LA^{2}I^{2}F at SemEval-2025 Task 5: Reasoning in Embedding Space – Fusing Analogical and Ontology-based Reasoning for Document Subject Tagging

Andrea Salfinger, Luca Zaccagna, Francesca Incitti,
Gianluca G. M. De Nardi, Lorenzo Dal Fabbro, Lauro Snidaro
University of Udine, Italy
{andrea.salfinger, francesca.incitti, lauro.snidaro}@uniud.it
{zaccagna.luca, denardi.gianlucagiuseppemaria, dalfabbro.lorenzo}@spes.uniud.it

Abstract

The *LLMs4Subjects* shared task invited system contributions that leverage a technical library's tagged document corpus to learn document subject tagging, i.e., proposing adequate subjects given a document's title and abstract. To address the imbalance of this training corpus, team LA^{2}I^{2}F devised a semantic retrieval-based system fusing the results of ontological and analogical reasoning in embedding vector space. Our results outperformed a naive baseline of prompting a llama 3.1-based model, whilst being computationally more efficient and competitive with the state of the art.

1 Introduction

Motivation. To assist human librarians in cataloging documents with suitable subject tags from modern libraries' comprehensive subject indices (Turvey and Letarte, 2014), the SemEval-2025 shared task "LLMs4Subjects" (D'Souza et al., 2025) explores the feasibility of Large Language Model (LLM)-based automated subject tagging. The systems should exploit a human-tagged document corpus from the Leibniz University's Technical Library (TIBKAT) to learn subject tagging (Golub, 2021): Given a document's title and abstract as input, the system should propose the top-k best-matching subjects from the ontology Gemeinsame Normdatei[1] (GND) for both English (EN) and German (DE) documents.

Example Document 1

Title: Gender and creative labour
Abstract: Introduction – Sexism, segregation and gender roles – Flexibility and informality – Image-making and representation – Boundary-crossing – Notes on contributors

GND Subject Labels: Geschlechterforschung (gender studies); Künstler (artist); Kulturbetrieb (cultural sector)

[1] https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html

Example Document 1 from the training corpus illustrates the problem, showing the "ground truth" GND subject labels assigned by the librarians (original DE labels and their translations shown).

Approach. This paper describes the solution developed by the team of the *Laboratory of Applied Artificial Intelligence and Information Fusion* (LA^{2}I^{2}F), from the University of Udine, Italy, which focused on exploring novel ways to address the imbalance of the provided datasets, since highly specific subject terms typically are assigned to few documents only. Our *semantic retrieval-based information fusion system* combines complementary reasoning strategies in embedding vector space (Incitti et al., 2023): (i) an *analogical reasoning* branch proposing subject tags from the most similar documents in the training data (conceivable as case-based reasoning in vector space), and an (ii) *ontological reasoning* branch proposing subject tags from the GND based on their semantic similarity with the current document's title and abstract.

Results & Insights. While analogical clearly outperformed ontological reasoning, the *fusion* of both strategies yielded the best performance in our ablation study, surpassing a "naive" baseline of prompting an unmodified `llama 3.1:8B` model (Dubey et al., 2024) (thus agnostic of both the provided ontology and training data) which maps the extracted subjects to the GND via fuzzy string matching. Our solution achieves an average (avg.) *Recall* of **0.58** (rank $3/11$ across all submitting teams) on the reduced subject index *tib-core-subjects* and **0.48** (rank $6/11$) on the full index *all-subjects* in the task's quantitative evaluations. In the qualitative evaluation with human subject matter experts rating a sample of the proposed results, it obtained an avg. *Precision* of **0.43** on technically correct subjects, and **0.25** on correct and also relevant subjects (rank $8/13$ in both cases), demonstrating competitive performance compared to the state of the art assessed in (D'Souza et al., 2025).

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2413–2423

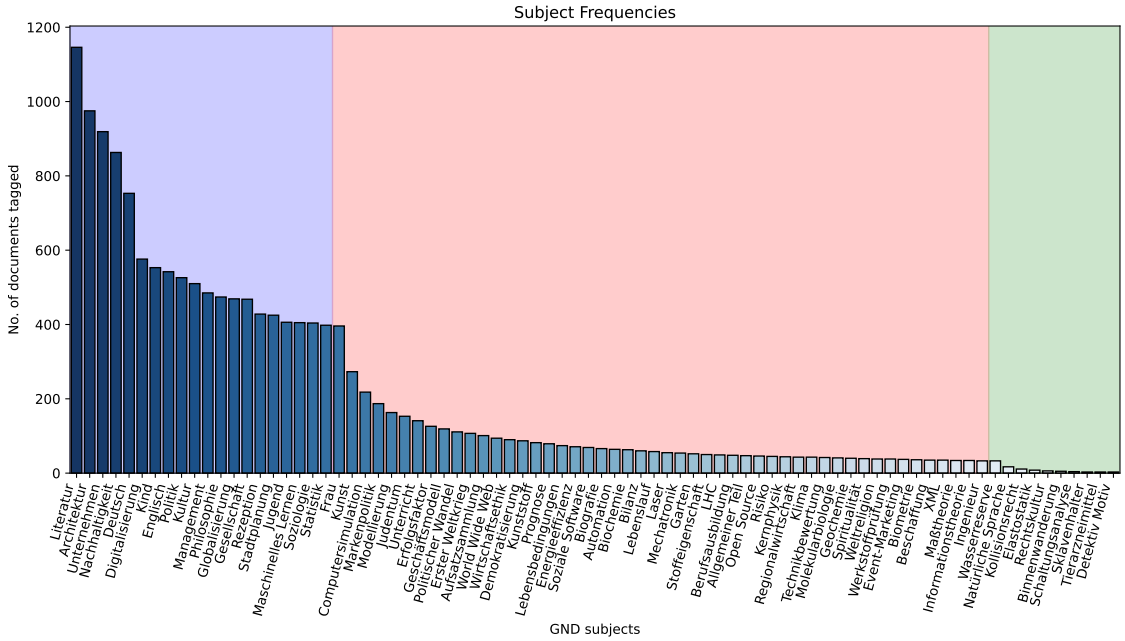

Figure 1: Ordered frequencies of subject terms. The blue shaded area represents the 20 most frequent subjects, the red and green one depict the following GND subjects sampled at sampling rates of 1:200 and 1:1000, resp.

2 Background

Data Sets. Document subject tagging is challenging for librarians due to the sheer size of the subject ontologies involved: For instance, the excerpt from the GND ontology, commonly utilized by German libraries, to be used for this shared task comprises more than 204K subjects, including additional information like their synonyms and description. Task participants should leverage a human-tagged document collection from TIBKAT for training their automated subject tagging systems, comprising five document types (article, book, conference proceedings, report, thesis) in two languages (EN and DE). This data has been split into a *development* (13666 documents) and a *training* set (81937 documents), both including the human experts' assigned GND subjects as ground truth labels, and the *test* data set (27986 documents) to tag. Task organizers made sure that the distributional characteristics between these splits were preserved.

Related Work. Fig. 1 outlines the frequency characteristics of the training corpus: While some – typically taxonomically "higher-level" – subjects have been assigned to a large number of documents, many subjects are fine-grained, specific descriptions assigned to very few documents only. Such imbalanced datasets – with sparse labels – render the training of Machine Learning (ML) classifiers challenging (Kaur et al., 2019). Typically, only very few examples will be available for the specific subject terms, while the learning process will be dominated by those few extremely frequent labels, such as the ones on the left-hand side of Fig. 1 and the subjects belonging to the most frequent macro classes in Fig. 4, like "Informatik (Computer Science)". Training an ML classifier to predict sparse labels from such a large set of possible labels is known as *eXtreme Multi-Label Classification* (XMLC) (Dasgupta et al., 2023). Document subject tagging thus typically has been framed as XMLC problem, as in related work like the system Annif (Suominen et al., 2022) developed for Finnish libraries. Processing multilingual data imposes further challenges, since linguistic variations can affect classification performance (Suominen and Koskenniemi, 2022). The improved Natural Language Understanding (NLU) capabilities of the recently introduced LLMs open up alternative solutions by converting this task into a text generation problem. However, LLMs introduce their own challenges, notably the question of how to constrain their outputs to the subjects indeed occurring in the GND (ultimately, the sought-after output needs to be given in terms of the appropriate GND ID), and moreover LLMs' limitations like the strong variance in their outputs with respect to minor variations in the prompt (Salfinger and Snidaro, 2024; Incitti et al., 2024), which we also found to be an issue in our experiments for this shared task.

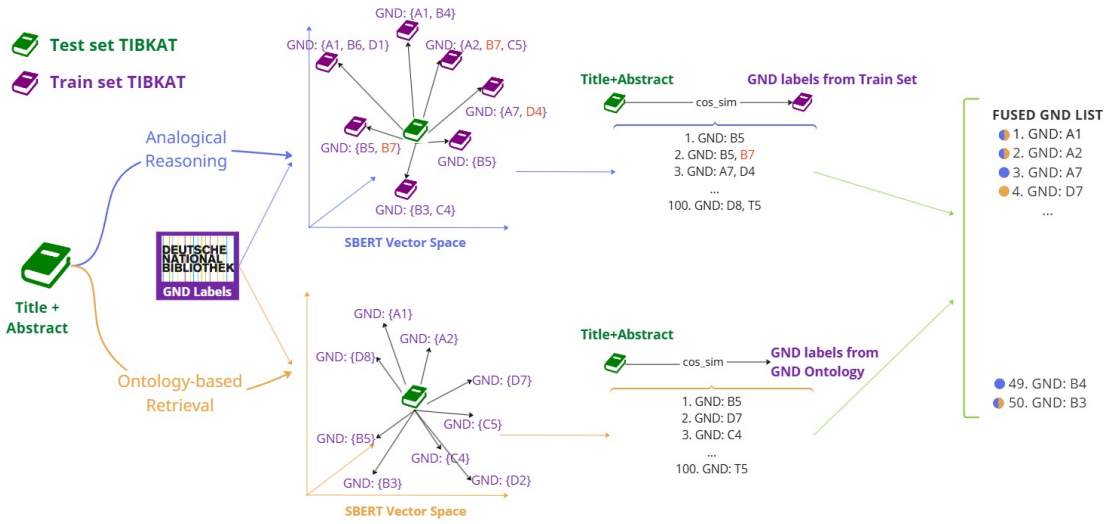

Figure 2: Processing architecture. For visualization purposes, we depict the two embedding vector spaces as 3D down-projections.

3 System overview

Motivation. Based on our insights gained from the data analysis, we aimed to address the imbalance and label-sparsity of the given datasets in a principled manner. Instead of formulating the problem as XMLC, as common in related work (see Sec. 2), we experimented with different reasoning strategies in the document and subject embedding space. These were motivated by the following findings, illustrated on Example Document 1:

1. The first GND subject label, "gender studies", could be inferred from the given title and abstract alone, with gender studies concepts concretely mentioned in the text.

2. Even for a human reader, it might not be easy to derive the other subject labels not explicitly mentioned in the text. Suitable subject tags thus could be inferred from librarians' reasoning on similar documents rather than relying solely on the text of the current document. This approach helps in identifying subject categories that are not explicitly mentioned (e.g., to link "creative labour" with the tags "artist" and "cultural sector"), which we frequently observed in the provided data.

We thus implemented both of these complementary reasoning strategies, as illustrated in Fig. 2 outlining our devised Information Fusion (IF) architecture, which we will describe in the following.

Embedding. For a given query document q (the document we seek to tag), we concatenate its title t and abstract a, forming our input text $q = [t\ a]$. We utilize the `all-mpnet-base-v2`[2] Sentence Transformer model based on MPNet (Song et al., 2020) – a multi-lingual model specifically designed for encoding sentences and paragraphs for tasks like semantic search and clustering – to map q into a 768-dimensional dense vector space. This Sentence Transformer model splits the input text into chunks fitting into its context window size, then converts each chunk into a numeric embedding vector. Finally, the mean vector is computed from these embedding vectors to form a single resulting vector space representation of q, which is then routed to two complementary "reasoning branches".

Ontological Reasoning. The *ontology-based retrieval* branch (lower half of Fig. 2) addresses finding 1., by assessing the semantic similarity of a document's title and abstract with the *subjects from the ontology*. To generate a semantic representation of the ontological concepts, we concatenate the GND subject term's name and its alternate names (such as synonyms or translations of these terms in other languages)[3] into a comma-separated list, which we map into the embedding vector space with the `all-mpnet-base-v2` Sentence Transformer model. This branch thus determines the semantic similarity between the query document q's title and abstract to all available subjects. It returns the $2k$ closest subject embeddings, which are taken as the list of $2k$ subjects returned from this branch. Hence, this

[2] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[3] Note that even more comprehensive embeddings could be created, e.g., by also including the GND descriptions or the explanations of linked Wikipedia articles, if provided. However, in our initial experiments on this, we did not find that including further information improved our results.

branch determines the *document-to-subject* similarities, and thus is purely driven by the semantic content and similarity of the query document's title and abstract with the given GND subjects.

Analogical Reasoning. Conversely, the *analogical reasoning* branch (upper row of Fig. 2) computes the *document-to-document* similarities. q's semantic similarity with *all other documents' titles and abstracts* is determined by computing the cosine similarity of q's and all training documents' embedding vectors. The rationale is that we expect documents on similar semantic content – irrespective of their titles'/abstracts' concrete textual surface forms (such as different synonyms utilized for referring to the same semantic contents) – to be mapped in the same sub-space of the embedding space, forming subject-oriented clusters. By embedding the query document q, we suppose that it will be mapped to the sub-space of topically related documents. Consequently, we assume that the subject tags assigned by domain experts to topically related documents from the training data will also be suitable for our query document q, and thus take the subject tags from the $2k$ most similar training documents as the output list of subject proposals from this branch, ranked according to their documents' similarity with q. More specifically, if document d is the closest training document to q in embedding space, we allocate its n assigned ground-truth subjects to ranks 1 to n of the resulting subject list. Since the shared task organizers clarified that the ground-truth subjects assigned to a document are not ranked, no order can be determined for the subjects *within* a document. We rank these subjects as listed in the training document, but the ordering between subjects from the same document thus can be considered random, which is reflected by all subjects stemming from the same training document sharing the same distance/similarity values. Next, the m subject labels from the second-closest training document are assigned to ranks $n+1$ to $n+m$, and so forth, until all $2k$ documents' subjects have been ranked. We then deduplicate the resulting ranked list by retaining only the first ranked occurrence of each subject (assuming that the order *across* the most similar documents matters – we assume that the earlier a subject appears, the more relevant it might be), eventually returning only the top-$2k$ ranked subjects. This strategy thus essentially implements *analogical reasoning* – or in other terms a form of *case-based reasoning in embedding space*: It identifies the most related documents from the human expert-labeled corpus under the assumption that subjects proposed for semantically similar documents will also be suitable subjects for q. Analogical inference hence allows to identify subjects not explicitly referred to in q's title and abstract, such as higher-level taxonomic subject descriptors assigned by the human domain experts. Such analogical reasoning appears thus suitable for emulating the experts' decision making in a case-based manner. This mitigates the problem of the skewed training data we observe in Fig. 1: No matter whether a document has a multitude or a limited set of similar instances in the training data – as long as *some* good examples do exist in the training data, this case-based reasoning will leverage only those for its subject inference, which thus does not get dominated by more frequent instances of other classes.

Fusion. A final *fusion* step combines the outputs retrieved from both branches: Both similarity-ranked lists are joined and re-ranked based on the *total order* of their elements' distances to q across both lists. Since the same embedding vector space, i.e., embedding model, is utilized in both branches, these distances are directly comparable. Again, duplicate subjects are removed, with only the first ranked occurrence being kept. This deduplication step is also the rationale behind the design decision of returning the top-$2k$ closest subjects from each branch, to ensure that at least k subjects are retained after deduplication and across-branch ranking. The top-k subjects of this ranked list are returned as the final result of k proposed subjects, ordered according to presumed relatedness to q.

4 Experimental setup

To estimate our models' generalization error and perform model selection, we created an internal validation split by sampling 10% of the provided training data, utilizing the remaining 90% for creating our models. This also allowed us to conduct ablation studies to measure the impact of the individual reasoning branches and the fusion approach. For producing our test set submission, we combined the provided training and development split for "training" our models, which in our approach only means embedding these corpora with `all-mpnet-base-v2`, utilized via the Sentence Transformers/SBERT framework (Reimers and Gurevych, 2019). Evaluation metrics comprise *Precision*, *Recall*, and *F1 measure* assessed for dif-

ferent cut-offs of top-k subjects[4]. In addition to the quantitative evaluation on existing *ground truth data* (which typically includes few ground truth subjects per document, usually 5-7 subjects), task organizers sampled 122 test documents across 14 subject classifications for the qualitative evaluation. The generated top-20 subject labels output by the participant systems then were marked by human subject matter experts as either *correct*, *technically correct but irrelevant*, or *incorrect* subject labels.

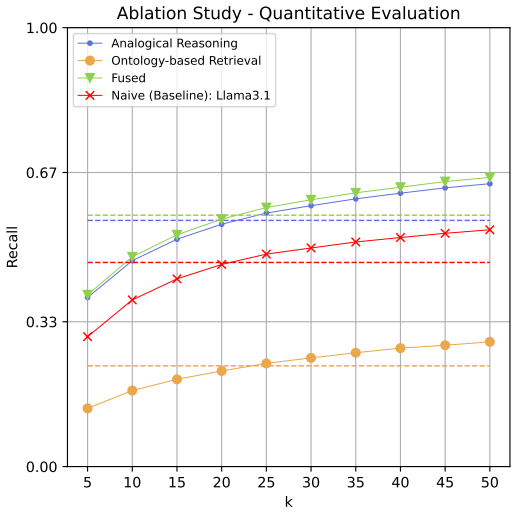

Figure 3: Ablation study comparing the individual branches, the naive baseline, and the fused results. Dashed lines represent the average values across the varied top-k values (solid lines).

5 Results

Ablation Study. Fig. 3 compares the individual performance of our two reasoning branches and the fusion approach on our own validation set. Analogical reasoning outperforms the naive baseline, and also the ontology-based retrieval branch by a large margin. Fusing the results of both branches – analogical reasoning and ontology-based retrieval – improves the metrics slightly further, suggesting that both branches extract complementary information, which is also backed by empirical evidence: Example Document 2 (document ID: 3A1649002734) shows one output from our IF system with subject predictions contributed from both reasoning branches. This highlights the feasibility of fusing the outputs of both ontological and analogical reasoning, which may retrieve complementary information. The curve of the fusion approach also surpasses the naive `llama3.1:8B`-based model, which presumably has more advanced NLU capabilities but does not incorporate the information from the ontology and the training data in its reasoning. Thus, fusing the outputs of simpler, domain-specific models – geared to the problem domain and focused on one specific reasoning approach at a time – is capable of outperforming a larger, generic and computationally far more costly model.[5]

Example Document 2

Title: Bone Densitometry in Growing Patients : Guidelines for Clinical Practice
Abstract: Focuses on the use of Clinical Densitometry in the pediatric and adolescent populations. This book, suitable for clinicians and technologists involved in the care of children and adolescents, provides expert-based guidelines to assist practitioners in performing dual-energy x-ray absorptiometry in younger patients and in interpreting the data.
GND Subject Labels: Kind (child); Osteodensitometrie (osteodensitometry);

Color coding: correct labels identified with
- ontological reasoning
- analogical reasoning

Quantitative Evaluation. Fig. 5 compares the performance obtained by our approach in the shared task's quantitative evaluation to other teams' submitted solutions. Our semantic retrieval-based reasoning fusion achieved an avg. *Recall* of **0.58** (rank $3/11$ across all submitting teams) on the reduced index *tib-core-subjects*[6], and **0.48** (rank $6/11$) on the full subject index *all-subjects*[7]. Since we employed the same model using the entire GND for submitting to both test collections, the strong performance on *tib-core-subjects* is particularly noteworthy, as limiting our predicted subjects to *tib-core-subjects* presumably would have further improved our results. Tables 1 and 3 present our language- and document type-level results, both revealing a superior performance of our model on English than on German texts. Table 2 shows the results we have obtained on our internal validation set used for model selection for *all-subjects*, in which we observe higher Recall for most document types.

Qualitative Evaluation. In the qualitative evaluation, where human subject matter experts rated a sample of the submitted test results, our IF approach obtained an average *Precision* of **0.43** on

[4] based on varying k from 5 to 50 (quantitative eval.) and from 5 to 20 (qualitative eval.), in increments of 5.

[5] On a GPU RTX A5000, avg. execution time (internal validation split): 'Analogical Reasoning': 0.32s, 'Ontology-based Retrieval': 0.74s, 'fused': 1.03s, 'Naive (Baseline): Llama3.1': 15.95s.

[6] min: 0.06, max: 0.66, $\mu = 0.41$, $\sigma = 0.2$

[7] min: 0.13, max: 0.63, $\mu = 0.44$, $\sigma = 0.17$

technically correct subjects[8] (case 1), and **0.25** on correct and also relevant subjects[9] (case 2), corresponding to rank $8/13$ in both cases (see Fig. 6 for a comparison to other teams' performance). Figs. 7 and 8 plot the obtained Recall, Precision and F1 measure of our approach for the different subject classes, according to the two cases analyzed. As expected, we observe a consistent trend across subject classes, with humanities' subjects like linguistics and history performing worse than the natural and technical sciences, which tend to use more specific custom terminology more directly related to their subject tags in their titles and abstracts, representing an easier inference problem.

Error Analysis & Future Work. As we see in Fig. 3, the IF approach is mostly driven by analogical reasoning: The final similarity ranking is dominated by document-document similarities, with similar embedded documents generally having smaller distances to q than its most similar embedded subjects. In general, even if the ontological branch retrieves a relevant subject in its top ranks, it thus may not appear at the same rank in the fused list. This is dissected in Fig. 9, depicting the fractions of subjects in the fused list proposed by the ontological branch (after fusion and de-duplication) on our *dev* set across different values of k. While we observe a low contribution *on average* (with a median of subjects from the ontological branch greater zero only for $k > 40$), the strong presence of outliers indicates that there are *specific documents* for which most and sometimes even all proposed subject labels originate from the ontological branch. Ontological retrieval complements analogical reasoning by allowing to handle novel documents which do not have good related "exemplars" in the training data yet, which thus could not be covered with analogical reasoning. In total, we observe that in 53.43 % of documents in the *dev* set, the ontological branch indeed contributes at least one subject label (irrespective of its correctness). However, for only 5.31 % of documents, ontology-based retrieval identified at least one ground truth subject, thereby increasing the performance of the fusion approach. Whilst this might appear low, we also find that in 28.17 % of documents, both branches retrieved duplicate labels, with one having been subsequently discarded by de-duplication (which is typically the lower-ranked ontological one). However, we note that a subject candidate simultaneously proposed by both reasoning branches actually would increase confidence in its relevance – for future work, we thus plan to factor this in by up-ranking such cases in the final fusion and re-ranking step.

Moreover, as illustrated by the GND IDs marked in red in Fig. 2, assuming that *all* subjects from similar documents fit q is a strong assumption which can lead to false positives, since this might not apply to *all* subjects from a similar document. This issue becomes evident in the comparably worse qualitative evaluation results, suggesting future work on further filtering the subjects output by the analogical branch.

6 Conclusion

Our proposed system explores innovative solutions to the challenging problem of document subject tagging: By operating in embedding space for identifying the semantically most similar documents and subjects, the corpus imbalance on the different subjects is not an issue: No matter whether a cluster/manifold is densely or sparsely populated, only the distances to the closest training documents and subjects are factored in. The relative frequency of documents per subject hence does not dominate or bias the "learning" process, unlike with training a neural network-based model with gradient descent or related ML classifiers. If the semantically closest document has similar contents and matching subject tags, those will be consistently proposed as the top-ranked subjects. Hence, our approach also corresponds to an interpretable and transparent strategy in terms of Explainable AI (xAI) (Longo et al., 2024). Since our model "training" only requires the embedding of the GND subjects and labeled training corpus, our approach is computationally efficient and suitable for "online learning" – if new labeled documents should be incorporated in the model, those only need to be embedded into the vector space, no "retraining" is required for dynamically growing the training corpus.

From the shared task's evaluations, we conclude that our semantic retrieval-based IF system is competitive with state-of-the-art XMLC approaches such as Annif, as comparatively evaluated in (D'Souza et al., 2025).

For future work, we aim to enhance our fusion by improving re-ranking, for instance by up-ranking results identified by both reasoning strategies.

[8] min: 0.07, max: 0.60, $\mu = 0.41$, $\sigma = 0.17$

[9] min: 0.04, max: 0.38, $\mu = 0.24$, $\sigma = 0.11$

Acknowledgments

This research was funded in whole, or in part, by the Austrian Science Fund (FWF) [Grant J4678-N].

This work was partially funded by the European Union – NextGenerationEU National Recovery and Resilience Plan (NRRP) M4C2 Inv. 3.3 D.M. 352/2022 and D.M. 630/2024. The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them. We also thank LimaCorporate, affiliate of Enovis Corporation, for supporting this research.

We sincerely thank Nicola Ferraresi, Ignazio Gasperi, Francesco Gorgone, Leonardo Ruoso and Michele Somero for their valuable contributions to this work.

References

Arpan Dasgupta, Siddhant Katyan, Shrutimoy Das, and Pawan Kumar. 2023. Review of extreme multilabel classification. *arXiv preprint arXiv:2302.05971*.

Jennifer D'Souza, Sameer Sadruddin, Holger Israel, Mathias Begoin, and Diana Slawig. 2025. SemEval-2025 Task 5: LLMs4Subjects - LLM-based Automated Subject Tagging for a National Technical Library's Open-Access Catalog. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1082–1095, Vienna, Austria. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Koraljka Golub. 2021. Automated Subject Indexing: An Overview. *Cataloging & Classification Quarterly*, 59:1–18.

Francesca Incitti, Andrea Salfinger, Lauro Snidaro, and Sri Challapalli. 2024. Leveraging LLMs for Knowledge Engineering from Technical Manuals: A Case Study in the Medical Prosthesis Manufacturing Domain. In *27th International Conference on Information Fusion (FUSION 2024)*. ISIF.

Francesca Incitti, Federico Urli, and Lauro Snidaro. 2023. Beyond word embeddings: A survey. *Information Fusion*, 89:418–436.

Harsurinder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. 2019. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM computing surveys (CSUR)*, 52(4):1–36.

Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, Richard Jiang, Hassan Khosravi, Freddy Lecue, Gianclaudio Malgieri, Andrés Páez, Wojciech Samek, Johannes Schneider, Timo Speith, and Simone Stumpf. 2024. Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Andrea Salfinger and Lauro Snidaro. 2024. Probing the consistency of situational information extraction with large language models: A case study on crisis computing. In *2024 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA 2024)*, Montreal, Canada.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and Permuted Pre-training for Language Understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.

Osma Suominen and Ilkka Koskenniemi. 2022. Annif analyzer shootout : comparing text lemmatization methods for automated subject indexing. *The Code4Lib Journal*, (54).

Osma Suominen, Mona Lehtinen, and Juho Inkinen. 2022. Annif and Finto AI : Developing and Implementing Automated Subject Indexing. *JLIS*, (1).

Michelle R Turvey and Karen M Letarte. 2014. Cataloging or knowledge management: Perspectives of library educators on cataloging education for entry-level academic librarians. In *Education for Cataloging and the Organization of Information*, pages 165–187. Routledge.

A Appendix

A.1 Data Analysis

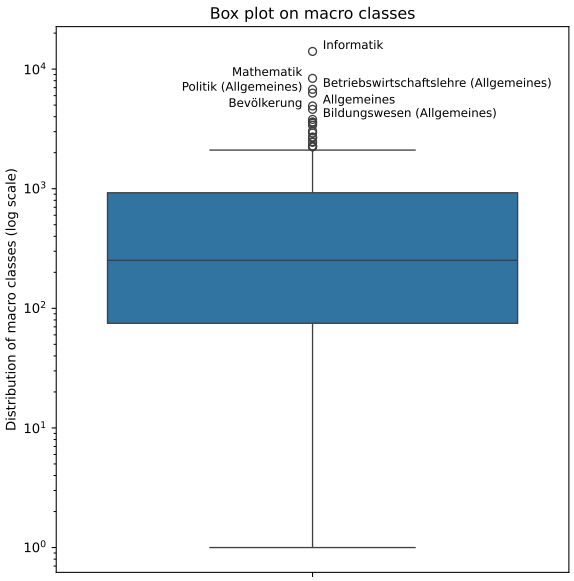

Figure 4: Plot on the "macro classes" that are also given for each subject term (indicating its "parent term" in the taxonomy).

A.2 Results

Language	Type	Recall	Precision	F1-score
DE	Article	0.000	0.000	0.000
	Book	0.478	0.062	0.103
	Conference	0.405	0.067	0.107
	Report	**0.512**	0.066	0.110
	Thesis	0.312	0.057	0.090
EN	Article	**0.831**	0.109	0.179
	Book	0.539	0.070	0.116
	Conference	0.591	0.083	0.136
	Report	0.521	0.069	0.115
	Thesis	0.420	0.071	0.113
Overall	Average	0.482	0.065	0.108

Table 1: Condense representation of achieved metrics on test evaluation all-subjects split. The results were produced by the organizers. The Overall Average is computed using a weighted average to avoid the influence of the data imbalance.

Language	Type	Recall	Precision	F1-score
DE	Article	0.000	0.000	0.000
	Book	**0.584**	0.067	0.109
	Conference	0.500	0.069	0.109
	Report	0.561	0.064	0.104
	Thesis	0.365	0.059	0.091
EN	Article	**0.829**	0.117	0.184
	Book	0.625	0.073	0.120
	Conference	0.682	0.086	0.137
	Report	0.554	0.070	0.114
	Thesis	0.426	0.072	0.110
Overall	Average	0.573	0.070	0.113

Table 2: Condense representation of achieved metrics on our internal all-subjects split. The Overall Average is computed using a weighted average to avoid the influence of the data imbalance.

Language	Type	Recall	Precision	F1-score
DE	Article	NaN	NaN	NaN
	Book	0.576	0.079	0.130
	Conference	0.433	0.080	0.125
	Report	**0.662**	0.085	0.142
	Thesis	0.339	0.071	0.107
EN	Article	**0.706**	0.164	0.242
	Book	0.633	0.079	0.132
	Conference	0.691	0.096	0.158
	Report	0.580	0.081	0.132
	Thesis	0.458	0.083	0.130
Overall	Average	0.579	0.080	0.132

Table 3: Condense representation of achieved metrics on the tib-core split. The results were produced by the organizers. The Overall Average is computed using a weighted average to avoid the influence of the data imbalance. NaN values in the first row mean that there were no documents in that subclass.

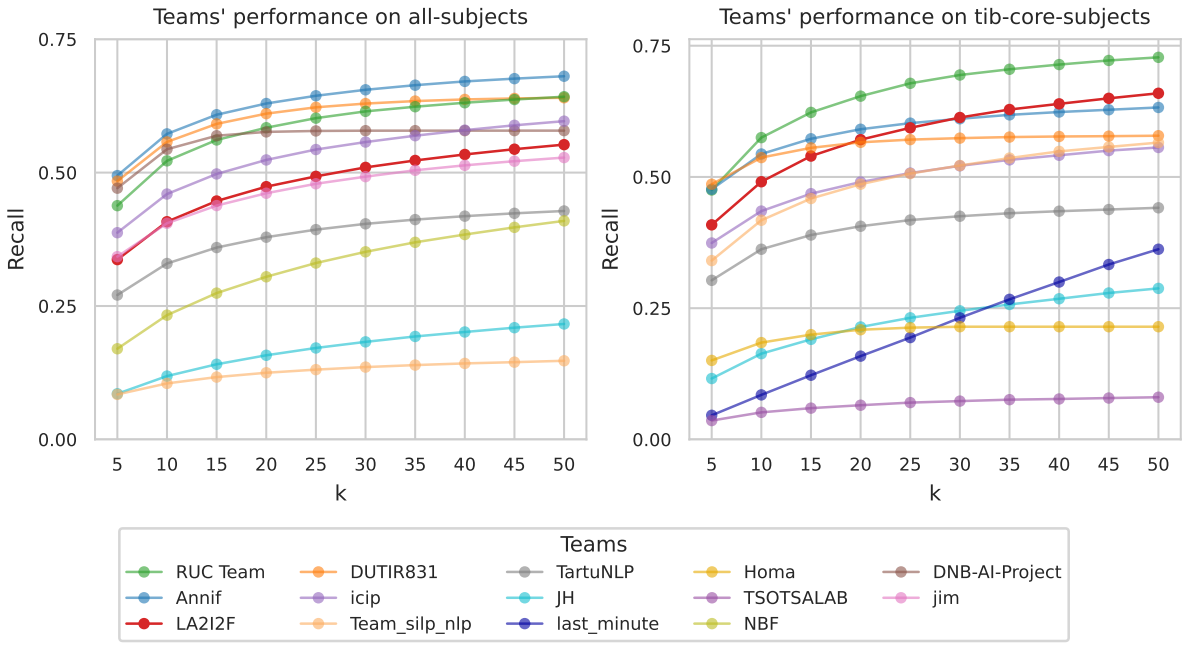

Figure 5: Performance comparison of our proposed approach (LA2I2F) to all other teams' submitted solutions for the quantitative evaluation. Note that not all teams submitted to both test collections.

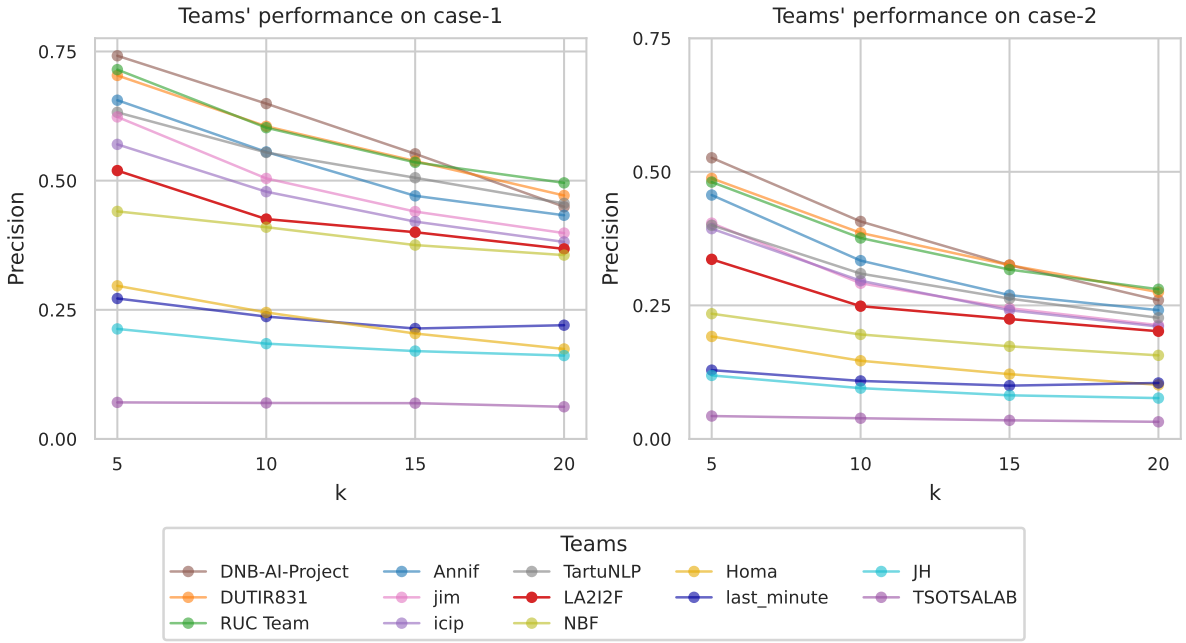

Figure 6: Performance comparison of our proposed approach (LA2I2F) to all other teams' submitted solutions for the qualitative evaluation.

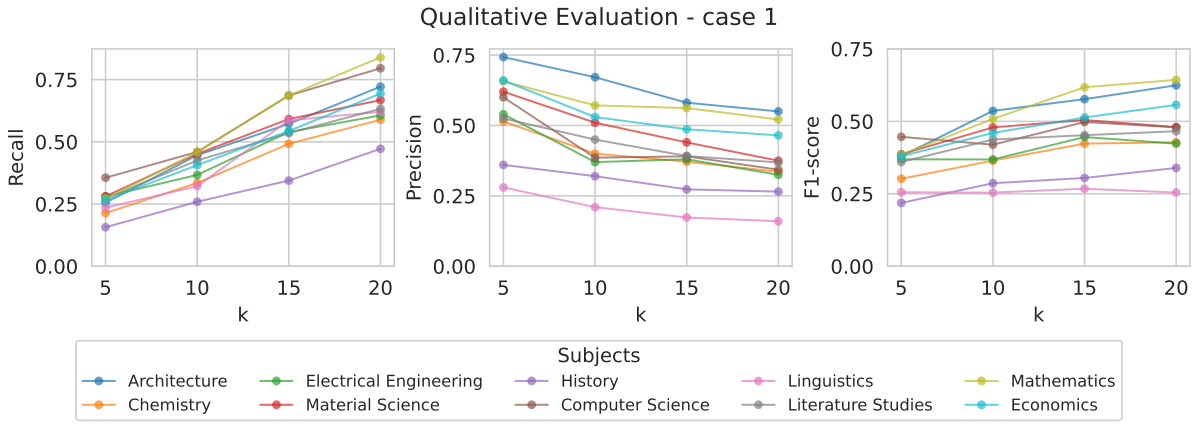

Figure 7: The qualitative evaluation of our approach on case 1 takes into consideration both types of label ratings assigned by the human experts: *correct* subjects and *technically correct, but irrelevant* subjects, and counts both of them as correct labels for determining the share of correct system outputs. The most relevant metric in this case is Precision, rating the correctness of the predicted labels as judged by the human expert.

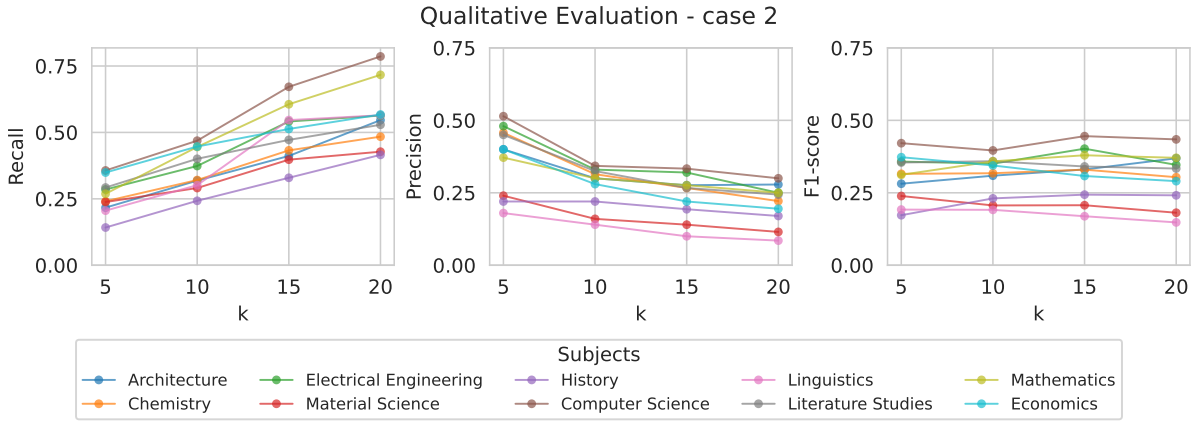

Figure 8: The qualitative evaluation of our approach on case 2 takes into consideration just those which the human experts marked as *correct*, thus, discarding subjects which are *(technically) correct but irrelevant*. As we observe, excluding the *(technically) correct but irrelevant* subjects leads to a significant drop in Precision values.

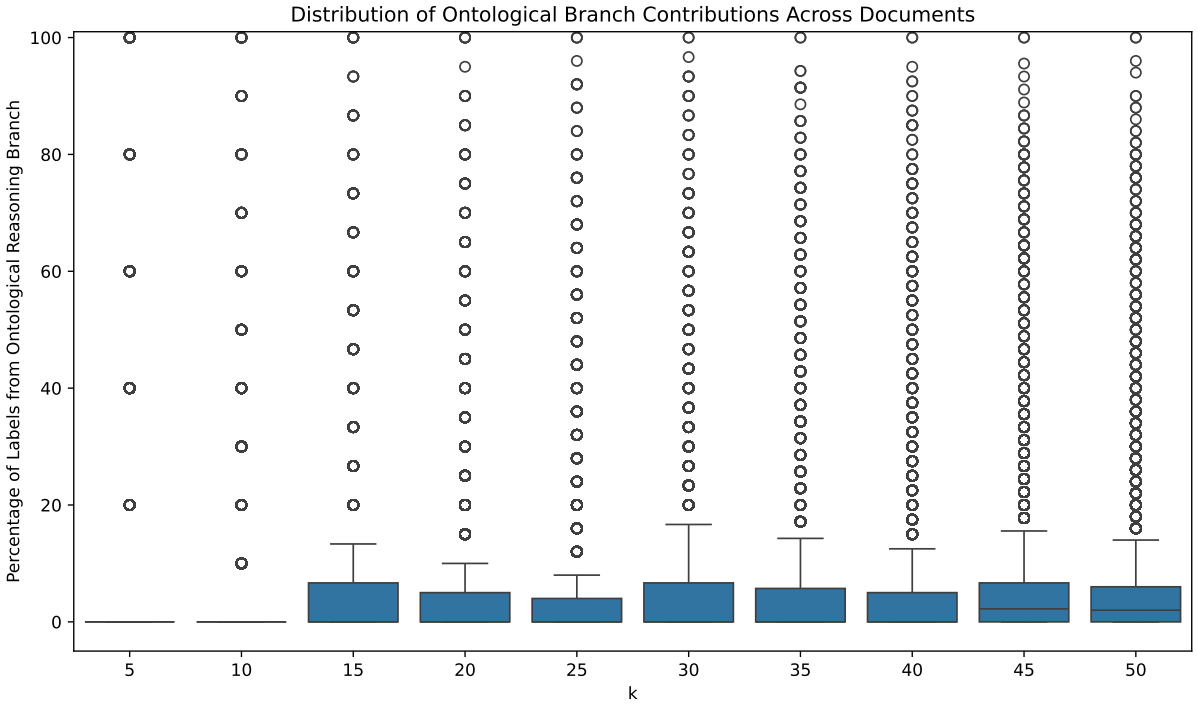

Figure 9: Fractions of subject labels in the fused list proposed by ontology-based retrieval, across varied top-k. We observe a low contribution *on average* (with a median of subjects from the ontological branch only greater zero only for $k > 40$), but the strong presence of outliers indicates documents for which most or even all proposed subject labels originate from the ontological branch.