) \wedge (k \preceq j \text{ in } \mathcal{T}) \implies y_j = 1 \tag{1}$$

The application of Natural Language Processing to organizational classification transforms the problem from a manual task into a text mining or representation learning challenge. The computational task, entity classification, requires mapping an input text $\mathcal{N}$ (in our case, the name of the organization) to the predefined set of labels $\mathcal{Y}$.

2.2. Related Tasks

Entity classification is commonly integrated within more complex NLP tasks as one of their step. Named Entity Recognition (NER), Entity Linking (EL), also referred to as Named Entity Disambiguation (NED) and Entity Normalization. NER is formulated as a sequence-labeling and span-extraction task. Its primary objective is to locate the start and end boundaries of an entity mention within a continuous string of unstructured text. EL goes a step further: once an entity span is recognized it must map that textual span to a unique, canonical identifier within a target Knowledge Base (KB), such as a Wikidata QID or a DBpedia URI. Entity Classification, operates within the boundaries of both, categorizing the identified span into a predefined taxonomic class. In modern natural language processing contexts, this classification extends far beyond the traditional, coarse-grained categories of the past (such as Person, Organization, or Location) into fine-grained typologies. This subtask can aid restricting the Knowledge Graph candidates for EL. Joint-learning models are designed to optimize a single, shared representation space where the loss functions of both Named Entity Recognition and Entity Linking are calculated and minimized simultaneously during the backpropagation phase avoiding compounding errors [46]. Contrastive Learning and regular expressions [34] have been used to guide these learning objectives. Contrastive Learning in particular attempts to deal with the issue of anisotropy making separation difficult within embedding spaces.

While the European Commission has aggressively adopted Linked Data principles to increase the accessibility and interoperability of Open Government Data, the metadata accompanying these massive datasets remains notoriously fragmented and heterogeneous. Resolving metadata to unified Knowledge Graphs presents a multi-faceted NLP challenge including the challenges of Temporal Dynamics of Administration, Absence of Evolution-Aware Ontologies, Metadata Quality Deficits, Linguistic Heterogeneity and Organizational Disambiguation [52]. 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 [80]. 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 [53, 81]. Integrating complex KGs into embedding-based classifiers augments how NLP models process linguistic data. Rather than relying solely on the statistical distribution of words in a text corpus, models employ advanced Knowledge Graph Embeddings (KGE). Methods such as TransE RotatE ComlEx and derivatives model relations in the KG as mathematical operations such as vector displacements, rotations or products between entity embeddings. RDF2Vec instead generates entity representations through random walks over the graph. In the context of entity classification, these embeddings offer an alternative route to encoding class semantics. However a key issue is that of aligning embedding spaces of instances and labels if they use different encoders. In our study we will keep to name-only as input for our methods which will stress the capabilities of many of the current best performing paradigms in similar tasks and shed some light in alternative approaches when we are under low-context conditions. However, the Contrastive Learning objective is ideal for this restriction.

2.3. Datasets and Benchmarks

As the scope of Information Extraction broadens across sectors, the demand for granular entity classification increases. Real-world applications rarely require simple extraction of a generic Person or Location; but identification of a Politician versus an Athlete, or Hospital versus a School. To benchmark the capability of methods to navigate complex taxonomic spaces, the Few-NERD dataset has emerged as the premier global evaluation standard [15]. Derived from the text of Wikipedia, it is a two-level hierarchical ontology. with 8 broad, coarse-grained categories (such as Organization, Location, Person, and Product) subsequently subdivided into 66 subcategories (such as Education or Government organizations). Methods such as GLiNER or NuNER [4, 59, 74, 75] treat labels as text, not as rigid categories. Other approaches to this benchmark share encoder-based architectures, enriched label and taxonomy descriptions that transfer to unseen types, and contrastive learning objectives [31, 35, 71]. While the Few-NERD dataset primarily tests taxonomic granularity, MultiCoNER V2 (Multilingual Complex Named Entity Recognition) dataset and its associated SemEval-2023 [27] shared task evaluate deep learning models against the chaotic, unstructured realities of global, real-world data. Top-performing systems is PAI [45], which integrates external entity information to improve performance and [33] which studies custom LLM heads and losses for training the final layer. MultiNERD [61] and DynamicNER [44] propose multi-granular, multi-language NER datasets. More specialized is AdminSet-NER [56], focuses on the administrative domain and is composed of unstructured French public administration documents. Moving on to Entity Linking we can find datasets that bridge text to Wikidata entities such as Mewsli-9 and T-REx [5, 16]. For multilingual entity linking, mGENRE [7] reformulates the task as autoregressive sequence generation over entity names, mReFinED [42] jointly trains mention detection, entity typing, and entity disambiguation in a single multilingual model and ReLIK [51] a Retriever-Reader architecture that first fetches candidate entities from a knowledge base index, and a DeBERTa-v3 Reader then contextualizes input text to predict which spans link to which entities.

Finally,

Span-based sequence labeling approaches used in NER are not applicable in our low-context, name-only setting and we are yet to identify datasets that approach these more context-restrictive environments. We propose that Named Entity Recognition benchmarks could be reasonably extended to structured, JSON-based evaluation datasets. Whereas literature surveyed above reveals that entity classification increasingly relies on contextual enrichment, whether through knowledge-graph triples, retrieved Wikipedia passages, or multi-sentence descriptions. This enrichment is effective precisely because it compensates for the semantic signal absence of short forms. Our

setting, however, imposes a deliberate and practically motivated constraint: only the organization name is available at inference time. Public procurement records, administrative registries, and similar noisy datasets routinely suffer from data quality issues and lack accompanying descriptions. Within this constraint we situate three complementary strategies. **Rule-based heuristics** require no training data and exploit the observation that public-sector names carry lexical signals: institutional designators, toponym patterns, and domain keywords. Their inclusion is justified not only as a baseline but as a common standard in NER tasks. Classification via **Natural Language Inference** exploits the finding [39] that NLI models trained on MultiNLI-derived corpora generalize to arbitrary label spaces when candidate classes are cast as hypotheses. DeBERTa-v3-based NLI models have emerged as the dominant backbone across reviewed systems in MultiCoNER, Few-NERD, and CPV classification alike and their cross-lingual variants make them a natural fit for our scope. Crucially, this approach requires no labeled examples from our target taxonomy and may be adapted to any taxonomy, relying instead on the semantic signal carried by class labels and model pre-training, consistent with the zero-shot paradigm explored for CPV assignment [47]. **Generative few-shot classification** complements the discriminative NLI approach, completing the spectrum from zero labeled data to minimal supervision. Finally, **embedding-based methods** address the the absence of a representation space limitation shared by the above approaches. We evaluate embedding representations using progressively more flexible classifiers. First, we assess the geometry of *frozen pretrained encoders* by measuring cosine distances to class examples in the induced embedding space, analogously to the prototype-space analyses conducted for NER datasets. Second, we attach a *linear classification head* directly to the encoder, the standard discriminative baseline adopted across the reviewed systems, from DeBERTa-based token classifiers in MultiCoNER [69] to the Softmax output layers in BERT-KG [81], providing a strong supervised upper bound under full label availability. Several reviewed systems note the anisotropy of pretrained transformer embeddings as an obstacle to classification in dense label spaces; contrastive objectives have been proposed as the remedy both in NER [34] and in Few-NERD [35]. We adopt Supervised Contrastive loss fine-tuning.

We deliberately exclude approaches that require information beyond the name. Methods that retrieve Wikipedia passages [69], inject KG triples at inference [81], or read financial filings [53] are reviewed because they establish the state of the art on the richer inputs that are unavailable in our setting, and because understanding where context helps is essential for interpreting where our name-only methods fall short. Similarly, joint NER–EL architectures and hierarchical text classifiers trained on large labeled corpora are surveyed to frame the broader task, not as direct comparators. Despite their theoretical appeal, we do not pursue KGE-based class representations in this work. The fundamental barrier is embedding space alignment: our organization name encoder produces representations in a name-based semantic space, while KGE methods produce representations in a structurally-based relational space. The standard remedy is to learn linear projection that maps one space onto the other using a set of anchor entities whose identity is known in both spaces [80]. However, a well-chosen anchor entity is itself a canonical representative of its class, which raises the question of whether the projection adds information beyond simply using that anchor as a class prototype directly. Further, before encoding the KG structure of the labels first (which may not add too much information besides examples of that class in our use case), we might look into textual descriptors of the label. Our evaluation, spanning flat and nested label structures and multiple languages, is designed to isolate the contribution of each strategy under this constraint, which can later be foregone to enhance our methods with extra context.

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 generalizable, and semantically grounded datasets for model training. 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.

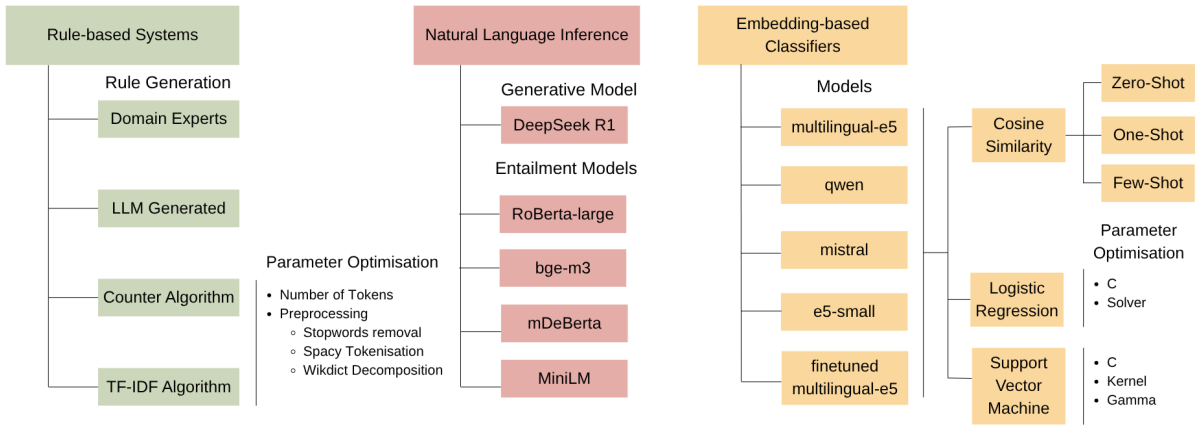


Figure 1. We have evaluated three methodological techniques: rule-based systems, natural language inference and embedding-based classifiers. Each family has several methods, and some test different models and parameter configurations. Parameter configurations are optimized and selected on a validation set.

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 pre-processing pipelines were tested on rule systems to assess their impact on classification outcomes, and 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.

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.

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. It was selected as our primary data source due to: European Union multilingual coverage, explicit ontological typing for a wealth of classes (including subclasses), and open licensing, which enables reproducible large-scale experimentation. 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. At the time of data extraction and experimentation, Wikidata did not provide a language-independent preferred label for organizations. To mitigate this limitation, we implemented a heuristic that defaulted to the majority language of each respective EU Member State. While we acknowledge that this compromise simplifies the linguistic diversity within countries, our underlying framework was designed to be capable of processing minority languages; the primary constraint was the absence of a resolution protocol for determining when to prioritize specific languages as primary. This issue has since been addressed by the introduction of the 'mul' (multilingual) language code, which provides a ground-truth for the preferred language of an organization's name.

To ground the choice of Wikidata as an authoritative source we have conducted a very tentative estimation of its coverage of selected organisational classes across the European Union. For this we have made use of Eurostat metrics on number of hospital beds per inhabitant [20], pupil enrolments in primary education [23], and pupil enrolments in lower and upper secondary education [21, 22]. The number of administrative units at local level per country was taken from the Eurostat Local Administrative Units (LAU) register for 2025 [24]. For schools, the number of educational institutions per country was estimated by dividing total enrolments by a conservative assumption of 400 students per school. For hospitals, institution counts were estimated by dividing total hospital beds by a conservative assumption of 150 beds per hospital. These figures represent order-of-magnitude estimates intended only to contextualise the relative coverage of Wikidata, and should not be interpreted as authoritative institution counts. The resulting estimated institution counts per country are summarised in Table 2. The results of this coverage estimation reveal a mixed picture across regions. For hospitals, Wikidata overestimates in certain countries the number of institutions relative to our Eurostat-derived figures in several countries,for which authoritative national sources confirm a substantially lower count of hospitals per inhabitant when small clinics and outpatient facilities are excluded. We hypothesise that Wikidata conflates hospitals in the strict sense with smaller clinical facilities such as day clinics, medical centres, and outpatient units, inflating the apparent count. Although Wikidata is also known to contain entries for former hospitals, this factor alone is unlikely to fully explain the observed overestimation; rather, it points to a broader issue of categorical ambiguity in Wikidata's class hierarchy, where the boundary between *hospital* and adjacent classes is inconsistently applied. This constitutes one of the known data quality issues of the platform. For schools, the picture is more encouraging. Estimated coverage ranges roughly between 20% and 50% depending on the country. While this range is wide, it is broadly consistent with what one would expect from a community-curated knowledge base: large countries with well-resourced Wikidata communities (e.g. Germany, France) tend to have higher coverage, while smaller or less-resourced countries show sparser representation. Crucially, the direction of error is likely to be underestimation rather than overestimation for schools, making the Eurostat-derived figures a more reliable upper bound. For local governments, we use the Eurostat LAU register as a proxy, equating one LAU entry to one local government entity. This assumption is violated in several cases. Lithuania in particular shows a significant outlier, where the Wikidata count of entities classified as local governments far exceeds the corresponding LAU count. We attribute this to a hereditary issue in subclass hierarchy: certain administrative entities that are subclasses of *local government*) in Wikidata do not correspond to distinct LAUs in the Eurostat register. This violates the one-to-one mapping assumption underlying our estimation, and is noted as a known limitation of the current approach. But overall, the country mean coverages (with top values clipped to 100%) seem very sensible with 35.42% of estimated local governments, 73.20% of estimated hospitals and 29.42% and 19.67% of estimated primary and secondary schools respectively covered in our dataset.

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.

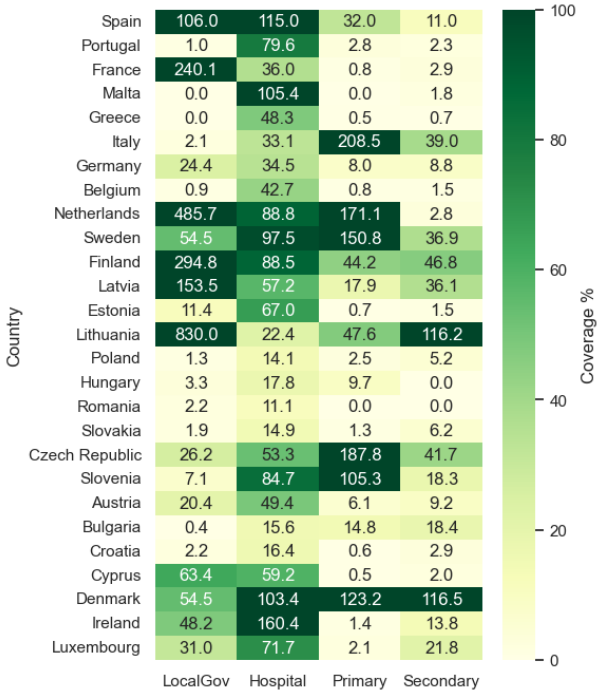


Figure 2. This heatmap display in percentage the ratio of extracted entities from Wikidata corresponding to selected classes or subclasses over estimation of real organizations. The estimation for Loval Governments is takes as the number of Local Administrative Units divisions per country, and for the rest we have taken Eurostat's number of hospital beds per inhabitant and enrolled students and applied a conservative estimation of beds or students per hospital or school respectively. Overall, the dataset covers significant parts of estimated organizations. In some cases it even surpasses the estimation. This is most likely due to mismatches in Wikidata labeling extra entities caused by class overlap, rather than underestimating the real number, but it must be considered that these are still rough estimates.

  - 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.

  In initiatives like the European Health and Digital Executive Agency's *Public Procurement Assessment in the Healthcare Sector* project, accurate differentiation between these subclasses is required to accurately profile healthcare services as part of the study to optimize procurement processes, effectively allocate medical resources, and enhance healthcare service delivery across Europe.
- **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.

  Understanding the prevalence and nature of joint procurement among local councils is vital for enhancing public spending efficiency. A study highlighted existing cross-border joint public procurement initiatives in Europe, identifying best practices and challenges . By exploiting these collaborations, administrations can identify opportunities for cost savings and the development of standardized procurement practices. As an example, the development of digital tools is easily applicable across similar administrations, though they are seldom doing so. This case would be the perfect application of joint procurement.

– **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

  Analyzing the geographical distribution of primary and secondary schools is essential for informed educational planning. For instance, research has examined educational attainment and inequality within European regions, emphasizing the role of spatial characteristics in educational systems . By understanding these patterns, authorities can address disparities in educational access, allocate resources effectively, and plan infrastructure developments to meet the needs of diverse populations.

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. While our choice of the official WDQS endpoint ensures alignment with the canonical RDF representation of Wikidata and is maintained directly by the Wikimedia Foundation. We acknowledge there exist high-performance alternatives like the University of Freiburg's QLever [2], operating over full Wikidata dumps. It can support more complex analytical queries with fewer runtime restrictions. However, such services have different update cycles, availability guarantees, and reproducibility considerations compared to the official endpoint. Due to hardware-specific memory constraints, hosting full Wikidata dumps was deemed unfeasible for this study. Instead, we prioritized a targeted extraction strategy designed to be reproducible across any organizational class. To further support transparency and reproducibility, we provide our curated dataset—comprising the results retrieved directly from the SPARQL endpoint—in our public repository[83]. While utilizing official infrastructure avoids the computational overhead associated with self-hosting, we recognize that the latter might have offered a more robust alternative given the various limitations encountered during the process [38]. To mitigate these constraints, specifically query result caps and request timeouts, queries were performed iteratively in discrete, manageable 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. These runtime issues encountered suggest that, in retrospect, a high-performance endpoint would have offered a more straightforward solution for obtaining the dataset.

This 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 [83]. The final dataset exhibits an unbalanced distribution that we assume is representative of the reality of this problem. It is represented in Figure 18. 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.

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. Within this work, low-context refers specifically to the linguistic isolation of the input strings; the entities' only feature are names, without surrounding text or sentences as is common in Named Entity Recognition Tasks. Our approaches are organized into three families: **(i) rule-based heuristics**, leveraging interpretable country and domain-specific keyword patterns. Rule-based methods have been used since the establishment of early industrial standards and are well-known in NLP tasks for providing a transparent, expert-grounded baseline that ensures high interpretability. By utilizing these systems, we attempt to capture explicit institutional markers that remain traceable to regulatory and regional naming conventions.

While rules provide clarity, they are often insufficient for "cold-start" scenarios without prior examples: **(ii) zero-shot classification using Natural Language Inference (NLI)**, which enables label prediction using large language models (LLM) without training. This allows us to categorize entities across languages instantly, providing a flexible framework that adapts to evolving classification targets without the overhead of traditional supervised learning. This technique has become a modern cornerstone for addressing zero-shot scenarios.

Finally, **(iii) lightweight embedding-based classifiers**, including logistic regression and support vector machines (SVMs) trained on static semantic representations complete our approach. Embedding methods exploit the latent semantic structure of text and have gained popularity since the introduction of large-scale pretrained transformers, which have revolutionized text classification by capturing contextual relationships. We utilize these representations to learn decision boundaries that partition the latent semantic space into classes.

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. There are also the added benefits of environmental sustainability and computational efficiency of smaller models that are sufficient for the proposed tasks. 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, our approach proves 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.

Consequently, 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. All experiments were conducted on a localized hardware setup with significant memory constraints. The primary computational bottleneck was a system equipped with 16GB of unified RAM; this limitation necessitated the use of lightweight architectures or aggressive quantization techniques. Furthermore, the environment was restricted to 100GB of available solid-state storage for data persistence which mandated careful management of the data footprint, particularly regarding the high-dimensional vector embeddings and model parameters.

Similarly, our methodology excludes remote deployment of LLMs and proprietary models. This decision was reflects this need for reproducibility and long-term stability; remote LLM services are subject to version drift and

availability constraints, making exact replication difficult. Running inference locally enabled systematic evaluation at scale without dependence on external quotas or cost constraints. Thus, a strict on-premise deployment scenario is enforced, mirroring the operational reality of many public sector and security-sensitive organizations where reliance on third-parties is often precluded by data sovereignty requirements.

Our methodology also compares different classification structures across domains, recognizing that the taxonomy complexity of target labels influences task 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.

We adopt the Massive Text Embedding Benchmark (MTEB) [49] 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. DeepSeek-R1 [30] was selected for its open nature, strong multilingual capabilities, competitive performance, and efficient local deployment compared to other LLMs of similar scale. 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. In our case we utilize DeepSeek-R1-Distill-Qwen-7B, the 7B distilled variant, deployed locally via the Ollama framework to ensure control over inference, reproducibility, and adherence to data constraints. The model was queried with the standardized prompt shown in the annex, listing 3 mirroring the survey distributed to experts, obtaining 5

keywords per class and country. 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 and automatic generative classification

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 . This approach, though it may not capture the canonical language of every name, ensures a common standardization.

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. Therefore, we evaluated task performance with and without spaCy-based tokenization and lemmatization using language-specific models, but, even though they may help the overall classification task, we still 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.

[1]Translations: *Bund* = federal, *Gymnasium* = high school, *Realgymnasium* = science-oriented high school, *Gleis* = track, *Dorf* = village.

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 reporting results to countries where the least frequent label has at least 30 examples. While meaningful patterns can emerge with as few as 5 instances, this threshold avoids over interpreting noise and ensures metrics do not exhibit unacceptable variance.

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. As our first approach 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.

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[11], a multilingual extension of RoBERTa. This model is fine-tuned on the XNLI (Cross-lingual Natural Language Inference) corpus[**?**], 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.
- **`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 [8]. 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 [39].
- **`MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7`**[4] is a multilingual adaptation of the DeBERTa-v3 architecture[32]. The model is fine-tuned on a multilingual NLI corpus combining XNLI, MultiNLI[70], and WANLI[43], along with high-quality translations.
- **`MoritzLaurer/multilingual-MiniLMv2-L6-mnli-xnli`**[5] is a compact multilingual NLI model based on the MiniLMv2 architecture[68], 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, entailment-based methods are not recommended for this strategy and preliminary results indicated low

[2] https://huggingface.co/joeddav/xlm-roberta-large-xnli
[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

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.

performance, likely due to confusion introduced by the added context. Instead, to test NLI with few-shot examples we test a generative method as well. Similarly to the generation of rules we employ DeepseekR1:7B wth a custom prompt (Annex, listing 4). Target classes are provided as well as a series of examples sampled equally from all countries. Unlike entailment-based methods, which rely on strict logical relationships, generative classification leverages the LLM's ability to interpret and contextualize ambiguous or nuanced inputs. This method is particularly valuable for tasks requiring multi-label assignments or handling overlapping roles (e.g., distinguishing between primary and secondary schools), though admittedly it benefits from richer contexts. Nonetheless it can serve as a representative baseline of generative classification methods and it shares the domain adaptability of the entailment method.

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.

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. In the case of logistic regression, functions as a linear probe test, which is a common practice to evaluate the quality of learned features.

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 [49].
- **`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.
- **`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.

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.

Having established baseline classification results with frozen pre-trained embeddings, we turn to the question of whether adapting the embedding space to the specific characteristics of the task and dataset can yield further

[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

improvements by optimally structuring the representation space for organizational classification. We investigate fine-tuning the encoder through a learning objective that optimizes downstream classification. Further, we compare if existing embedding models capture the nature of the organization class, addressing the question of whether the name contains taxonomic attributes in its semantic features.

Several established methods were considered for this purpose. SetFit [64] is a few-shot fine-tuning framework that couples contrastive training with an integrated classification head, making it less suitable here since our goal is to produce improved embeddings that remain comparable across the same classifier architectures used in our other experimental conditions. SimCSE [9] offers both supervised and unsupervised variants and has been shown to produce highly discriminative sentence embeddings, of which we turn into the supervised variant of SimCSE. While applicable in principle, it does not naturally accommodate multi-label settings and does not explicitly leverage the relational structure between categories. A more specialized alternative we considered is Hierarchical Joint Supervised Contrastive Learning (HJCL) [72], a Hierarchy-aware method that weights the contrastive signal between label embeddings according to their semantic distance in the taxonomy. While HJCL demonstrates modest improvements over standard SupCon in benchmarks with deep taxonomies, it is unnecessary for our simpler taxonomy and not directly comparable to the framework shared by other methods as it inputs the label hierarchy graph as well. Given these considerations, we adopt Supervised Contrastive Learning [37] as the fine-tuning objective.

Contrastive learning was originally developed in the context of self-supervised representation learning for images, where the core idea is to train an encoder to pull together an anchor and a semantically related positive sample in embedding space while simultaneously pushing apart the anchor from negative samples drawn from the remainder of the batch. Since the original problem does not have labels, the original paradigm is self-supervised: positive pairs are constructed through data augmentation of the same instance. [37] extend this paradigm to the fully supervised setting with the Supervised Contrastive (SupCon) loss, which rather than pairing an anchor only with an augmented version of itself, all samples sharing the same class label within a batch are treated as positives. This allows the loss to simultaneously pull together entire clusters of same-class instances while pushing apart samples from different classes, resulting in a more globally coherent and discriminative representation space . The core difference with SimCSE is the latter uses one positive, many negatives per anchor, structurally equivalent to the N-pairs loss. SupCon loss uses many positives per anchor. In our setting this distinction is particularly meaningful: an anchor such as a Local Government Administration has multiple valid positives in a batch (other local governments from different countries), collapsing that to a single chosen positive would discard supervision signal and possibly prioritize regional features instead.

However, one adaptation is necessary: the standard SupCon loss formulation was designed for single-label classification. Our setting introduces an important complication: organization names may belong to multiple categories simultaneously. A *university hospital*, for instance, carries both the `hospital` and `university_hospital` labels, or a *primary school* can also be a *secondary school*. We resolve this by defining the positive relationship as follows: two samples x_i and x_j with labels $\mathbf{y}_i, \mathbf{y}_j \in \{0,1\}^K$ are treated as a positive pair if and only if they share at least one active label, that is, $\exists k \in K$ such that their dot product $\mathbf{y}_i \cdot \mathbf{y}_j \geqslant 1$. To illustrate with a concrete batch example: suppose a batch contains a primary school. Any other organization in the batch that is also a primary school, be it also a secondary school or a hospital (not likely) or any other class, will be treated as a positive pair. Organizations with no active label among the five categories never form positive pairs with any sample, but remain present in the denominator of every anchor's softmax, functioning as hard negatives that tighten the labeled clusters. We reproduce here the resulting multi-label Supervised Contrastive Loss for an anchor a in batch $\mathcal{A} \subset \mathcal{X}$:

$$\mathcal{L}^{\text{sup}} = \sum_{a \in \mathcal{A}: |\mathcal{P}(a)|>0} \frac{-1}{|\mathcal{P}(a)|} \sum_{p \in \mathcal{P}(a)} \log \frac{\exp(\mathbf{z}_a \cdot \mathbf{z}_p \,/\, \tau)}{\sum_{b \in \mathcal{A}} \exp(\mathbf{z}_a \cdot \mathbf{z}_b \,/\, \tau)} \tag{2}$$

where $\mathbf{z}_x$ is the L2-normalised mean-pooled encoder output of a sample $\forall x \in \mathcal{X}$, $\mathcal{P}(i)$ denotes the positive pairs and τ is the temperature hyperparameter.

We fine-tune `intfloat/multilingual-e5-base`, our baseline model, using the SupCon objective, following a two-stage training protocol: in the first stage the encoder is optimised end-to-end with the contrastive loss; in the second stage the encoder weights are frozen and the classification heads are trained on top of the resulting

representations, exactly mirroring the evaluation setup of our frozen-embedding baselines and ensuring comparability across conditions. Training proceeds for up to 200 epochs with early stopping triggered after 10 consecutive epochs without improvement in validation contrastive loss. We use AdamW with a learning rate of 2×10^{-6}, and temperature set to $\tau = 0.5$ after observing quick convergence to a solution with more aggressive parameters. Gradient norms are clipped to 1.0 for stability. Batches of size 64 are constructed using a country-balanced sampler that interleaves samples across countries in round-robin fashion, ensuring that no single country dominates a batch and that the within-batch distribution of positive and negative pairs is geographically diverse. The dataset is partitioned into train, validation, and test sets at a 50/25/25 ratio. This setup converged to the lowest validation loss after 8 epochs.

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 Although large quantities of training data can be extracted for some classes from Wikidata, its availability is highly uneven and cannot be assumed a priori for all organizational types and countries. Even in our examples, class availability across all countries and languages is not guaranteed nor its unbiased distribution. As an example, for all three considered domains we have more than 1000 instances per class, but in many countries we are left with less than a dozen. Since a large part of conclusions centers on analyzing present conventions on naming we cannot exclude any of these cases due to non available data. In the case of Malta only six hospitals do exist, but that does not entail they cannot exhibit characteristic linguistic patterns. We have anticipated this data scarcity issue and integrated as a design choice to develop and evaluate methods that generalize across arbitrary class definitions, rather than optimizing performance for a fixed, well-populated label set. While relaxing this assumption would likely improve results for specific classes, it comes at the cost of generalization of the methods to specific use-cases. While it does not affect our other multilingual techniques, in the case of rule generation we do partition training per country for TF-IDF and Counter keyword generation. Therefore, only in these cases, we restrict performance evaluation to countries-class partitions for labeled data exceeds a threshold above 30 instances. Partitions below 30 will not be reported in average F1 metrics.

To benchmark the performance of automatic 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. We evaluate this baseline only on the answered regions.

Further, we will perform pairwise-statistical tests between experiments to determine if differences in F1 scores are statistically significant. For rule-based methods we will use the correctness tables and permutation tests. If a certain parameter is to be evaluated that has several experiments associated (eg. a certain type of preprocessing) we will run a stratified permutation test where each experiment is a strata. Further evaluation of our rule-based technique will include analysis of the corpus, language features and their correlation with F1 scores. In the case of NLI we instead use the classification result confidence for the permutation tests. We do not run all pair combinations but designate `xlm-roberta` as the baseline, as it is a stable and well-established model in inference tasks. For embeddings, we use correctness tables, and our selected baseline is `multilingual-e5-large`: it is a well-established and widely adopted multilingual embedding model that has demonstrated strong and consistent performance across a broad range of retrieval and classification tasks. We also provide three complementary geometric analysis of embedding quality: class separability, isotropy and label-embedding alignment.

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

We have conducted a total of 122 experiments, not including parameter optimization runs. Each experiment varies in either: domain, method, taxonomy structure, pre-processing, inference or embedding method or classifier head. Experiments can be identified by a unique ID with four segments separated by dashes. The first identifies the three domains, the second the technique: the underlying methodology, either rule-based, natural language inference or embedding-based. The third is the method referring to specific algorithms. The fourth is an increasing integer that for the rest of the configuration parameters: model, taxonomy structure, number of examples and classifier head. Some smaller hyperparameters are auto-selected in some cases from a validation set. As discussed throughout the evaluation, our three target domains exhibit differences in classification performance (Figure 3), reflecting varying levels of structural complexity and lexical regularity in organization names.

Domain	Structure	Mean F1	Var F1
Medical	3-class	0.2734	0.0218
Medical	nested-class	0.4391	0.0462
Administrative	2-class	0.7245	0.0221
Education	3-multiclass	0.5371	0.0516

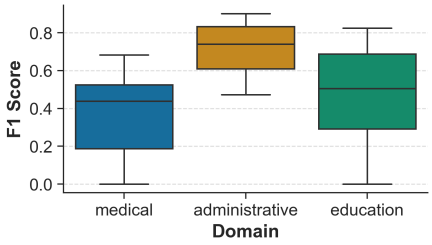


Figure 3. The table shows F1 scores for the total 122 experiments across domains and structures. Score differences correlate to the complexity of the structure. In particular, for the medical domain we can propose the task as a nested taxonomy or two independent classes at classification time. The optimal structure will depend for each technique. On the right, a boxplot aggregates F1 metric distribution for all-rule based experiments per domain, this time different structures are aggregated together.

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 [85]. Datasets and generated embeddings can be found on [83].

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 are contained in the annex, table 5 and figure 18a.

Performance can 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, or education where terms like *hospital*, *clinic*, and *school* are both semantically distinctive and cross-linguistically consistent. But in our dataset 24 languages are present.

Our gold-standard experiment based on a manually curated rule set demonstrated limited generalization across the dataset, despite being informed by regulatory frameworks and institutional naming norms. Our findings suggest

that these expert-informed linguistic rules significantly underperform in the face of real-world variability and complexity. A contributing factor could be the ambiguity of Wikidata labels, which frequently do not reflect 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. We have selected as benchmark our implementation of a simple counter algorithm that selects the words most present in the training target classes. We have not opted for expert rules as they are only available for the medical domain and three countries and lacking the possibility to compare to all experiments.

In terms of classification structure, we tested a nested approach for the medical domain. It proved a 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. To confirm this, we have tested a two-sided stratified permutation test. Strata are the pair of experiments that only differ in structure. They are six strata in total (3 pre-processing $\times$ 2 rule generation algorithms). The result confirms that the nested structure significantly outperforms simultaneous three-class prediction ($\Delta F1 = 0.0049$, $p < 0.0001$[10]). Therefore on the final reported experiments we will only include nested structure.

In our comparison of pre-processing strategies, we observed only marginal but significant improvements in classification performance. There are some implementation challenges to deploy these techniques due to the integration of multiple language-specific tokenization models and dictionaries. SpaCy-based tokenization slightly outperformed no pre-processing achieving statistically significant results on a stratified permutation test ($\Delta F1 = 0.0089$, $p < 0.0001$). Similarly, Wikdict-based compound word decomposition is also significantly better than no pre-processing at the aggregate level ($\Delta F1 = 0.0055$, $p < 0.001$). Between tokenization and decomposition, tokenization reports better F1 scores ($\Delta F1 = 0.0145$, $p < 0.001$). However, these are the aggregated findings and the motivation for decomposition pre-processing was to address specific languages such as german. After running stratified permutation tests filtering the correctness tables by country we find interesting results (Figure 4). Decomposition only performs better for the following languages: german, dutch and finnish. These languages have a high use of compound words. Notably, Austria does not share Germany's trend result indicating that even if they share the same language, regional naming conventions may have a lesser presence of compound words. For another 7 tokenization does outperform decomposition and for the rest the results are inconclusive. Still, we will use SpaCy tokenization for the following comparisons as the domains where and margins by which decomposition is better are very limited.

We therefore now compare the three main methods of rule-based approaches: LLM generated rules, counter and TF-IDF algrithm generated rules using the nested taxonomy for the medical domain and tokenization. We compare them running pairwise permutation tests between all pairings with the results shown in figure 6. 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. 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.

Among all rule-based methods, the best-performing approaches were the keyword selection algorithms, with the counter-based algorithm outperforming LLM rules ($\Delta F1 = 0.0166$, $p < 0.001$) but particularly the refined TF-IDF best-k-words method. This technique outperforms LLM generated keywords by a greater margin ($\Delta F1 = 0.0204$, $p < 0.001$) and performs very similar to the counter algorithm with a marginal improvement ($\Delta F1 = 0.0038$, $p < 0.001$) across domains, with all comparisons surviving Holm correction. 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.

[10] All permutation tests run 10000 permutations. Due to the high volume of test data, most executed statistical tests throughout this study will report significant results.

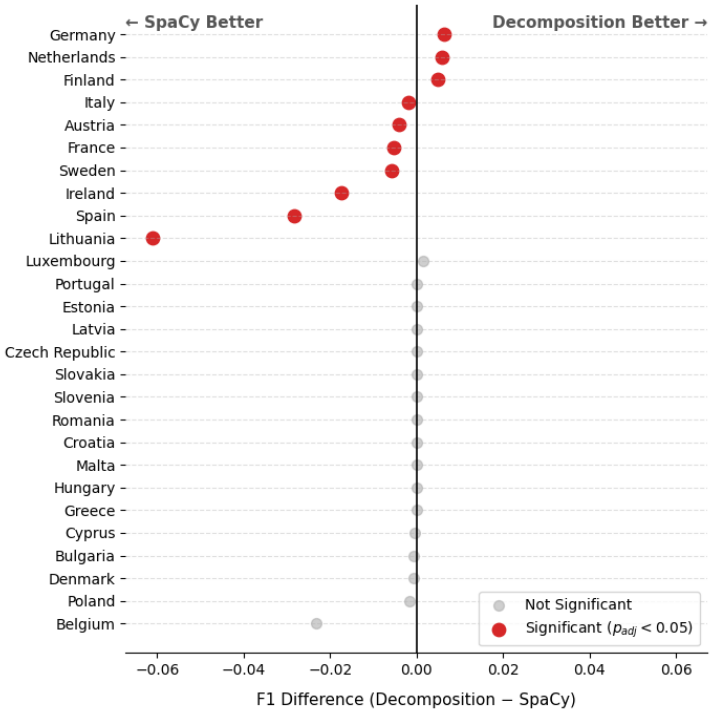


Figure 4. This plot shows $\Delta F1$ between experiments run with SpaCy tokenization and Wikdict decomposition. Negative differences indicate tokenization performs better and positive ones decomposition. Bonferroni-corrected p-values below 0.05 of the country-filtered two-sided stratified permutation tests are highlighted in red to show significance. We can observe german. dutch and finnish perform better with decomposition pre-processing, while seven other countries do so with tokenization. Most countries are inconclusive.

To conclude this discussion we present some analytical results on the corpus. It comprises 182,519 organizational entities totalling 572,685 tokens, with 150,056 unique: a type-token ratio (TTR) of 0.262. This indicates substantial lexical repetition. On average, entities span 23.9 characters and 3.1 tokens (std. 2.0), suggesting names are predominantly short 3-word expressions. The hapax ratio of 0.719 reveals that roughly 72% of unique tokens appear only once, pointing to a long tail of rare naming terms. Morphological diversity is low, entropy values are high indicating a broad and unpredictable distribution, which is unsurprising given the multilingual nature of the corpus (table 5. At the country level, there is considerable variation driven primarily by corpus size and language morphology. Morphologically rich languages stand out clearly: Estonia and Croatia exhibit exceptionally high suffix and prefix diversity. Romania and Poland produce the longest average entity names, while Finland and Sweden yield the shortest tokens on average (table11. We also show Spearman correlations between mean F1 scores per country and linguistic features. We have extracted language family and syntactical features from `lang2vec` library and corpus statistics to establish the most correlated features with high and low F1 scores, as shown in table 6. The strongest signal is syntactic: languages where the possessor follows the noun (e.g., Romance and Scandinavian) are associated with substantially higher F1, while the reverse order (Germanic and Slavic languages) is equally strongly associated with lower F1. This is almost certainly not a direct causal relationship and most likely reflects differences between language families. Corpus richness consistently helps: countries with more total tokens, more unique tokens, and more entities in the corpus all achieve higher F1. Higher n-gram entropy also correlates positively, indicating that lexically diverse corpora expose the model to a wider variety of surface forms for the same organization types. On the negative side, high morphological diversity and high compression ratio both hurt performance. Among language families, Indo-European and North/East Scandinavian languages correlate with better outcomes, while West Germanic (Franconian, Macro-Dutch) and South-Western Slavic subfamilies fare worse. Taken together, the results suggest that classification F1 is jointly driven by training data volume in entities and tokens, predictability of institutional keywords, and the morphological simplicity of the target language.

Method	Mean F1
Expert Rules	0.1480
LLM Generated Rules	0.3720
Counter Algorithm	0.4861
TF-IDF	**0.6126**

Figure 5. This table displays F1 Scores for expert rules (in available domain and countries), LLM generated, counter and TF-IDF algorithms averaged across the three domains. Only nested structure and tokenization configuration is considered. Highest F1 method in bold.

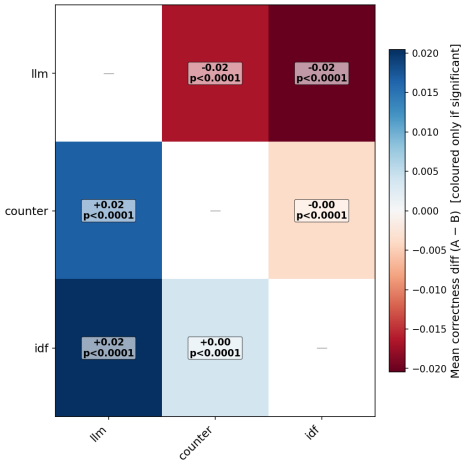


Figure 6. The heatmap displays $\Delta F1$ between the LLM generated, counter algorithm generated and TF-IDF generated rules, all run with tokenization and nested structure (if applicable). Full results across all experiments against the baseline can be found in the annex, figure 18a. Experiment A is displayed on the x axis and experiment B on the y axis. The difference is calculated $F1(A) - F1(B)$. Positive values indicate the c axis experiment performs better and viceversa. Bonferroni-corrected non-significant p-values obtained from pairwise permutation tests are not shown.

Metric	Value
Mean Char Length	23.869
Std Char Length	14.096
Mean Token Count	3.138
Std Token Count	1.950
ttr	0.262
Hapax Ratio	0.719
Suffix Diversity	0.029
Prefix Diversity	0.029
Token Entropy	13.574
Bigram Entropy	16.193
Trigram Entropy	16.597
Mean Compression Ratio	0.341
Unique tokens	150056
Total tokens	572685
Total entities	182519

Table 5

The table shows overall lexical statistics over the whole corpus. Results per country are shown in Table 11

Figures 19, 20 and 21 in the annex present the keywords that inform our tested best classifier configuration on the three domains. We show keyword coverage rates of the true positive instances and, on the same scale to favor comparison, the number of false positives for that keyword. In the medical domain (Figure 19), generic hospital terms achieve strong coverage with remarkably low false positive rates in most countries — *nemocnice* (Czech Republic, 67.7%), *szpital* (Poland, 65.7%), *hospital* (Ireland, 85.3%; Portugal, 85.0%; Spain, 71.1%), *μί* (Greece, 70.5%) — confirming that hospitals are lexically well-anchored across Europe. Complementary nouns such as

Feature	Type	ρ	p-value
Positively correlated with F1			
Possessor after noun	Syntax	+0.389	<0.001***
Total tokens in corpus	Corpus size	+0.186	<0.001***
Number of entities	Corpus size	+0.182	<0.001***
Trigram entropy	Corpus lexical	+0.178	<0.001***
Unique tokens	Corpus size	+0.161	<0.001***
Indo-European family	Language family	+0.137	<0.001***
North Germanic subfamily	Language family	+0.137	<0.001***
East Scandinavian subfamily	Language family	+0.137	<0.001***
Bigram entropy	Corpus lexical	+0.111	0.002**
F Czech-Slovak	—	+0.109	0.003**
Negatively correlated with F1			
Macro-Dutch subfamily	Language family	−0.133	<0.001***
West Germanic subfamily	Language family	−0.149	<0.001***
South-Western Slavic subfamily	Language family	−0.156	<0.001***
Molise-SKB subfamily	Language family	−0.156	<0.001***
Franconian subfamily	Language family	−0.162	<0.001***
Type-token ratio (TTR)	Corpus lexical	−0.196	<0.001***
Prefix diversity	Corpus morphol.	−0.198	<0.001***
Suffix diversity	Corpus morphol.	−0.217	<0.001***
Compression ratio	Corpus lexical	−0.232	<0.001***
Possessor before noun	Syntax	−0.389	<0.001***

Table 6

Spearman rank correlations (ρ) between country-level F1 and linguistic/corpus features. $\rho > 0$: higher feature value associates with higher F1; $\rho < 0$: the opposite. Only top and bottom 10 features with significant results ($p < .05$) shown. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

centre, *clínica* complete the coverage. A notable cross-linguistic pattern is the use of hagionyms[11]: *sint* in Belgium (26.5%), *szent* in Hungary (29.8%), and in Bulgaria (11.5%) all reflect Catholic and Orthodox naming traditions where hospitals are named after saints, but these terms carry elevated false positive rates. University hospitals share the same pattern using compound words such as *universitetshospital* or just academic such as *universitario*. Though some countries exhibit keywords of city names or proper nouns which yield elevated false positive rates. This is due to a lack of training data on these classes indicating training volume requirements should be more strict for building a practical application.

In the administrative domain, the picture is more uneven, with some countries achieving near-perfect coverage through a single unambiguous term: *taryba* (Lithuania, 94.3%), *gemeenteraad* (Netherlands, 99.4%), *rådhus* (Denmark, 87.0%), *câmara* (Portugal, 90.9%). Others struggle with either low coverage or false positive contamination. Administrative vocabulary overlaps with general language: *commune* covers 40.7% of French local governments but generates a 43% false positive rate, *regionale* and *parco* in Italy similarly, which indicates natural parks may be typed incorrectly in Italy as regional governments or a subclass of them. Some countries rely on highly specific keywords: German *verwaltungsgemeinschaft*, Finnish *kunnanvaltuusto* and *kaupunginvaltuusto*, Austrian *gemeindeamt*, precise but low-frequency. We see Austria and Germany, though sharing the same language, have significant deviations in naming patterns.

In the education domain there is confusion, as expected, with words that refer to both primary and secondary schools: high coverage but high false positive rates. But more specific terms appear for some regions: *lycée* in France (70.4% of secondary schools), *škola* in Czech Republic (87.6% of primary schools), *grundschule* (Germany, 44.1% of primary schools), *volksschule* (Austria, 63.2% of primary schools). The *gymnasium* family for secondary

[11] From onomastics: the name of a saint.

schools is the most consistent cross-linguistic anchor in the dataset, appearing across Germany, Slovakia, Slovenia, Croatia, Hungary, Sweden, Denmark, Austria, Greece, Lithuania, the Netherlands, and Slavic languages. Spain's uses official acronyms *ceip* (38.2%) and *ies* (31.9%), which could in this case also help identufy the proportion of public schools as it is an indicator of that as well. Italy's secondary schools have personal names, the most common appearing as keywords: *leonardo*, *vinci*, *bosco*, reflecting the Italian convention of naming schools after historical figures.

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 specially 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 and allowing lexical pattern mining, for tasks where domain semantics are linked to naming conventions. Naming patterns in EU public organizations tends to be very descriptive and generally standardized across regions. In some cases they are not trivial but high-coverage keywords appear that could potentially be useful for graph population and cleaning. However, cross-border differences are too pronounced to generalize to a multilingual method. 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 non-organizations. Performance also deteriorates with sparse labeled data, as LLM-generated rules have shown limited reliability in our experiments due to the very specific nature of naming patterns. 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.

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 the annex: Table 8 and Figure 18b. To evaluate if a few-shot paradigm improves performance we initially test prompting those inference-tuned models. However, the added examples confuse the models and do not perform well. Instead we opt for a generative LLM approach.

In contrast to our rule-based experiments, the nested classification structure did leads to much degraded performance in the NLI setting and is subsequently discarded from testing. We propose this behavior is caused by a combination of class imbalance and error propagation. Indeed, setting thresholds for confidences is non-trivial in this setup and to keep the zero-shot paradigm, its optimization on a training set was discarded. Instead, in our threshold-agnostic setup, the most probable label is selected with an 'others' class added. This can lead to misclassifications. Additionally, the NLI model may struggle to resolve fine-grained distinctions when presented with hypothesis classes that are semantically close but hierarchically nested. 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.

Similarly to the rule-based experiments we run pairwise, stratified by domain, two-sided, permutation tests. Instead of traditional binary correctness, which relies on majority confidence, we evaluate the continuous confidence. This measures the proximity of the model's posterior probabilities to the ground-truth labels. For an instance $x \in \mathcal{X}$ with ground-truth binary labels $\mathbf{y} \in \{0,1\}^K$ across K classes, and model confidence scores $\hat{\mathbf{y}} \in [0,1]^K$, the instance

Method	Mean F1
XLM-RoBERTa	**0.5606**
BGE-M3	0.5119
mDeBerta	0.5327
MiniLM	0.2974

Figure 7. This table displays F1 Scores for XLM-RoBERTa, BGE-M3, mDeBerta and MiniLM NLI models averaged across the three domains. Only flattened structure ais considered. Highest F1 method in bold.

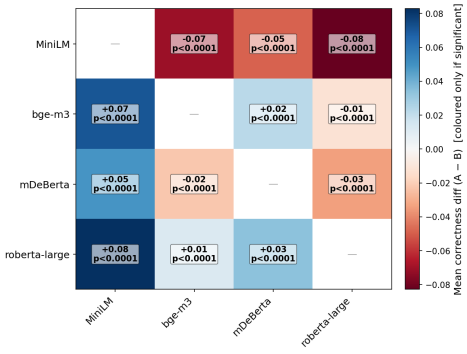


Figure 8. The heatmap displays $\Delta F1$ between the XLM-RoBERTa, BGE-M3, mDeBerta and MiniLM NLI models, all run with majority confidence classification and flattened structure for the medical domain. Full results across all experiments against the baseline can be found in the annex, figure 18b. Experiment A is displayed on the x axis and experiment B on the y axis. The difference is calculated $F1(A) - F1(B)$. Positive values indicate the c axis experiment performs better and viceversa. Bonferroni-corrected non-significant p–values obtained from pairwise permutation tests are not shown.

score $S(x)$ is defined as:

$$S(x) = \frac{1}{K} \sum_{k \in K} (1 - |y_k - \hat{y}_k|)$$

This metric rewards models for high-confidence correct predictions and low-confidence incorrect predictions, providing a more defined signal for model comparison. The pairwise comparison of NLI models indicates that larger models deliver superior classification performance. `roberta-large` is the strongest model overall, whose size makes up for the majority-english training as is not the case with bge-m3 and mDeBerta. It outperforms MiniLM by 0.08 points in mean correctness, mDeBerta by 0.03, and bge-m3 by 0.01. At the other end, MiniLM, as expected being the small model, is consistently the weakest as well. Though, the gaps between the top three are modest which suggests that multilingual models bge-m3 and mDeBerta are competitive and may be preferable in practice.

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 common in the literature 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 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.

Regarding generative LLM classification, we adopt the few-shot prompting strategy detailed in Listing 4. We construct domain-specific few-shot contexts by sampling labeled instances from the training split in a round-robin fashion, ensuring equal representation across both countries and classes. The model is instructed to emit all class predictions in a domain simultaneously given a name as input. To ensure comparability with the discriminative experiments, we adopt the same domain-based train/test splits throughout. Generative classification incurred by far the highest computational cost of all methods evaluated. An initial configuration capping the model's reasoning budget at 718 output tokens proved insufficient: the model failed to produce well-formed classifications within that window. Raising the parameter to 2048 tokens achieved some results but caused inference time to spike to over 72 hours for a single domain. Given this cost, and that performance did not surpass other approaches (Mean F1: 0.4791), we do not explore further other generative options. However it does surpass the simple rule generative approach which indicates that this paradigm performs better with more degrees of liberty to decide per instance. The results suggest that, for the name-only classification setting, the reasoning overhead of generative models confers no measurable benefit over encoder-based discriminative approaches, while imposing prohibitive latency for any real-world deployment scenario.

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 the annex: Tables 9 and 10 and Figure 18c.

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 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

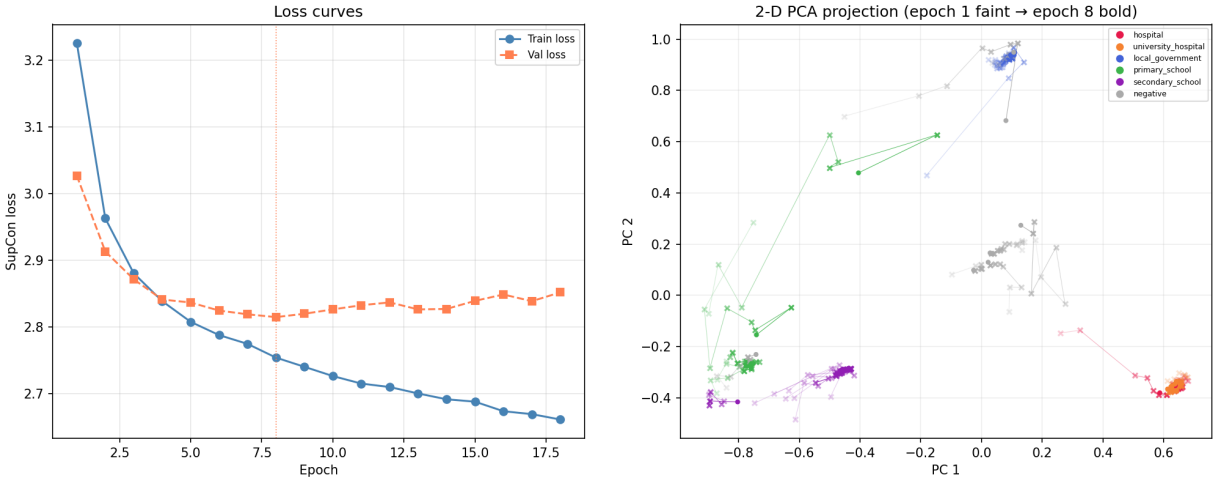


Figure 9. The left figure shows training and validation Supervised Contrastive losses. We have used a learning rate of 2×10^{-6} and temperature 0.6 with batches of size 64. Patience is 10, and the optimal found epoch is number 8. On the right a sample of 8 instances per class (including negative instances) is shown during the training process up until epoch 8. Opacity encodes the epoch index, and consecutive representations of the same instance are connected by a line. We observe that all classes are rapidly pulled together, and that neighboring classes, ie. those sharing some instances in the dataset, remain similarly close in the projection. Notably, the nested university hospital and hospital classes exhibit a higher degree of grouping than primary and secondary schools. This appears to validate our adaptation of the Supervised Contrastive Loss to avoid penalizing such nested or overlapping class instances.

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 limit that boundary adjustment for the classifier experiments.

As the final component of our embedding-based techniques, we move from static embeddings to a finetuned encoder trained with Supervised Contrastive Learning Loss on our target classes. Optimal validation loss is reached quickly so we switch to a low learning rate and an extended hyperparameter search, obtaining our best results in just 8 epochs. Figure 9 shows the loss curves and, most importantly, the trajectory of a sample of points on a 2-dimensional Principal Component Analysis plot. Points belonging to the same class cluster together, while other organizations are pushed towards the center. This appears to suggest that, contrary to our initial expectations, this method may struggle to generalize to unseen classes, at least given the large number of classes with limited examples each. This is consistent with the nature of the Contrastive Loss objective, which pushes negatives apart and pulls positives together without incorporating any broader semantic signal. On the positive side, embedding metrics are greatly improved by this method. Therefore, when target classes are known in advance and some labeled training data is available, this approach appears to be a viable option.

For testing we run four comparisons using the same stratified permutation tests specified in other families. We rely on correctness tables and test:

- We test our five models against each other across all classifier methods and domains (both cosine-based and classifier heads).
- Then we test the five classifier methods (cosine-similarity to label or 0-shot, cosine-similarity to random class prototype or one-shot, cosine-similarity to country prototypes or few-shot, logistic regression head and support vector machine head) against each other across all embedding models and domains.

Across all evaluated embedding models (Figure 11), performance was relatively consistent, with comparable results achieved across domains. Only the smaller `e5-small` model had a significant performance drop of 0.11 to the baseline `Multilingual-E5-Large`, which 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. Interestingly, between Mistral and Qwen

Model	Mean F1
finetuned-me5	**0.6384**
mistral	0.5796
qwen	0.5454
multilingual-e5	0.5298
e5-small	0.4165

Figure 10. This table displays F1 Scores for Multilingual-E5, Small Multilingual-E5, Supervised Contrastive Learning Finetuned Multilingual-E5, Mistral and Qwen models averaged across the three domains and five classifier methods. Highest F1 method in bold.

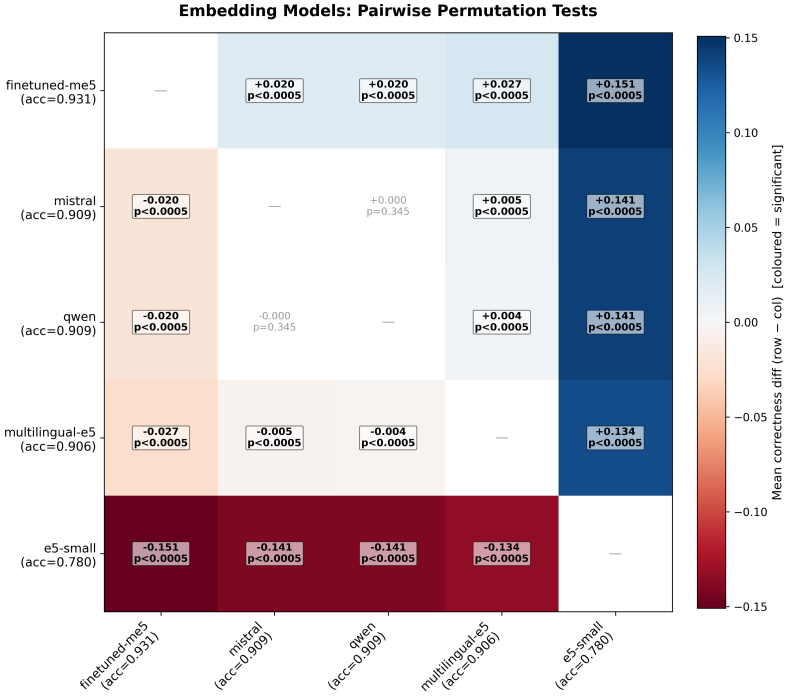


Figure 11. The heatmap displays mean $\Delta F1$ between the Multilingual-E5, Small Multilingual-E5, Supervised Contrastive Learning Finetuned Multilingual-E5, Mistral and Qwen models. Full results across all experiments against the baseline can be found in the annex, figure 18c. Experiment A is displayed on the x axis and experiment B on the y axis. The difference is calculated $F1(A) - F1(B)$. Positive values indicate the c axis experiment performs better and vice versa. Bonferroni-corrected non-significant p-values obtained from pairwise permutation tests are not shown.

no significant difference was observed 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. We have to consider our finetuned e5 model to again see a significant increase in F1 performance. These results clearly indicate finetuning an embedding model is much more important and less computationally demanding than utilizing larger models and embedding vectors for classification tasks. In terms of classification methods (Figure 13) we observe, as expected, that SVM heads are the clear winner with a 0.27 increase over the baseline cosine-similarity classification. Then comes Logistic Regression. Between the similarity methods, the 0-shot paradigm is the clear winner. This indicates the label is a bigger representative in the embedded semantic space than an example. Between having several examples or one, there is a clear difference in favor of providing several class prototypes (one per country), which suggests, regional variance is translated to distance between embeddings. 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. We can interpret the LR head scores as a Linear Probe test. Our mean score of 0.69 indicates that classes within embedded space are reasonably well linearly separated, at least in our reduced taxonomy.

To conclude this section, we present an analysis of the embedding spaces produced by our models in terms of class separability, isotropy (how uniformly representations are distributed across dimensions) and label-embedding alignment in figure 14. FDR (Fisher´s Discriminant Ratio) measures how well the embedding space separates distinct classes. The dramatic increase from base models to finetuned-me5 suggests that task-specific training substantially reshapes the geometry of the space. This is particularly relevant in a corpus with many fine-grained and over-

Method	Mean F1
Clf (SVM)	**0.7647**
Clf (LR)	0.6966
Sim (0-shot)	0.4447
Sim (few-shot)	0.4303
Sim (1-shot)	0.3735

Figure 12. This table displays F1 Scores for 0-Shot, One-Shot and Few-Shot cosine similarity and LR and SVM Classifier heads averaged across the three domains and five embedding models. Highest F1 method in bold.

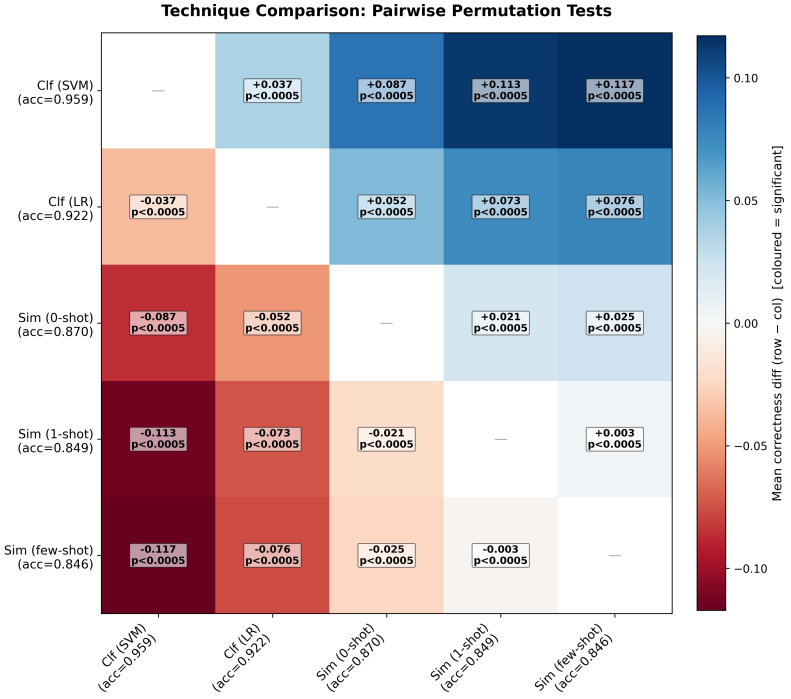


Figure 13. The heatmap displays mean $\Delta F1$ between 0-Shot, One-Shot and Few-Shot cosine similarity and LR and SVM Classifier heads. Full results across all experiments against the baseline can be found in the annex, figure 18c. Experiment A is displayed on the x axis and experiment B on the y axis. The difference is calculated $F1(A) - F1(B)$. Positive values indicate the c axis experiment performs better and vice versa. Bonferroni-corrected non-significant p-values obtained from pairwise permutation tests are not shown.

lapping organizational categories. Isotropy measures how uniformly embeddings are distributed across the space. Near-zero values across all models confirm the well-known anisotropy problem in similar tasks of the literature, where representations collapse into a narrow cone rather than spanning the full space. This means models are not exploiting their full representational capacity, and similarity computations may be dominated by a few dominant directions. The high FDR of the finetuned model despite equally near-zero isotropy suggests that finetuning does not fix anisotropy but rather learns to pack discriminative class information more efficiently into those same dominant directions, which is a effective but inefficient solution. Mean Reciprocal Rank (MRR) and Recall@k assess retrieval quality over neighbors. The setup is a kNN retrieval over entity embeddings: each entity is used as a query against the full corpus, and the ranking based on cosine similarity to all other entities. Even though kNN is not considered as a classification algorithm due to its dependance on class distribution, here it serves as a comparative measure of the neighboring instances classes between embedding models. MRR then captures how consistently an entity's closest neighbor in embedding space belongs to the same organizational class and Recall@k measures whether at least one same-class entity appears in the k nearest neighbours, a softer test of class neighbourhood cohesion. The slight divergence between the two for finetuned-me5 hints that finetuning sharpens the model's confidence in its top prediction at the cost of some broader coverage, a reasonable trade-off for entity disambiguation tasks where the top-ranked candidate is typically used directly. Taken together, the metrics suggest that while all models retrieve competently, only finetuning meaningfully restructures the embedding space geometry, which may matter most for robustness on rare or ambiguous entity types.

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

Model	FDR	Isotropy	MRR	Recall@10
e5-small	5.5388	0.0000	0.9178	0.9898
multilingual-e5	21.5714	0.0000	0.9521	0.9950
qwen	27.3535	0.0000	0.9420	0.9922
mistral	39.1506	0.0000	0.9556	0.9940
finetuned-me5	67.6518	0.0000	0.9598	0.9908

Figure 14. This table displays Fisher´s Discriminant Ratio (FDR), Isotropy, Mean Reciprocal Rank (MRR) and Recall@k using kNN for e5-small, multilingual-e5, qwen, mistral and our SupCon finetuned multilingual-e5 encoder models. It can be seen class separability (FDR) increases greatly on finetuning and all other measures remain similar. Anisotropy remains a problem even after finetuning consistent to what whas found in [9]. This behavior was proposed as intrinsic to attention mechanisms in [29].

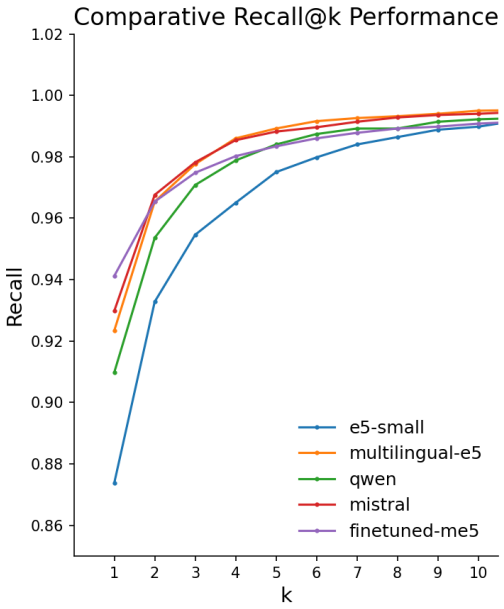


Figure 15. This line plot shows the evolution of recall@k for our encoding models. Most models begin with a good recall@1 with the winner being our finetuned me5. As k increases, finetuned-me5 is overtaken by mistral and later qwen, which points to a meaningful difference in how these models structure their neighborhoods. Finetuning pulls together the core of each class cluster potentially at the expense of its periphery that becomes disadvantageous at higher k.

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.

5. Results and Conclusions

5.1. Results

This study addressed the multilingual classification of public sector organization names across healthcare, administrative, and educational domains. Performance varied substantially across domains due to their taxonomy complexity. We evaluate three simple taxonomies, but still, the medical domain proved most challenging

We show that lightweight rule-based methods, particularly TF-IDF keyword extraction, can achieve strong performance in low-context, multilingual settings. However, pre-processing remains a challenge for multilingual standardized compound handling and tokenization, requiring adjustment. Each language requires different pre-processing to reach optimal results, but for a small performance drop a universal tokenizer like SpaCy outperforms no pre-processing and Wikdict-based decomposition overall.. Regional differences continue to play a significant role in language processing, but our tests report that TF-IDF based keyword extraction outperforms counter-based algorithms by a small margin. Expert generated rules fail to generalize or understand the corpus and perform poorly for classification. Algorithmic processing of texts still proves essential in understanding proper names where lexical signal is stronger than semantic or syntactic ones. 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. The interpretability of TF-IDF based rule methods provides clear outcomes that no other method can match. In this context, analyzing corpus keywords provides a summary of naming patterns easily that is easily understandable.

Natural Language Inference (NLI) models did not achieve competitive zero-shot performance, require careful calibration and are sensitive to fine-grained distinctions between semantically close classes. For unsupervised we recommend embedding-based cosine similarity instead. Small language models offer a comparative efficiency advantage to large generative ones, but their main limitation is adapting the task effectively to an entailment problem without providing supervision. This approach could prove better for name classification when more metadata is available. The best model proved `XLM-RoBERTa` but the difference is not too great from smaller multilingual models like `BGE-M3`, which we recommend. Further training of smaller language models is necessary to improve performance in Universal NLI tasks. The generative few-shot approach achieved similar macro-F1 under the 2048-token budget. This confirms that reasoning capacity alone does not compensate for the absence of a structured classification objective when inputs are limited to short name strings.

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, but smaller models do suffer from degraded results, which indicates there might be a threshold above which an embedding model size is sufficient for class representation tasks. Instead, performance was primarily dependent on the quality of the downstream classifier, suggesting that embedding-based pipelines can scale efficiently if well optimized. This family provides an alternative option for the zero-shot approach based on cosine similarity to labels. It is more computationally intensive but this can be expanded using finetuned embeddings, or clearer decision boundaries with trained classifier heads. Within similarity experiments, the addition of examples did not improve over using class embeddings, which indicates semantic representations of instances cluster around the class representation and show a promising approach for universal classification based on cosine distances. Similarly, regional differences appear to be represented in smaller clusters within as few-shot approaches perform better than one-shot approaches. The Supervised Contrastive Fine-tuned e5 model achieves a ery significant increase in performance, even without classifier being comparable to the best SVM+Mistral experiments. We attribute this to the dual training regime which proves the effectiveness of both choice and variety of training objectives.

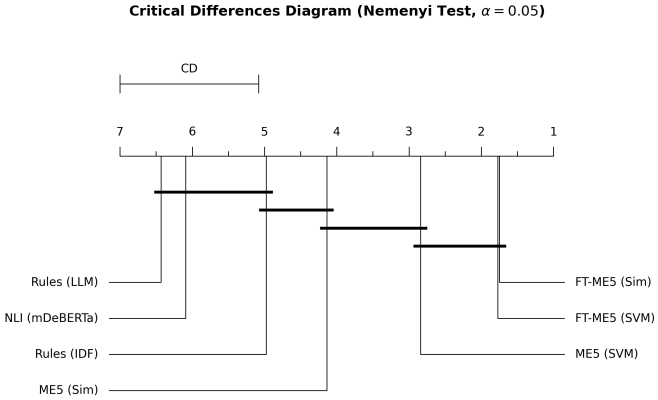
Upon manual inspection, we observed a highly unequal distribution of instances across countries, which may point to systemic coverage biases or incomplete curation within certain regions. This uneven representation can negatively affect model generalization. Semantic drift has also been observed when inspecting rule-classifiers keywords, primarily due to the mislabeling of instances. Several of our results point to issues with data quality within the Knowledge Graph that must be tackled regionally. 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.

To provide a unified comparison across classifiers, we present a Critical Difference (CD) diagram [14] in Figure 16b. To ensure a sufficient number of independent observations for the statistical test, we treat each country-domain combination as an independent dataset, yielding 72 observations per method (24 countries $\times$ 3 domains.

Returning to RQ1 on whether NLP techniques trained on Knowledge Graph data remain effective on name organization classification. We have carried an extensive suite of experiment across three families and report F1 scores between 0.61 and 0.83 for best rule-based experiments; between 0.47 and 0.77 for NLI experiments and between 0.68 and 0.90 for embedding-based experiments, where the range in each case reflects variance across domains rather than across model variants. The consistent ordering (embedding-based > rule-based > NLI) holds across mean F1. The three families offer different properties aside from performance, most notably, interpretable rules and omission of training data. Taken together, these results confirm that Knowledge Graph-derived supervision, training targets for classifiers, or contrastive anchors provides a viable and consistent signal for organization name classification across domains and languages. The answer to RQ1 is therefore affirmative, with the caveat that NLI zero-shot as tested using entailment and generative models does not provide strong results. Therefore training is

	MR	MED	MAD
Rules (LLM)	6.432	0.511	0.076
NLI (mDeBERTa)	6.091	0.689	0.115
Rules (IDF)	4.977	0.874	0.067
ME5 (Sim)	4.136	0.921	0.056
ME5 (SVM)	2.841	0.984	0.007
FT-ME5 (SVM)	1.773	0.989	0.006
FT-ME5 (Sim)	1.750	0.990	0.007

(a) Summary statistics from the tests. Mean Rank (MR) is the average per-dataset rank across all observations, lower is better. MED and MAD report the median and median absolute deviation of F1 scores.



(b) The diagram, ranks classifiers by their mean F1 across all domains and countries, and draws a horizontal bar, the critical difference, connecting any pair of methods whose performance difference is not statistically significant.

Figure 16. Friedman-Nemenyi analysis and Critical Difference Diagram across all country-domain pairs. Each classifier is ranked separately on each dataset (country-domain pair) and mean ranks inform position on the diagram's axis. Classifiers rank distributions are tested using Friedman to determine whether any significant difference exists among the classifiers as a group. As significant differences are detected, the post-hoc pairwise comparisons are conducted using the Nemenyi test, which computes a critical difference value. In the diagram, a horizontal bar connects all pairs that do not exceed this threshold and should therefore be treated as statistically equivalent; methods outside each other's bar are significantly distinguishable. The configurations entered into the comparison have been selected according to optimal results in the statistical test carried on the discussion: TF-IDF with tokenization, LLM-generated keywords, RoBERTa-large for NLI, cosine similarity on multilingual-e5 as the zero-shot embedding baseline, an SVM head on multilingual-e5 as the supervised embedding configuration, and both of those repeated on SupCon fine-tuned multilingual-e5 embeddings.

still necessary for a robust classifier and Wikidata proves essential in this task of achieving a Universal Classifier by providing flexible taxonomies against which we can query and optimize training to provide optimal results.

RQ2 asks whether interpretable patterns can be derived from cross-country analysis of the corpus itself and experimental scores of different methods. The keyword and language feature analysis across 27 countries reveals strong regional variation in naming conventions: all organization types present strong lexical standardized keywords. Some of these point to potentially misclassified entities in some Wikidata coutries. Therefore it is essential to consider regional variations when accounting for data quality in the platform. Some other notable insights is the use of names of catholic saints for hospitals, of historical figures for schools and toponyms for governments. Decomposition of words proves useful in only some languages such as german, finish and dutch. Strong performance in rule-based methods is positively correlated with corpus volume and lexical diversity but also language features such as trigram entropy. Performance is negatively correlated with prefix and suffix diversity and compression ratio. Between TF-IDF selection of keywords and a simple counter, performance varies across countries. TF-IDF favours the Netherlands, France, Slovenia, Austria, Spain, Denmark and Germany where institutional designators tend to be stable, morphologically simple, and class-specific, where inverse document frequency component that the other algorithm lacks, is higher. Counter performs better for Lithuania, Ireland, Czech Republic, Italy and Poland. The methodologies employed revealed distinct trade-offs. Rule-based methods offered interpretability and efficiency but struggled with dependency in training coverage. NLI-based approaches did not perform well with minimal labeled data, and 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.

5.2. 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 guide the development of more generalizable embedding space finetuning objectives.

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. But as they currently stand, the methods presented in this work represent a broad comparison study and could be better task-adapted. They may not be optimal enough to warrant direct contribution to Wikidata or analogous knowledge base. Although automated entity classification holds clear promise as a module within broader tasks such as Entity Linking, Named Entity Recognition or Graph Population and Cleaning pipelines, several open challenges remain: the construction of larger, more representative training corpora that span the full diversity and complexity of taxonomies; the development of generalizable encoder models whose classification geometry transfers reliably to unseen domains; and the detection and characterization of emergent properties that would signal a system is ready for deployment in a live graph setting. We regard these as the natural next steps, and we hope the comparative evaluation conducted here provides a concrete baseline against which future, contribution-ready systems can be measured. In spite of this, 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.

The problem addressed in this work can be understood as a sequence of larger challenges, each adding a layer of complexity. We begin with the general problem of classifying a short, low-context text span, not embedded within text like common tasks of Named Entity Recognition (the setting studied in FewNERD [15] and related benchmarks) into a fine-grained label space. Further hierarchical taxonomies adds the constraint that labels are not independent, introducing consistency requirements. Our nested versus flat experiments provide a preliminary approach, but dedicated hierarchical objectives remain unexplored. Finally, the innermost and most practically demanding layer is the multilingual, cross-region setting, where all of the above challenges compound with morphological diversity and uneven resource availability. As part of our evaluation we have developed a small 100-entity manually curated dataset which we considered presenting as another contribution. Within *Organization* they have *Government* and *Education* and within *Building*, *Hospital*. Besides the taxonomy mismatch of *Hospital*, contributing to FewNERD would require sentence-level annotated spans, which our dataset does not contain. We propose that Named Entity Recognition benchmarks does not need to be bound to the sentence-span paradigm, and structured, JSON-based

evaluation datasets constitute a legitimate alternative. A benchmark of this kind could progressively incorporate additional structured attributes common across administrative, procurement, and knowledge graph datasets: address, description, NACE codes, webpage, etc. This design would directly target the low-context retrieval stage that precedes full NER and entity linking.

5.3. Conclusions

latex

5.4. Conclusions

References

[1] A. Ancona, R. Cerqueti and G. Vagnani, A novel methodology to disambiguate organization names: an application to EU Framework Programmes data, *Scientometrics* **128**(8) (2023), 4447–4474. doi:10.1007/s11192-023-04746-x.

[2] H. Bast and B. Buchhold, QLever: A Query Engine for Efficient SPARQL+Text Search, in: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 647–656–. ISBN 9781450349185. doi:10.1145/3132847.3132921.

[3] M. Birti, A. Maurino and F. Osborne, Optimizing Large Language Models for ESG Activity Detection in Financial Texts, in: *Proceedings of the 6th ACM International Conference on AI in Finance*, ICAIF '25, ACM, 2025, pp. 856–863–. doi:10.1145/3768292.3770371.

[4] S. Bogdanov, A. Constantin, T. Bernard, B. Crabbé and E. Bernard, NuNER: Entity Recognition Encoder Pre-training via LLM-Annotated Data, 2024. https://arxiv.org/abs/2402.15343.

[5] J.A. Botha, Z. Shan and D. Gillick, Entity Linking in 100 Languages, 2020. https://arxiv.org/abs/2011.02690.

[6] camillop, company-classification-dataset-synt-v3, Hugging Face, 2024, Accessed: 2026-04-06. https://huggingface.co/datasets/camillop/company-classification-dataset-synt-v3.

[7] N.D. Cao, L. Wu, K. Popat, M. Artetxe, N. Goyal, M. Plekhanov, L. Zettlemoyer, N. Cancedda, S. Riedel and F. Petroni, Multilingual Autoregressive Entity Linking, 2021. https://arxiv.org/abs/2103.12528.

[8] 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, in: *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins and V. Srikumar, eds, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 2318–2335. doi:10.18653/v1/2024.findings-acl.137. https://aclanthology.org/2024.findings-acl.137/.

[9] T. Chen, S. Kornblith, M. Norouzi and G.E. Hinton, A Simple Framework for Contrastive Learning of Visual Representations, *CoRR* **abs/2002.05709** (2020). https://arxiv.org/abs/2002.05709.

[10] 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*, E. Riloff, D. Chiang, J. Hockenmaier and J. Tsujii, eds, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2475–2485. doi:10.18653/v1/D18-1269. https://aclanthology.org/D18-1269/.

[11] 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, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter and J. Tetreault, eds, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747. https://aclanthology.org/2020.acl-main.747/.

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

[13] E. Demidova, A. Dsouza, S. Gottschalk, N. Tempelmeier and R. Yu, Creating knowledge graphs for geographic data on the web, *ACM SIGWEB Newsletter* **2022**(Winter) (2022), 1–8–. doi:10.1145/3522598.3522602.

[14] J. Demsar, Statistical Comparisons of Classifiers over Multiple Data Sets, *Journal of Machine Learning Research* **7** (2006), 1–30.

[15] N. Ding, G. Xu, Y. Chen, X. Wang, X. Han, P. Xie, H. Zheng and Z. Liu, Few-NERD: A Few-shot Named Entity Recognition Dataset, in: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li and R. Navigli, eds, Association for Computational Linguistics, Online, 2021, pp. 3198–3213. doi:10.18653/v1/2021.acl-long.248. https://aclanthology.org/2021.acl-long.248/.

[16] H. Elsahar, P. Vougiouklis, A. Remaci, C. Gravier, J. Hare, F. Laforest and E. Simperl, T-REx: A Large Scale Alignment of Natural Language with Knowledge Base Triples, in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis and T. Tokunaga, eds, European Language Resources Association (ELRA), Miyazaki, Japan, 2018. https://aclanthology.org/L18-1544/.

[17] K. Enevoldsen, I. Chung, I. Kerboua, M. Kardos, A. Mathur, D. Stap, J. Gala, W. Siblini, D. Krzemiński, G.I. Winata, S. Sturua, S. Utpala, M. Ciancone, M. Schaeffer, G. Sequeira, D. Misra, S. Dhakal, J. Rystrøm, R. Solomatin, Ömer Çağatan, A. Kundu, M. Bernstorff, S. Xiao, A. Sukhlecha, B. Pahwa et al., MMTEB: Massive Multilingual Text Embedding Benchmark, 2025. https://arxiv.org/abs/2502.13595.

[18] 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.

[19] European Parliament and Council of the European Union, Regulation (EC) No 2195/2002 on the Common Procurement Vocabulary (CPV), as amended by Commission Regulation (EC) No 213/2008, 2008, OJ L 74, 15.3.2008; CPV version in force since 17 September 2008. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CONSLEG:2002R2195:20080915.

[20] Eurostat, Hospital beds by type of care (`hlth_rs_bds1`), 2024, Accessed: April 2026.

[21] Eurostat, Pupils and students enrolled in lower secondary education by programme orientation, sex and type of institution (`educ_uoe_enrs01`), 2024, Accessed: April 2026.

[22] Eurostat, Pupils and students enrolled in upper secondary education by programme orientation, sex and type of institution (`educ_uoe_enrs04`), 2024, Accessed: April 2026.

[23] Eurostat, Pupils enrolled in primary education by sex and type of institution (`educ_uoe_enrp04`), 2024, Accessed: April 2026.

[24] Eurostat, Correspondence table LAU – NUTS 2024, EU-27 and EFTA / candidate countries, 2025, 2025, Accessed: April 2026.

[25] Eurostat, *NACE Rev. 2.1: Statistical Classification of Economic Activities in the European Union – 2025 Edition*, Publications Office of the European Union, Luxembourg, 2025, Established by Commission Delegated Regulation (EU) 2023/137. https://ec.europa.eu/eurostat/documents/3859598/21633320/KS-GQ-24-007-EN-N.pdf.

[26] Executive Office of the President, Office of Management and Budget, North American Industry Classification System, United States, 2022, Technical Report, U.S. Census Bureau, 2022. https://www.census.gov/naics/reference_files_tools/2022_NAICS_Manual.pdf.

[27] B. Fetahu, Z. Chen, S. Kar, O. Rokhlenko and S. Malmasi, MultiCoNER v2: a Large Multilingual dataset for Fine-grained and Noisy Named Entity Recognition, 2023. https://arxiv.org/abs/2310.13213.

[28] I. García-Ferrero, J.A. Campos, O. Sainz, A. Salaberria and D. Roth, IXA/Cogcomp at SemEval-2023 Task 2: Context-enriched Multilingual Named Entity Recognition Using Knowledge Bases, in: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, A.K. Ojha, A.S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar and E. Sartori, eds, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1335–1346. doi:10.18653/v1/2023.semeval-1.186. https://aclanthology.org/2023.semeval-1.186/.

[29] N. Godey, É. Clergerie and B. Sagot, Anisotropy Is Inherent to Self-Attention in Transformers, in: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Y. Graham and M. Purver, eds, Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 35–48. doi:10.18653/v1/2024.eacl-long.3. https://aclanthology.org/2024.eacl-long.3/.

[30] D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi, X. Zhang, X. Yu, Y. Wu, Z.F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao et al., DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning, *Nature* **645**(8081) (2025), 633–638–. doi:10.1038/s41586-025-09422-z.

[31] Q. Guo, Y. Dong, L. Tian, Z. Kang, Y. Zhang and S. Wang, BANER: Boundary-Aware LLMs for Few-Shot Named Entity Recognition, in: *Proceedings of the 31st International Conference on Computational Linguistics*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B.D. Eugenio and S. Schockaert, eds, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 10375–10389. https://aclanthology.org/2025.coling-main.691/.

[32] P. He, X. Liu, J. Gao and W. Chen, DeBERTa: Decoding-enhanced BERT with Disentangled Attention, 2021. https://arxiv.org/abs/2006.03654.

[33] K.V. Holla, C. Kumar and A. Singh, Large Language Models aren't all that you need, 2024. https://arxiv.org/abs/2401.00698.

[34] C. Hu, Joint contrastive learning and belief rule base for named entity recognition in cybersecurity, *Cybersecurity* **7** (2025). doi:10.1186/s42400-024-00206-y.

[35] H. Hwang, Y. Jung, C. Lee and W. Go, A Nested Named Entity Recognition Model Robust in Few-Shot Learning Environments Using Label Description Information, *Applied Sciences* **15**(15) (2025), 8255. doi:10.3390/app15158255. https://www.mdpi.com/2076-3417/15/15/8255.

[36] H. Jani, Predicting NACE Codes Using Web Page Texts: A Machine Learning Approach for Enterprises in Austria and Germany, Master's thesis, Johannes Kepler Universität Linz, Linz, Austria, 2025, Faculty of Social Sciences, Economics and Business, Institut für Business Analytics und Technology Transformation. https://resolver.obvsg.at/urn:nbn:at:at-ubl:1-95426.

[37] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu and D. Krishnan, Supervised Contrastive Learning, 2021. https://arxiv.org/abs/2004.11362.

[38] A.N. Lam, B. Elvesæter and F. Martin-Recuerda, Evaluation of a Representative Selection of SPARQL Query Engines Using Wikidata, in: *The Semantic Web – ESWC 2023: 20th International Conference, Proceedings*, C. Pesquita, H. Skaf-Molli, V. Efthymiou, S. Kirrane, A.-C. Ngonga Ngomo, D. Dumas, J. Liuo, H. Alani, J. McCusker and I. Santana-Perez, eds, Lecture Notes in Computer Science, Vol. 13870, Springer Nature Switzerland, Cham, 2023, pp. 679–696. ISBN 978-3-031-33455-9. doi:10.1007/978-3-031-33455-9_40.

[39] M. Laurer, W. van Atteveldt, A. Casas and K. Welbers, Building Efficient Universal Classifiers with Natural Language Inference, 2024. https://arxiv.org/abs/2312.17543.

[40] W. Leyh, Interlinking Standardized OpenStreetMap Data and Citizen Science Data in the OpenData Cloud, 2017. doi:10.1007/978-3-319-60366-7_9.

[41] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie and M. Zhang, Towards General Text Embeddings with Multi-stage Contrastive Learning, 2023. https://arxiv.org/abs/2308.03281.

[42] P. Limkonchotiwat, W. Cheng, C. Christodoulopoulos, A. Saffari and J. Lehmann, mReFinED: An Efficient End-to-End Multilingual Entity Linking System, in: *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino and K. Bali, eds, Association for Computational Linguistics, Singapore, 2023, pp. 15080–15089. doi:10.18653/v1/2023.findings-emnlp.1007. https://aclanthology.org/2023.findings-emnlp.1007/.

[43] A. Liu, S. Swayamdipta, N.A. Smith and Y. Choi, WANLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation, *CoRR* **abs/2201.05955** (2022). https://arxiv.org/abs/2201.05955.

[44] H. Luo, Y. Jin, X. Li, X. Liu, R. Chen, T. Shang, K. Wang, Q. Wen and Z. Liu, DynamicNER: A Dynamic, Multilingual, and Fine-Grained Dataset for LLM-based Named Entity Recognition, 2026. https://arxiv.org/abs/2409.11022.

[45] L. Ma, K. Lu, T. Che, H. Huang, W. Gao and X. Li, PAI at SemEval-2023 Task 2: A Universal System for Named Entity Recognition with External Entity Information, in: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, A.K. Ojha, A.S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar and E. Sartori, eds, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 744–750. doi:10.18653/v1/2023.semeval-1.102. https://aclanthology.org/2023.semeval-1.102/.

[46] P.H. Martins, Z. Marinho and A.F.T. Martins, Joint Learning of Named Entity Recognition and Entity Linking, *CoRR* **abs/1907.08243** (2019). http://arxiv.org/abs/1907.08243.

[47] F. Moiraghi, M. Palmonari, D. Allavena and F. Morando, Zero-Shot Hierarchical Classification on the Common Procurement Vocabulary Taxonomy, 2024. https://arxiv.org/abs/2405.09983.

[48] MSCI and S&P Dow Jones Indices, Global Industry Classification Standard (GICS®) Methodology, Technical Report, MSCI and S&P Dow Jones Indices, 2024, First published January 2020, last updated August 2024. https://www.msci.com/indexes/documents/methodology/0_MSCI_Global_Industry_Classification_Standard_GICS_Methodology_20240801.pdf.

[49] N. Muennighoff, N. Tazi, L. Magne and N. Reimers, MTEB: Massive Text Embedding Benchmark, 2023. https://arxiv.org/abs/2210.07316.

[50] National Center for Charitable Statistics, National Taxonomy of Exempt Entities (NTEE) Codes, Accessed: 2026-04-05. https://nccs.urban.org/project/national-taxonomy-exempt-entities-ntee-codes.

[51] R. Orlando, P.-L. Huguet Cabot, E. Barba and R. Navigli, ReLiK: Retrieve and LinK, Fast and Accurate Entity Linking and Relation Extraction on an Academic Budget, in: *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins and V. Srikumar, eds, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 14114–14132. doi:10.18653/v1/2024.findings-acl.839. https://aclanthology.org/2024.findings-acl.839/.

[52] J. Portisch, O. Fallatah, S. Neumaier, M.Y. Jaradeh and A. Polleres, *Challenges of Linking Organizational Information in Open Government Data to Knowledge Graphs*, 2020, pp. 271–286. ISBN 978-3-030-61243-6. doi:10.1007/978-3-030-61244-3_19.

[53] M. Rizinski, A. Jankov, V. Sankaradas, E. Pinsky, I. Mishkovski and D. Trajanov, Comparative analysis of NLP-based models for company classification, *Information* **15**(2) (2024), 77.

[54] I. Savin, K. Chukavina and A. Pushkarev, Topic-based classification and identification of global trends for startup companies, *Small Business Economics* **60**(2) (2023), 659–689, Epub 2022 Mar 1. doi:10.1007/s11187-022-00609-6.

[55] P. Scharpf, C. Breitinger, A. Spitz, N. Meuschke, A. Greiner-Petter, M. Schubotz and B. Gipp, Entity Linking with Wikidata: A Systematic Literature Review, *ACM Comput. Surv.* **58**(9) (2026). doi:10.1145/3795134.

[56] T. Sebbag, S. Quiniou, N. Stucky and E. Morin, AdminSet and AdminBERT: a Dataset and a Pre-trained Language Model to Explore the Unstructured Maze of French Administrative Documents, in: *Proceedings of the 31st International Conference on Computational Linguistics*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B.D. Eugenio and S. Schockaert, eds, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 392–406. https://aclanthology.org/2025.coling-main.27/.

[57] L. Siciliani, E. Tanzi, P. Basile and P. Lops, Automatic Generation of Common Procurement Vocabulary Codes, in: *Proceedings of the Ninth Italian Conference on Computational Linguistics (CLiC-it 2023)*, F. Boschetti, G.E. Lebani, B. Magnini and N. Novielli, eds, CEUR Workshop Proceedings, Venice, Italy, 2023, pp. 396–402. ISBN 979-12-550-0084-6. https://aclanthology.org/2023.clicit-1.47/.

[58] A. Soylu, O. Corcho, B. Elvesæter, C. Badenes-Olmedo, T. Blount, F. Yedro 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, *Semantic Web* **13**(2) (2022), 265–291. doi:10.3233/SW-210442.

[59] I. Stepanov and M. Shtopko, GLiNER multi-task: Generalist Lightweight Model for Various Information Extraction Tasks, 2024. https://arxiv.org/abs/2406.12925.

[60] K. Sun, Y. Hu, Y. Ma, R.Z. Zhou and Y. Zhu, Conflating point of interest (POI) data: A systematic review of matching methods, *Computers, Environment and Urban Systems* **103** (2023), 101977. doi:10.1016/j.compenvurbsys.2023.101977. https://www.sciencedirect.com/science/article/pii/S0198971523000406.

[61] S. Tedeschi and R. Navigli, MultiNERD: A Multilingual, Multi-Genre and Fine-Grained Dataset for Named Entity Recognition (and Disambiguation), in: *Findings of the Association for Computational Linguistics: NAACL 2022*, M. Carpuat, M.-C. de Marneffe and I.V. Meza Ruiz, eds, Association for Computational Linguistics, Seattle, United States, 2022, pp. 801–812. doi:10.18653/v1/2022.findings-naacl.60. https://aclanthology.org/2022.findings-naacl.60/.

[62] N. Tempelmeier and E. Demidova, Linking OpenStreetMap with Knowledge Graphs - Link Discovery for Schema-Agnostic Volunteered Geographic Information, *CoRR* **abs/2011.05841** (2020). https://arxiv.org/abs/2011.05841.

[63] N. Tempelmeier, S. Gottschalk and E. Demidova, GeoVectors: A Linked Open Corpus of OpenStreetMap Embeddings on World Scale, *CoRR* **abs/2108.13092** (2021). https://arxiv.org/abs/2108.13092.

[64] L. Tunstall, N. Reimers, U.E.S. Jo, L. Bates, D. Korat, M. Wasserblat and O. Pereg, Efficient Few-Shot Learning Without Prompts, 2022. https://arxiv.org/abs/2209.11055.

[65] A. Vidali, N. Jean and G.L. Pera, Unlocking NACE Classification Embeddings with OpenAI for Enhanced Analysis and Processing, 2024. https://arxiv.org/abs/2409.11524.

[66] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder and F. Wei, Improving Text Embeddings with Large Language Models, 2024. https://arxiv.org/abs/2401.00368.

[67] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder and F. Wei, Multilingual E5 Text Embeddings: A Technical Report, 2024. https://arxiv.org/abs/2402.05672.

[68] W. Wang, H. Bao, S. Huang, L. Dong and F. Wei, MiniLMv2: Multi-Head Self-Attention Relation Distillation for Compressing Pretrained Transformers, *CoRR* **abs/2012.15828** (2020). https://arxiv.org/abs/2012.15828.

[69] X. Wang, Y. Shen, J. Cai, T. Wang, X. Wang, P. Xie, F. Huang, W. Lu, Y. Zhuang, K. Tu, W. Lu and Y. Jiang, DAMO-NLP at SemEval-2022 Task 11: A Knowledge-based System for Multilingual Named Entity Recognition, in: *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh and S. Ratan, eds, Association for Computational Linguistics, Seattle, United States, 2022, pp. 1457–1468. doi:10.18653/v1/2022.semeval-1.200. https://aclanthology.org/2022.semeval-1.200/.

[70] A. Williams, N. Nangia and S.R. Bowman, A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference, *CoRR* **abs/1704.05426** (2017). http://arxiv.org/abs/1704.05426.

[71] Y. Xiao, J. Zou and Q. Yang, Advancing Few-Shot Named Entity Recognition with Large Language Model, *Applied Sciences* **15**(7) (2025), 3838. doi:10.3390/app15073838. https://www.mdpi.com/2076-3417/15/7/3838.

[72] S. Yu, J. He, V. Gutiérrez-Basulto and J.Z. Pan, Instances and Labels: Hierarchy-aware Joint Supervised Contrastive Learning for Hierarchical Multi-Label Text Classification, 2024. https://arxiv.org/abs/2310.05128.

[73] A. Zangari, M. Marcuzzo, M. Rizzo, L. Giudice, A. Albarelli and A. Gasparetto, Hierarchical Text Classification and Its Foundations: A Review of Current Research, *Electronics* **13**(7) (2024), 1199. doi:10.3390/electronics13071199.

[74] U. Zaratiana, N. Tomeh, P. Holat and T. Charnois, GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer, 2023. https://arxiv.org/abs/2311.08526.

[75] U. Zaratiana, G. Pasternak, O. Boyd, G. Hurn-Maloney and A. Lewis, GLiNER2: An Efficient Multi-Task Information Extraction System with Schema-Driven Interface, 2025. https://arxiv.org/abs/2507.18546.

[76] Q. Zhang, Q. Su, W. Zhu and P. Yachun, HierPrompt: Zero-Shot Hierarchical Text Classification with LLM-Enhanced Prototypes, in: *Findings of the Association for Computational Linguistics: EMNLP 2025*, C. Christodoulopoulos, T. Chakraborty, C. Rose and V. Peng, eds, Association for Computational Linguistics, Suzhou, China, 2025, pp. 3846–3859. ISBN 979-8-89176-335-7. doi:10.18653/v1/2025.findings-emnlp.207. https://aclanthology.org/2025.findings-emnlp.207/.

[77] Y. Zhang, R. Yang, X. Xu, R. Li, J. Xiao, J. Shen and J. Han, TELEClass: Taxonomy Enrichment and LLM-Enhanced Hierarchical Text Classification with Minimal Supervision, 2025. https://arxiv.org/abs/2403.00165.

[78] Z. Zhang, W. You, T. Wu, X. Wang, J. Li and M. Zhang, A Survey of Generative Information Extraction, in: *Proceedings of the 31st International Conference on Computational Linguistics*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B.D. Eugenio and S. Schockaert, eds, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 4840–4870. https://aclanthology.org/2025.coling-main.324/.

[79] Y. Zhao, C. Dai, D. Niyato, C.F. Tan, K. Xiang, Y. Wang, Z. Yeo, D.T.Z. Loong, J.L. Zhaozhi and E.H.Z. HO, A Graph-Retrieval-Augmented Generation Framework Enhances Decision-Making in the Circular Economy, 2025. https://arxiv.org/abs/2506.04252.

[80] Y. Zhao, A. Zhang, R. Xie, K. Liu and X. Wang, Connecting Embeddings for Knowledge Graph Entity Typing, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter and J. Tetreault, eds, Association for Computational Linguistics, Online, 2020, pp. 6419–6428. doi:10.18653/v1/2020.acl-main.572. https://aclanthology.org/2020.acl-main.572/.

[81] Y. Zhong, Z. Zhang, W. Zhang and J. Zhu, BERT-KG: A Short Text Classification Model Based on Knowledge Graph and Deep Semantics, in: *Natural Language Processing and Chinese Computing*, L. Wang, Y. Feng, Y. Hong and R. He, eds, Springer International Publishing, Cham, 2021, pp. 721–733. ISBN 978-3-030-88480-2.

[82] R. Zhu, C. Shimizu, S. Stephen, C.K. Fisher, T. Thelen, K. Currier, K. Janowicz, P. Hitzler, M. Schildhauer, W. Li, D. Rehberger, A. Barua, A. Christou, L. Cai, A. Dalal, A. D'Onofrio, A. Eells, M. Faulk, Z. Liu, G. Mai, M.S. Mahdavinejad, B. Mecum, S.S. Norouzi, M. Shi, Y. Tian, S. Wang, Z. Wang and J. Zalewski, The KnowWhereGraph: A Large-Scale Geo-Knowledge Graph for Interdisciplinary Knowledge Discovery and Geo-Enrichment, 2025. https://arxiv.org/abs/2502.13874.

[83] Álvaro Fontecha del Ser, Dataset for Organization Names NLP: Multilingual Public Sector Entity Classification, 2024, Zenodo repository. doi:10.5281/zenodo.15312407.

[84] 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.

[85] 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

6.1. Prompts

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
```

Listing 3: LLM prompt used for keyword name generation.

```
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.
```

Listing 4: LLM prompt used for keyword name generation.

```
Given an organization name, decide which of the following classes it belongs to: [{class_list}] or none.
An organization can belong to multiple classes or none.
Here are examples from different countries:
{few_shot_block}
Respond with ONLY a JSON object like: {{{example_pairs}}}
using 1 for yes and 0 for no, for each class in [{class_list}]. Nothing else.
```

6.2. Experimental Results

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

Table 7: The table shows experimental metrics for rule-based techniques across all domains. Experiment IDs reflect: domain (edu=education) - technique (r=rules) - method - configuration (increasing integer). Experimental setup with **highest F1 per domain** in bold. Variations in structure, preprocessing, and tokenization are presented to identify optimal configurations.

ID	Technique	Method	Model	Structure	Domain	Accuracy	Recall	F1
med-n-0-0	nli	0_shot	roberta-large	3-class	medical	0.974863	0.644098	0.440823
med-n-0-1	nli	0_shot	bge-m3	3-class	medical	0.522354	0.650609	0.265277
med-n-0-2	**nli**	**0_shot**	**mDeBerta**	**3-class**	**medical**	**0.980167**	**0.593545**	**0.476352**
med-n-0-3	nli	0_shot	MiniLM	3-class	medical	0.522452	0.625190	0.123156
med-n-0-4	nli	0_shot	roberta-large	nested-class	medical	0.253989	0.809255	0.054539
med-n-0-5	nli	0_shot	bge-m3	nested-class	medical	0.045716	0.565684	0.041871
med-n-0-6	nli	0_shot	mDeBerta	nested-class	medical	0.149934	0.777962	0.053806
med-n-0-7	nli	0_shot	MiniLM	nested-class	medical	0.698137	0.739025	0.142762
med-n-few-0	nli	few_shot	DeepSeek-R1:7B	nested-class	medical	0.951589	0.684133	0.362167
adm-n-0-0	nli	0_shot	roberta-large	2-class	administrative	0.896647	0.826847	0.767278
adm-n-0-1	**nli**	**0_shot**	**bge-m3**	**2-class**	**administrative**	**0.909281**	**0.797062**	**0.772669**
adm-n-0-2	nli	0_shot	mDeBerta	2-class	administrative	0.800449	0.831855	0.674289
adm-n-0-3	nli	0_shot	MiniLM	2-class	administrative	0.635634	0.695312	0.527614
adm-n-few-0	nli	few_shot	DeepSeek-R1:7B	2-class	administrative	0.920000	0.500000	0.479167
edu-n-0-0	nli	0_shot	roberta-large	3-multiclass	education	0.833607	0.416544	0.473872
edu-n-0-1	**nli**	**0_shot**	**bge-m3**	**3-multiclass**	**education**	**0.831991**	**0.537146**	**0.497835**
edu-n-0-2	nli	0_shot	mDeBerta	3-multiclass	education	0.827772	0.419011	0.447585
edu-n-0-3	nli	0_shot	MiniLM	3-multiclass	education	0.711922	0.366130	0.241515
edu-n-few-0	nli	few_shot	DeepSeek-R1:7B	3-multiclass	education	0.788845	0.000100	0.000199

Table 8: The table shows experimental metrics for Natural Language Inference and Generative Few Shot Classification across all domains. Experiment IDs reflect: domain (edu=education) - technique (n=natural language inference) - method - configuration (increasing integer). Experimental setup with **highest F1 per domain** in bold. Different parameters like model choice and configuration are explored to determine optimal configurations.

ID	Domain	Technique	Method	Model	Structure	Prototypes	Distance	Accuracy	Recall	F1
med-e-similarity-0	**medical**	**embedding**	**similarity**	**multilingual-e5**	**3-class**	**0_shot**	**0.1**	**0.984168**	**0.304990**	**0.348230**
med-e-similarity-1	medical	embedding	similarity	multilingual-e5	3-class	1_shot	0.1	0.921418	0.408667	0.105151
med-e-similarity-2	medical	embedding	similarity	multilingual-e5	3-class	few_shot	0.05	0.978404	0.407954	0.220443
med-e-similarity-3	medical	embedding	similarity	qwen	3-class	0_shot	0.5	0.971403	0.530709	0.149797
med-e-similarity-4	medical	embedding	similarity	qwen	3-class	1_shot	0.5	0.954201	0.328241	0.185169
med-e-similarity-5	medical	embedding	similarity	qwen	3-class	few_shot	0.3	0.977911	0.509904	0.256171
med-e-similarity-6	medical	embedding	similarity	mistral	3-class	0_shot	0.2	0.980090	0.241478	0.194220
med-e-similarity-7	medical	embedding	similarity	mistral	3-class	1_shot	0.2	0.975926	0.113083	0.136733
med-e-similarity-8	medical	embedding	similarity	mistral	3-class	few_shot	0.15	0.975729	0.486291	0.253893
med-e-similarity-9	medical	embedding	similarity	e5-small	3-class	0_shot	0.1	0.981571	0.117828	0.173329
med-e-similarity-10	medical	embedding	similarity	e5-small	3-class	1_shot	0.1	0.953576	0.084934	0.050910
med-e-similarity-11	medical	embedding	similarity	e5-small	3-class	few_shot	0.1	0.816069	0.522673	0.073753
med-e-similarity-12	medical	embedding	similarity	finetuned-me5	3-class	0_shot	0.01	0.961911	0.469388	0.026048
med-e-similarity-18	medical	embedding	similarity	finetuned-me5	3-class	1_shot	0.05	0.979487	0.427971	0.343107
med-e-similarity-19	medical	embedding	similarity	finetuned-me5	3-class	few_shot	0.01	0.979553	0.559112	0.347172
adm-e-similarity-0	administrative	embedding	similarity	multilingual-e5	2-class	0_shot	0.15	0.770938	0.804084	0.639797
adm-e-similarity-1	administrative	embedding	similarity	multilingual-e5	2-class	1_shot	0.1	0.771179	0.574040	0.543606
adm-e-similarity-2	administrative	embedding	similarity	multilingual-e5	2-class	few_shot	0.05	0.919950	0.607796	0.654192
adm-e-similarity-3	administrative	embedding	similarity	qwen	2-class	0_shot	0.7	0.794364	0.766306	0.644378
adm-e-similarity-4	administrative	embedding	similarity	qwen	2-class	1_shot	0.5	0.891397	0.502398	0.486496
adm-e-similarity-5	administrative	embedding	similarity	qwen	2-class	few_shot	0.3	0.925626	0.697743	0.741955
adm-e-similarity-6	administrative	embedding	similarity	mistral	2-class	0_shot	0.3	0.919572	0.705013	0.737436
adm-e-similarity-7	administrative	embedding	similarity	mistral	2-class	1_shot	0.3	0.772348	0.713248	0.609657
adm-e-similarity-8	administrative	embedding	similarity	mistral	2-class	few_shot	0.15	0.935417	0.758484	0.793699
adm-e-similarity-9	administrative	embedding	similarity	e5-small	2-class	0_shot	0.1	0.900600	0.500223	0.474728
adm-e-similarity-10	administrative	embedding	similarity	e5-small	2-class	1_shot	0.1	0.898979	0.510050	0.496698
adm-e-similarity-11	administrative	embedding	similarity	e5-small	2-class	few_shot	0.05	0.911513	0.554500	0.574878
adm-e-similarity-12	administrative	embedding	similarity	finetuned-me5	2-class	0_shot	0.1	0.962284	0.941176	0.831819
adm-e-similarity-18	**administrative**	**embedding**	**similarity**	**finetuned-me5**	**2-class**	**1_shot**	**0.05**	**0.963270**	**0.936754**	**0.834844**
adm-e-similarity-19	administrative	embedding	similarity	finetuned-me5	2-class	few_shot	0.01	0.676945	0.890314	0.353266
edu-e-similarity-0	education	embedding	similarity	multilingual-e5	3-multiclass	0_shot	0.15	0.871706	0.614236	0.505434
edu-e-similarity-1	education	embedding	similarity	multilingual-e5	3-multiclass	1_shot	0.1	0.752082	0.434147	0.248134
edu-e-similarity-2	education	embedding	similarity	multilingual-e5	3-multiclass	few_shot	0.1	0.575447	0.679848	0.262027
edu-e-similarity-3	education	embedding	similarity	qwen	3-multiclass	0_shot	0.7	0.870782	0.581780	0.485113
edu-e-similarity-4	education	embedding	similarity	qwen	3-multiclass	1_shot	0.5	0.869994	0.315616	0.262263
edu-e-similarity-5	education	embedding	similarity	qwen	3-multiclass	few_shot	0.3	0.906132	0.228346	0.340588
edu-e-similarity-6	education	embedding	similarity	mistral	3-multiclass	0_shot	0.3	0.908266	0.350068	0.451101
edu-e-similarity-7	education	embedding	similarity	mistral	3-multiclass	1_shot	0.3	0.611930	0.565826	0.316307
edu-e-similarity-8	education	embedding	similarity	mistral	3-multiclass	few_shot	0.15	0.915368	0.402072	0.491348
edu-e-similarity-9	education	embedding	similarity	e5-small	3-multiclass	0_shot	0.15	0.671910	0.409485	0.213756
edu-e-similarity-10	education	embedding	similarity	e5-small	3-multiclass	1_shot	0.2	0.524017	0.533857	0.189732
edu-e-similarity-11	education	embedding	similarity	e5-small	3-multiclass	few_shot	0.1	0.782214	0.407158	0.291436
edu-e-similarity-12	education	embedding	similarity	finetuned-me5	3-multiclass	0_shot	0.2	0.953300	0.851818	0.795884
edu-e-similarity-18	education	embedding	similarity	finetuned-me5	3-multiclass	1_shot	0.3	0.951745	0.866248	0.793478
edu-e-similarity-19	**education**	**embedding**	**similarity**	**finetuned-me5**	**3-multiclass**	**few_shot**	**0.01**	**0.957772**	**0.764926**	**0.799415**

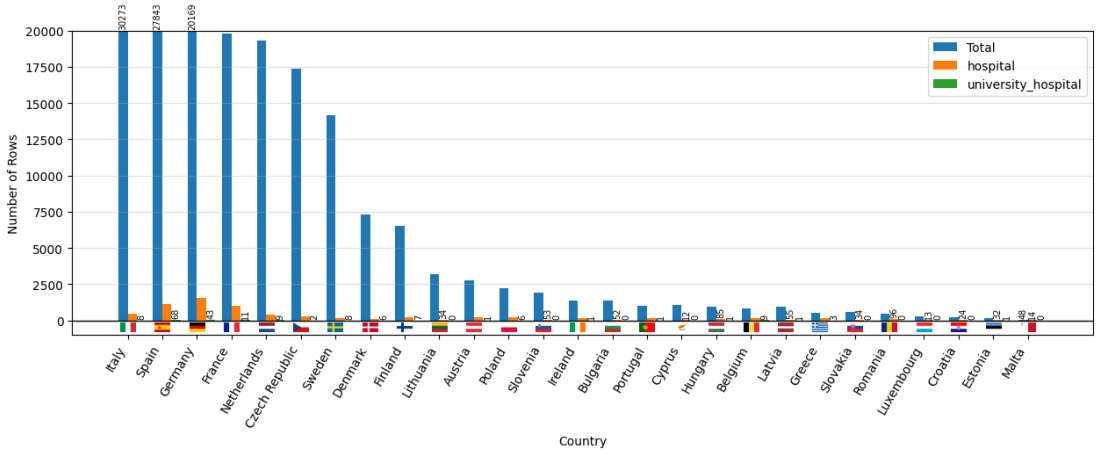
Table 9: The table shows experimental metrics for Embedding-based classification using distance to class protoypes across all domains. Experiment IDs reflect: domain (edu=education) - technique (n=natural language inference) - method - configuration (increasing integer). Experimental setup with **highest F1 per domain** in bold. Different parameters like model choice and number of prototypes are explored to determine optimal configurations. Distance parameter is optimized on a validation set.

ID	Domain	Technique	Method	Model	Classifier	Structure	C	solver	penalty	Accuracy	Recall	F1
med-e-classifier-0	medical	embedding	classifier	multilingual-e5	logreg	nested-class	10	liblinear	l1	0.979479	0.723484	0.562499
med-e-classifier-2	medical	embedding	classifier	qwen	logreg	nested-class	10	liblinear	l1	0.985538	0.783560	0.629873
med-e-classifier-4	medical	embedding	classifier	mistral	logreg	nested-class	10	liblinear	l1	0.986950	0.826420	0.643318
med-e-classifier-6	medical	embedding	classifier	e5-small	logreg	nested-class	1	lbfgs	l2	0.902455	0.734472	0.270243
med-e-classifier-8	medical	embedding	classifier	finetuned-me5	logreg	nested-class	10	liblinear	l1	0.979093	0.818082	0.513329
adm-e-classifier-0	administrative	embedding	classifier	multilingual-e5	logreg	2-class	10	liblinear	l1	0.938874	0.941253	0.862211
adm-e-classifier-2	administrative	embedding	classifier	qwen	logreg	2-class	10	liblinear	l1	0.944517	0.925343	0.868487
adm-e-classifier-4	administrative	embedding	classifier	mistral	logreg	2-class	10	liblinear	l1	0.950837	0.942878	0.883531
adm-e-classifier-5	administrative	embedding	classifier	e5-small	logreg	2-class	10	liblinear	l1	0.876291	0.874330	0.758679
adm-e-classifier-8	administrative	embedding	classifier	finetuned-me5	logreg	2-class	0.01	liblinear	l1	0.953211	0.960845	0.888877
edu-e-classifier-0	education	embedding	classifier	multilingual-e5	logreg	3-multiclass	10	liblinear	l1	0.863709	0.912029	0.714403
edu-e-classifier-2	education	embedding	classifier	qwen	logreg	3-multiclass	10	liblinear	l1	0.885593	0.891451	0.748070
edu-e-classifier-3	education	embedding	classifier	e5-small	logreg	3-multiclass	10	liblinear	l1	0.740400	0.821275	0.537592
edu-e-classifier-4	education	embedding	classifier	mistral	logreg	3-multiclass	10	liblinear	l1	0.898863	0.918018	0.775537
edu-e-classifier-8	education	embedding	classifier	finetuned-me5	logreg	3-multiclass	10	lbfgs	l2	0.907380	0.953533	0.791671
ID	Domain	Technique	Method	Model	Classifier	Structure	C	Kernel	Gamma	Accuracy	Recall	F1
med-e-classifier-1	medical	embedding	classifier	multilingual-e5	svm	nested-class	10	rbf	scale	0.987554	0.665972	0.624416
med-e-classifier-3	medical	embedding	classifier	qwen	svm	nested-class	1	rbf	scale	0.990139	0.765587	0.666347
med-e-classifier-5	**medical**	**embedding**	**classifier**	**mistral**	**svm**	**nested-class**	**10**	**rbf**	**scale**	**0.991947**	**0.672002**	**0.682510**
med-e-classifier-7	medical	embedding	classifier	e5-small	svm	nested-class	10	rbf	scale	0.979095	0.620456	0.537543
med-e-classifier-9	medical	embedding	classifier	finetuned-me5	svm	nested-class	10	rbf	scale	0.984813	0.827731	0.556991
adm-e-classifier-1	administrative	embedding	classifier	multilingual-e5	svm	2-class	10	rbf	scale	0.950992	0.952657	0.885668
adm-e-classifier-3	administrative	embedding	classifier	qwen	svm	2-class	1	rbf	scale	0.955703	0.943593	0.892814
adm-e-classifier-6	administrative	embedding	classifier	e5-small	svm	2-class	10	rbf	scale	0.942260	0.918194	0.862910
adm-e-classifier-7	**administrative**	**embedding**	**classifier**	**mistral**	**svm**	**2-class**	**10**	**rbf**	**scale**	**0.960816**	**0.944597**	**0.900882**
adm-e-classifier-9	administrative	embedding	classifier	finetuned-me5	svm	2-class	10	rbf	scale	0.959062	0.959172	0.899847
edu-e-classifier-1	education	embedding	classifier	multilingual-e5	svm	3-multiclass	10	rbf	scale	0.895371	0.936686	0.771529
edu-e-classifier-5	education	embedding	classifier	mistral	svm	3-multiclass	10	rbf	scale	0.924284	0.899582	0.823174
edu-e-classifier-6	**education**	**embedding**	**classifier**	**qwen**	**svm**	**3-multiclass**	**10**	**rbf**	**scale**	**0.929943**	**0.820452**	**0.824113**
edu-e-classifier-7	education	embedding	classifier	e5-small	svm	3-multiclass	10	rbf	scale	0.886457	0.872244	0.741813
edu-e-classifier-9	education	embedding	classifier	finetuned-me5	svm	3-multiclass	10	rbf	scale	0.911311	0.955301	0.799627

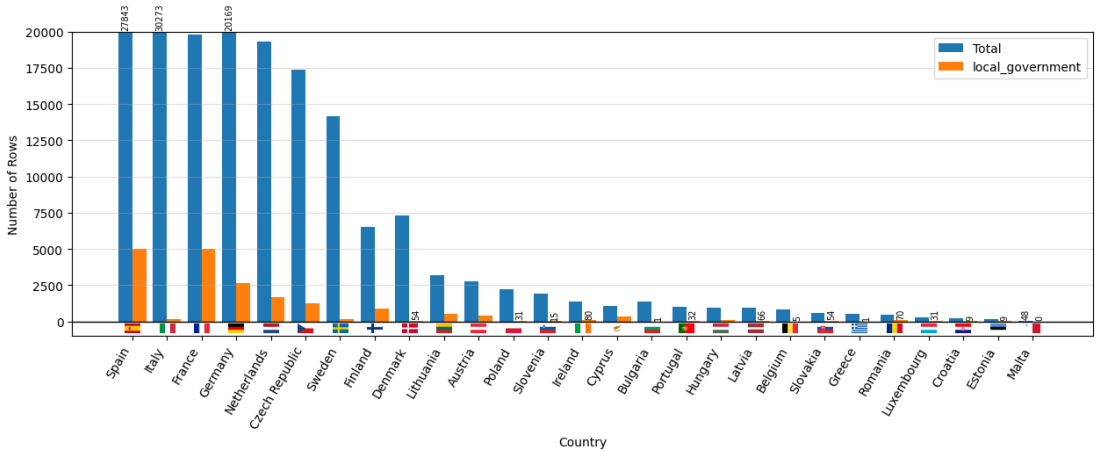
Table 10: The table shows experimental metrics for Embedding-based classification using a classifier head across all domains. Top half of the table corresponds to a logistic regressor head and top bottom to a support vector machine. Experiment IDs reflect: domain (edu=education) - technique (n=natural language inference) - method - configuration (increasing integer). Experimental setup with **highest F1 per domain** in bold. Different parameters like model choice and classifier head algorithm are explored to determine optimal configurations. Classifier parameters are optimized on a validation set.

	Char Len		Tok Count				Diversity		Entropy (H)						
Country	μ	σ	μ	σ	TTR	Hapax Ratio	Suff.	Pref.	Tok	Bi	Tri	CR	Uniq.	N	Ent.
France	23.391	14.670	3.035	2.058	0.330	0.813	0.056	0.059	10.065	11.840	12.718	0.344	19844	60066	19789
Germany	26.834	15.646	2.751	1.697	0.390	0.767	0.057	0.065	12.293	14.280	13.755	0.354	21637	55487	20169
Estonia	20.965	8.003	2.472	0.763	0.545	0.820	0.313	0.435	6.556	7.677	6.190	0.393	194	356	144
Latvia	24.252	12.420	3.111	1.678	0.387	0.732	0.198	0.253	8.771	10.124	9.747	0.324	1110	2868	922
Czech Republic	20.425	11.632	3.014	1.637	0.256	0.732	0.079	0.078	10.018	11.584	11.951	0.309	13422	52447	17399
Slovakia	17.150	10.875	2.501	1.328	0.272	0.716	0.178	0.203	6.106	6.916	8.156	0.232	394	1448	579
Slovenia	25.922	14.123	3.536	1.828	0.371	0.755	0.151	0.170	9.053	10.541	11.176	0.307	2486	6701	1895
Romania	32.374	14.727	4.452	1.984	0.385	0.778	0.230	0.263	7.970	9.241	9.401	0.332	810	2106	473
Bulgaria	25.851	15.675	3.678	1.959	0.446	0.769	0.243	0.273	9.781	11.055	10.891	0.257	2275	5102	1387
Croatia	23.439	12.378	3.184	1.494	0.566	0.771	0.373	0.427	7.852	8.328	7.778	0.426	353	624	196
Cyprus	22.449	9.801	2.720	1.044	0.470	0.831	0.258	0.319	8.566	9.433	9.599	0.270	1357	2886	1061
Malta	23.917	11.026	3.396	1.642	0.767	0.864	0.653	0.673	6.713	6.793	6.129	0.583	125	163	48
Ireland	20.354	10.537	2.947	1.400	0.384	0.754	0.183	0.203	8.967	10.634	10.253	0.353	1564	4070	1381
Hungary	27.293	13.988	3.244	1.619	0.477	0.801	0.268	0.264	9.145	10.247	9.910	0.348	1420	2978	918
Spain	27.392	15.391	4.108	2.362	0.183	0.642	0.037	0.038	9.888	13.062	14.333	0.327	20890	114365	27842
Belgium	23.576	14.119	2.867	1.887	0.613	0.838	0.310	0.375	9.690	10.286	9.739	0.444	1421	2319	809
Luxembourg	27.664	15.671	3.538	2.061	0.501	0.799	0.367	0.408	7.646	8.614	8.378	0.397	448	895	253
Finland	18.245	8.547	2.001	0.840	0.495	0.792	0.106	0.132	11.102	12.567	10.560	0.333	6466	13067	6530
Sweden	17.888	8.489	2.005	1.025	0.418	0.729	0.071	0.087	11.693	13.234	11.558	0.346	11857	28389	14156
Denmark	17.679	8.771	2.226	1.104	0.448	0.774	0.121	0.141	11.150	12.786	11.599	0.363	7294	16284	7314
Poland	33.717	19.623	4.421	2.545	0.339	0.744	0.146	0.164	9.529	11.025	11.366	0.303	3290	9717	2198
Lithuania	28.335	13.637	3.411	1.550	0.286	0.759	0.085	0.119	8.828	10.297	10.293	0.249	3137	10966	3215
Italy	25.973	15.189	3.695	2.157	0.213	0.664	0.031	0.033	11.172	14.205	14.832	0.359	23860	111863	30273
Austria	22.520	14.161	2.498	1.471	0.475	0.813	0.163	0.189	10.186	11.178	10.121	0.341	3263	6865	2748
Greece	28.465	12.189	3.509	1.430	0.492	0.799	0.244	0.336	8.624	9.547	9.527	0.256	943	1916	546
Portugal	28.836	13.446	4.453	2.174	0.304	0.710	0.164	0.201	8.246	10.130	10.265	0.314	1347	4426	994
Netherlands	22.346	11.944	2.817	1.371	0.321	0.736	0.069	0.073	11.019	12.947	13.223	0.351	17419	54311	19280

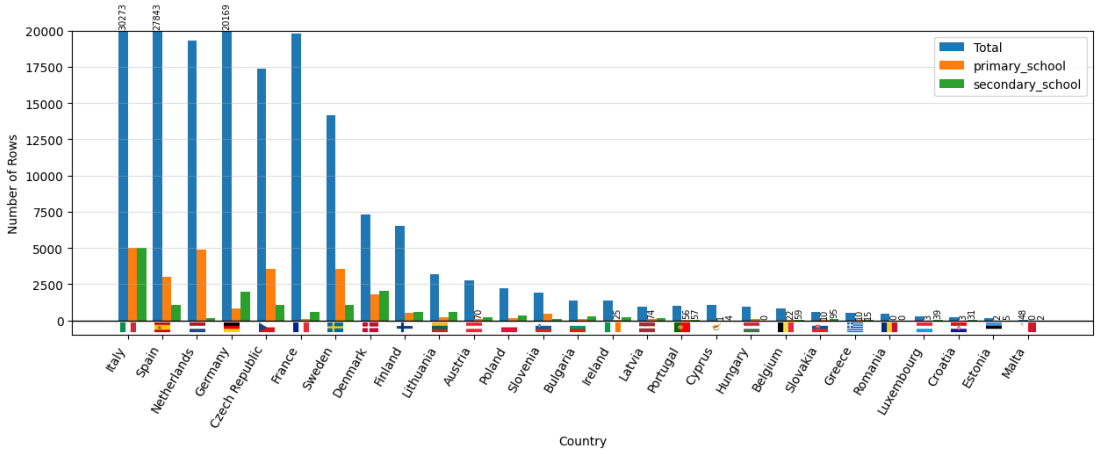
Table 11: Table shows lexical statistics for the corpus of entities per country. μ and σ denote mean and standard deviation; TTR: Type-Token Ratio; Suff./Pref.: Suffix/Prefix Diversity; Tok/Bi/Tri: Token, Bigram, and Trigram Entropy (H); CR: Compression Ratio; Uniq.: Unique Tokens; N: Total Tokens; Ent.: Total Entities.



(a) Medical domain

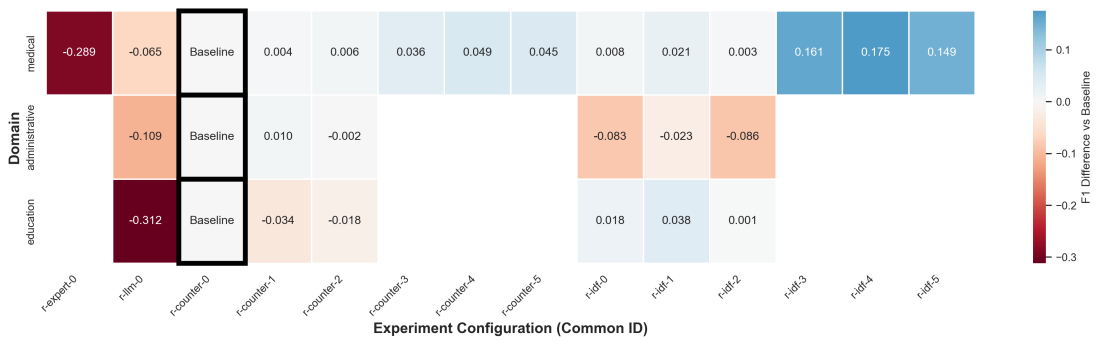


(b) Administrative domain

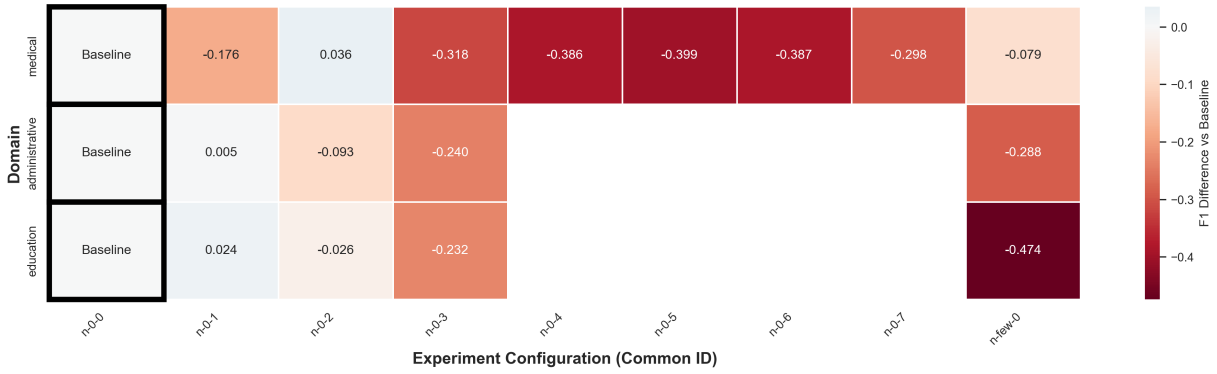


(c) Educational domain

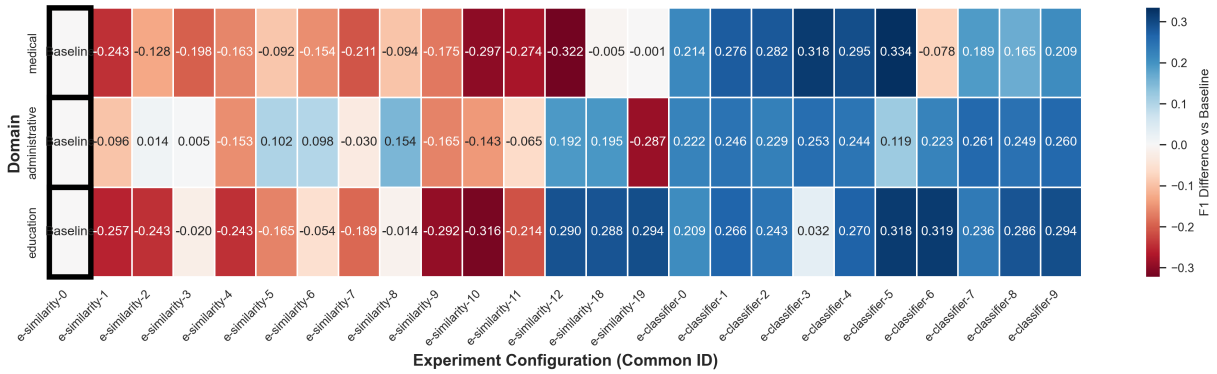
Figure 17. This plot shows the volume of entities per domain in our dataset. Each subfigure shows the distribution of entities by country and class in the Medical, Administrative, and Educational classes respectively. As mentioned, even with high data vlume overall, when partitioning our data into classes (specially minority ones like university hospitals) and most importantly, countries it can be seen that data scarcity is very much an issue. Fragmentation of data is unavoidable due to the intrinsic differences between regions.



(a) Rule-based family of experiments. Experiment configurations and complete scores can be read in table 5. Selected baseline is Counter keyword selection using without preprocessing.



(b) Natural Language Inference family of experiments. Experiment configurations and complete scores can be read in table 8. Selected baseline is zero-shot classification using RoBERTa large.



(c) Embedding-based family of experiments. Experiment configurations and complete scores can be read in tables **??**. Selected baseline is Multilingual E5 Closest cosine Similarity to label Classification.

Figure 18. These heatmaps show $\Delta F1$ Scores against selected domain baselines for all three families.

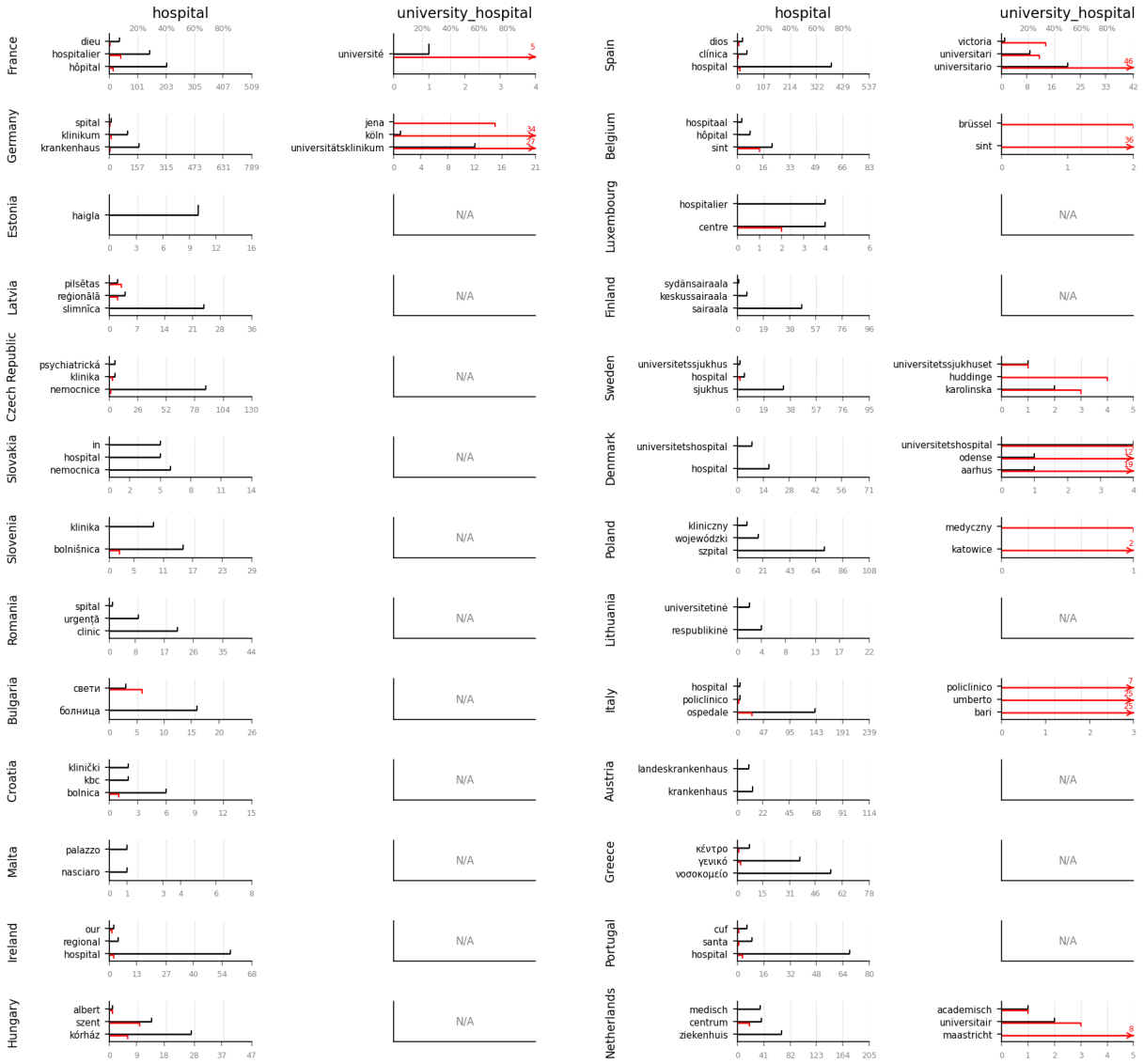


Figure 19. This plot presents the TF-IDF keyword selection and their coverage of true positive instances in the medical domain, broken down by country and class. It is a small-multiples plot where each cell displays the keywords selected for a specific partition of the domain: rows correspond to countries and columns to classes. To keep the figure legible, countries are split into two halves and displayed side by side. Within each cell, keywords are listed on the y-axis and the number of instances on the x-axis. The black line shows how many true positive instances are covered by each keyword, while the red line shows the number of false positives it induces. Crucially, the x-axis is scaled to the total number of true instances for that country–class pair, so lines can be read as implicit proportions, ie. coverage of the true positives for black, while false positives over true instances for red. Note that keyword coverages are not a disjoint sets, their sum going over the number of true instances; eg. an organization can be covered by multiple keywords. A red line extending beyond is indicated with an arrow and its number, indicating more false positives than there are true instances. This happens in partitions with few true instances and limited training data. The algorithm should avoid this cases if not for keyword selection being performed on a training set, while the plot reflects test set behavior. To read the plot, pick a country row and scan the keywords on the y-axis. For each one, the black line tells you how many true instances it covers. The red line tells you how much noise it brings in.

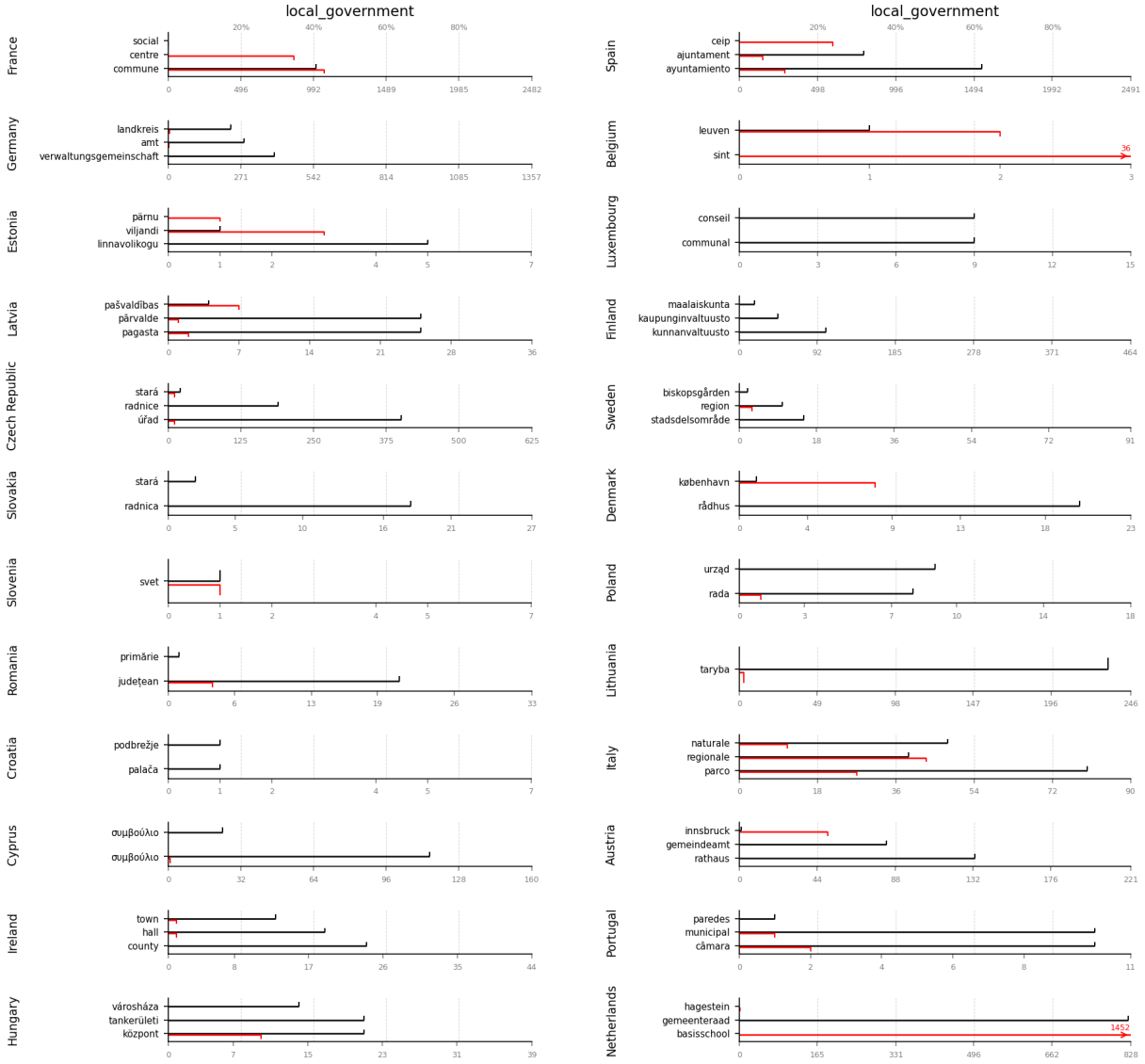


Figure 20. This plot presents the TF-IDF keyword selection and their coverage of true positive instances in the administrative domain, broken down by country and class. It is a small-multiples plot where each cell displays the keywords selected for a specific partition of the domain: rows correspond to countries and columns to classes. To keep the figure legible, countries are split into two halves and displayed side by side. Within each cell, keywords are listed on the y-axis and the number of instances on the x-axis. The black line shows how many true positive instances are covered by each keyword, while the red line shows the number of false positives it induces. Crucially, the x-axis is scaled to the total number of true instances for that country–class pair, so lines can be read as implicit proportions, ie. coverage of the true positives for black, while false positives over true instances for red. Note that keyword coverages are not a disjoint sets, their sum going over the number of true instances; eg. an organization can be covered by multiple keywords. A red line extending beyond is indicated with an arrow and its number, indicating more false positives than there are true instances. This happens in partitions with few true instances and limited training data. The algorithm should avoid this cases if not for keyword selection being performed on a training set, while the plot reflects test set behavior. To read the plot, pick a country row and scan the keywords on the y-axis. For each one, the black line tells you how many true instances it covers. The red line tells you how much noise it brings in.

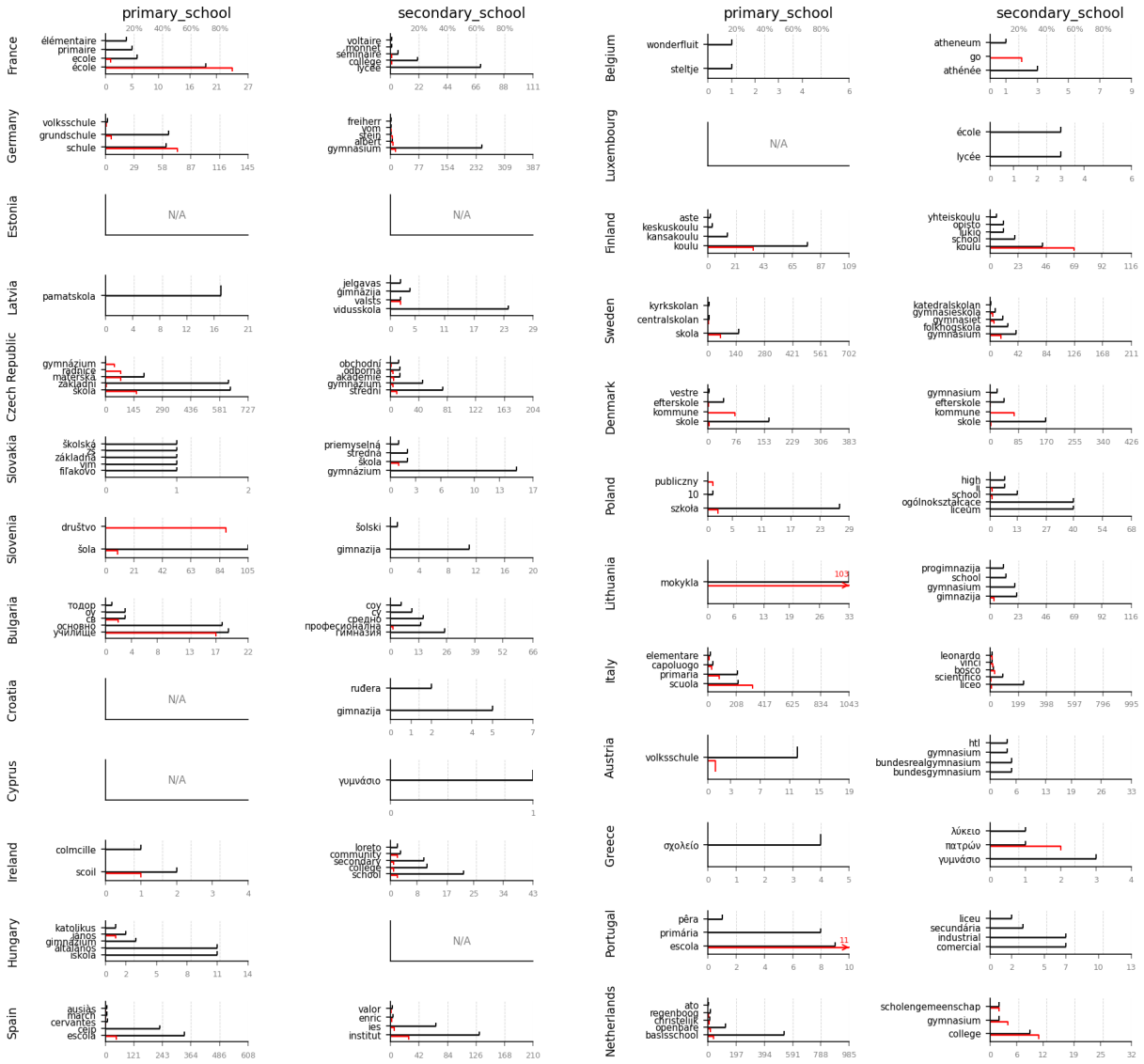


Figure 21. This plot presents the TF-IDF keyword selection and their coverage of true positive instances in the education domain, broken down by country and class. It is a small-multiples plot where each cell displays the keywords selected for a specific partition of the domain: rows correspond to countries and columns to classes. To keep the figure legible, countries are split into two halves and displayed side by side. Within each cell, keywords are listed on the y-axis and the number of instances on the x-axis. The black line shows how many true positive instances are covered by each keyword, while the red line shows the number of false positives it induces. Crucially, the x-axis is scaled to the total number of true instances for that country–class pair, so lines can be read as implicit proportions, ie. coverage of the true positives for black, while false positives over true instances for red. Note that keyword coverages are not a disjoint sets, their sum going over the number of true instances; eg. an organization can be covered by multiple keywords. A red line extending beyond is indicated with an arrow and its number, indicating more false positives than there are true instances. This happens in partitions with few true instances and limited training data. The algorithm should avoid this cases if not for keyword selection being performed on a training set, while the plot reflects test set behavior. To read the plot, pick a country row and scan the keywords on the y-axis. For each one, the black line tells you how many true instances it covers. The red line tells you how much noise it brings in.