Multi-Property Multi-Label Documents Metadata Recommendation based on Encoder Embeddings

Nasredine Cheniki
Publications Office
of the European Union
nasredine.cheniki@ext.ec.europa.eu

Vidas Daudaravicius
European Commission
Joint Research Centre
vidas.daudaravicius@ec.europa.eu

Abdelfettah Feliachi
Publications Office
of the European Union
abdelfettah.feliachi
@ext.publications.europa.eu

Didier Hardy
Publications Office
of the European Union
didier.hardy@publications.europa.eu

Marc Wilhelm Küster
Publications Office
of the European Union
marc.kuster@publications.europa.eu

Abstract

The task of document classification, particularly multi-label classification, presents a significant challenge due to the complexity of assigning multiple relevant labels to each document. This complexity is further amplified in multi-property multi-label classification tasks, where documents must be categorized across various sets of labels. In this research, we introduce an innovative encoder embedding-driven approach to multi-property multi-label document classification that leverages semantic-text similarity and the reuse of pre-existing annotated data to enhance the efficiency and accuracy of the document annotation process. Our method requires only a single model for text similarity, eliminating the need for multiple property-specific classifiers and thereby reducing computational demands and simplifying deployment. We evaluate our approach through a prototype deployed at the European Commission for daily operations, which demonstrates superior performance over existing classification systems. Our contributions include improved accuracy without additional training, increased efficiency, and demonstrated effectiveness in practical applications. The results of our study indicate the potential of our approach to be applied across various domains requiring multi-property multi-label document classification, offering a scalable and adaptable solution for metadata annotation tasks.

1 Introduction

Metadata facilitates navigation through extensive document collections, offering insights into data usage, retrieval, traceability, and reusability. It also refines search processes within large datasets. Document classification, also known as document annotation, is crucial for information retrieval applications and involves tagging documents with various metadata. This task is laborious, especially when multiple labels per document are required. The complexity increases with multi-property multi-label classification tasks, where each property may contain multiple labels.

Recent advances in document classification (Song et al., 2022; Chalkidis et al., 2019) using natural language processing have significantly improved efficiency, accuracy, and completeness of metadata. However, these methods typically necessitate the development and training of separate models for each classification property, which is resource-intensive and time-consuming. Furthermore, the need for continuous retraining to update these models with new properties and labels presents challenges in scalability and adaptability.

In this paper, we introduce an innovative approach for multi-property, multi-label document classification that is driven by encoder embeddings. Our method capitalizes on semantic-text similarity,

Proceedings of the Natural Legal Language Processing Workshop 2024, pages 233–242
November 16, 2024 ©2024 Association for Computational Linguistics

using pre-annotated datasets to streamline the annotation process. It stands out by eliminating the need for additional model training or fine-tuning, which greatly reduces computational requirements and eases deployment. Our proposed approach is highly applicable in use-cases where multi-label annotated data already exists which is often a use case in various enterprises.

We leverage pre-trained models like BERT(Devlin et al., 2018) to avoid fine-tuning, making our solution more scalable and flexible. Our prototype, tested at the European Commission for daily operations, has outperformed existing systems, enhancing efficiency and accuracy in document annotation. The success of our approach suggests its applicability in various settings that require sophisticated document classification.

The key contributions of our study are as follows:

– Improved Accuracy Without Additional Training: We leverage pre-trained embeddings to enhance the accuracy of document classification without the need for further training. This approach not only speeds up the process but also yields better accuracy in annotating documents with various properties and labels.
– Enhanced Efficiency: Our technique utilizes a single text similarity model instead of multiple classifiers tailored to specific properties. This greatly simplifies deployment in practical settings where there are often limitations on computational resources and time.
– Proven Practical Effectiveness: Our method's integration into the daily operations of the European Commission. Empirical results shows that our approach outperforms existing systems in document classification tasks.

2 Related Work

Classifying large collections of documents is time intensive and consuming task. However, recent breakthroughs in NLP and the development of large language models (LLMs) have greatly improved the efficiency of this process. Avram et al. (2021) proposed a framework for classifying documents according to the EuroVoc framework in 22 different languages[1] by fine-tuning advanced Transformer-based pretrained language models. This method has shown significant improvements in classification accuracy. Nonetheless, it requires individual training for each language, leading to high computational demands and difficulties in scaling, particularly when new descriptors or languages need to be added.

Suominen (2019) introduced Annif, a tool that automates the labor-intensive process of subject indexing for librarians. It uses a combination of existing tools and various NLP algorithms to boost accuracy and versatility for different types of documents. However, its effectiveness might be limited in environments with constantly changing content.

Chalkidis et al. (2019) developed a technique for classifying legal documents using a dataset annotated with EuroVoc labels, comprising 57,000 texts. They found that self-attention mechanisms and domain-specific embeddings notably improve classification performance. However, this method is computationally expensive, particularly for long documents, due to the inclusion of GRU units.

Chang et al. (2020) created the X-Transformer model to address issues in extreme multi-label text classification, which involves dealing with vast output spaces and tackling the problem of label sparsity. Their model surpasses traditional models in various benchmarks and achieves top-tier results. However, this model requires considerable GPU resources and has scalability issues when faced with large sets of labels due to memory limitations.

Wan et al. (2019) tackled the challenge of classifying long legal documents by breaking them down into smaller sections. They found that this segmentation, along with the use of BiLSTM networks and simpler architectures, made it easier to process lengthy texts. The effectiveness of this approach depends heavily on the quality of the initial segmentation, as poor segmentation can lead to complications in the model's implementation and fine-tuning, especially if it doesn't correctly reflect the thematic or semantic divisions within the documents.

3 Background and Definitions

In this section, we provide detailed definitions of the terms and key concepts used in this paper, including Document, Context, Metadata , semantic text similarity, k-nearest neighbors.

3.1 Document and dataset

A document in the context of this study consists of two main components: text and metadata. Let d_i

[1] In 2024, the number of supported languages in EuroVoc is 27.

denote a document that belongs to a dateset DT, which can be formally described as:

$$d_i = (T_i, M_i) \in DT$$

where T represents the text component of the document and M represents the metadata associated with the document.

3.1.1 Text

A text is a primary component of a document that refers to the plain, natural language content that conveys information. This includes sentences, paragraphs, titles, abstracts, and other narrative elements. The text can be further decomposed into specific contexts $C_1, C_2, \ldots, C_n$, where each C_i denotes a specific part of the text relevant to the analysis.

$$T = \{C_1, C_2, \ldots, C_n\}$$

3.1.2 Metadata

Metadata is structured information which provides additional context and attributes that help to categorize and identify the document.

The metadata M consists of various properties $P_1, P_2, \ldots, P_m$ and their corresponding sets of values, where each property P_i is an attribute of the document, and $\{V_{i1}, V_{i2}, \ldots, V_{in_i}\}$ are the values assigned to that attribute. A property can have multiple values.

$$M = \{(P_i, \{V_{i,j}\})\}$$

3.1.2.1 Classification Properties

Classes are predefined categories or labels that are assigned to documents based on their content. Within this study, each class is denoted by V_{ij}, where i represents the property index, and j denotes the specific class within that property. A document d_i can belong to one or more of these classes based on the corresponding property.

Let V_{ij} represent a specific class for property p_i. The membership of a document d in multiple classes is represented as follows:

$$d \in \bigcup_{i,j} V_{ij}$$

Here, the notation $\bigcup_{i,j} V_{ij}$ indicates the union of classes to which the document d may belong, emphasizing that a document can be associated with multiple classes across different properties.

3.2 Embedding

Embedding is a technique used to convert the context of a document into a vector in a continuous vector space. This vector representation captures the semantics of the context, allowing for various computational operations such as similarity measurements and clustering. Let E_m denote an embedding function based on model m that maps the context of a document C to a vector $\mathbf{v}$ in an n-dimensional continuous vector space. Formally, the embedding function E_m can be described as:

$$E_m : C_i \rightarrow \mathbb{R}^n$$

where:

- C_i is the context of the document, which can be a sentence, paragraph, or any specific part of the text.
- $\mathbb{R}^n$ is the n-dimensional continuous vector space.
- $\mathbf{v_i} = E_m(C_i)$ is the resulting n-dimensional vector that represents the semantics of the context C_i.

Embeddings are typically obtained using neural network models trained on large text corpora, such as DistilBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2020) or BERT (Reimers and Gurevych, 2019). The length of a context is delimited by the input size of such transformer language model.

3.3 K-Nearest Neighbors (K-NN)

K-Nearest Neighbors (K-NN) is a machine learning algorithm used to identify the most similar documents to a new document based on their embeddings.

For a given document d_x and its vector representation $\mathbf{v_x}$, K-NN aims to find the subset $S \subseteq D$ of k documents that are highly similar to d_x as measured by a specific distance metric μ (e.g., Euclidean distance).

$$S_{d_x} = \text{kNN}(\mathbf{v}_x, \mathbf{v}_i, \mu, k)$$

where:

- d_x: The new document for which we are finding the nearest neighbors.
- v_x: The vector embedding of the new document.
- $\mathbf{v}_i$: The vector embeddings of all documents in the dataset D.
- μ: The metric used for measuring distances (Euclidean or Manhattan).
- k: The number of nearest neighbors to retrieve.

The function maximizes the similarity between $\mathbf{v}_x$ and $\mathbf{v}_i$ to identify the nearest neighbors.

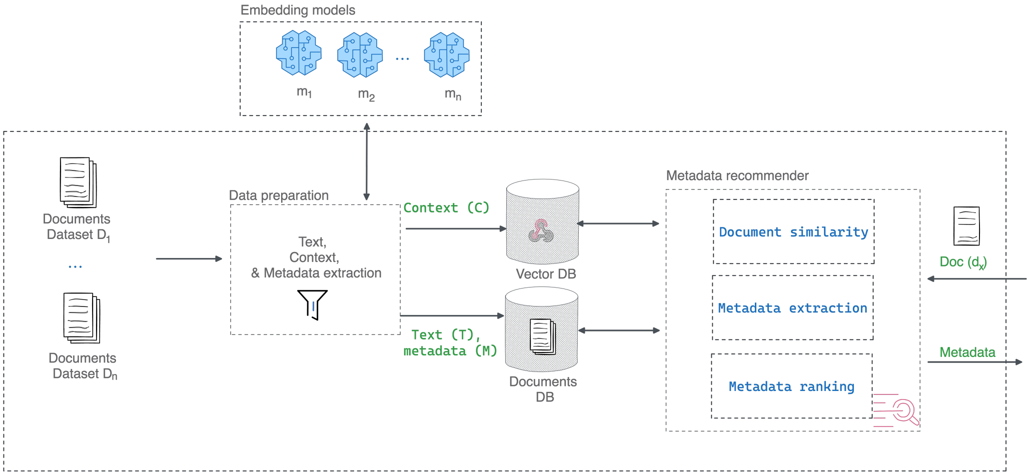


Figure 1: k–NN-based metadata replication framework

4 Multi-property Multi-label Documents Classification

In this section, we present our method for Multi-property Multi-label document classification. Figure 1 illustrates the key components of our approach, from document datasets to metadata recommendation.

4.1 Data Preparation

Initially, datasets retrieved from various sources undergo data cleaning and preparation. This step ensures that the data is consistent, accurate, and ready for processing. This preparation is crucial for effective feature extraction in later stages.

4.2 Context and Vector Embeddings

The number of contexts per document is not balanced among documents, sometimes can reach over 2000 contexts in a document, and it may lead a long document to be overvalued. Therefore, in this stage, only the first context C_1 is extracted from each document which is a title and a summary often. Context may include all textual content or parts of the document. Subsequently, this context is used in the embedding process, where an embedding model m converts the context into vector embeddings $\mathbf{v}_i$. These embeddings capture the semantics of the text and are stored in a vector database (Vector DB).

$$\mathbf{v} = E_m(C_1)$$

4.3 Metadata and Text Database

Parallel to embedding, metadata M_i and a first context C_1 from each document are extracted and stored in a vector database (vector DB). This database supports facilitates access to both the raw text and its associated metadata, ensuring that these elements are readily available for retrieval and analysis.

4.4 Metadata recommendation

This phase is central to our approach and involves several sub-processes designed to leverage the prepared data and embeddings for effective metadata recommendation.

4.4.1 Document Similarity

When a new document d_x is introduced, the system applies K-NN algorithm to find the n most similar documents from the Vector DB based on their vector embeddings $\mathbf{v}_i$. This process identifies k documents with the highest semantic similarity to d_x, suggesting a high potential relevance of their metadata for d_x.

$$\text{kNN}(\mathbf{v}_x, \mathbf{v}_i, \mu, k) \to \{d_1, d_2, \ldots, d_k\}$$

where: μ is the Euclidean distance metric used to measure the similarity between two vector embeddings. It is defined as:

$$\mu(\mathbf{v}_x, \mathbf{v}_i) = \sqrt{\sum_{j=1}^{n} (v_{xj} - v_{ij})^2}$$

4.4.2 Metadata Extraction

The metadata identified for document d_x in the previous stage is extracted and collected for further processing, where:

$$\text{Metadata}(d_1, \ldots, d_k; P) \to \{V_{P,1}^{(f_1)}, \ldots, V_{P,n}^{(f_m)}\}$$

where each $V_{P,i}$ is the values of a specific metadata property P and f_i is the frequency of occurrence of $V_{P,i}$ in the similar documents $d_1, d_2, \ldots, d_k$.

4.4.3 Metadata Ranking

Using a defined scoring function, such as frequency, the list of metadata associated with the retrieved top documents is then ranked. The scoring function can be formalized as follows:

$$\text{Scoring}(d_x, P) \rightarrow \left[\mathrm{V}_{P,1}^{(s_1)}, \mathrm{V}_{P,2}^{(s_2)}, \ldots, \mathrm{V}_{P,m}^{(s_m)}\right]$$

where $\mathrm{V}_{P,i}^{(s_i)}$ denotes the metadata value V_i of a property P, with a score s_i assigned based on the frequency of its occurrence in documents similar to d_x. The list is sorted in descending order of s_i, indicating that values with higher scores are deemed more relevant to d_x.

The ranked metadata from this comprehensive process is then used to classify the new document d_x.

5 Implementation

5.1 Overview

We conducted two evaluation experiments. The first experiment aimed to validate the hypothesis that 'similar documents should have similar metadata.' In this experiment, a random set of documents was selected and subjected to the metadata recommendation process. The metadata recommended by the process was then compared with the metadata previously attributed to these documents.

The second experiment involved deploying our prototype in a real-world scenario to collect user feedback for benchmark comparisons. This allowed us to directly compare the performance and effectiveness of our approach with existing annotation systems.

5.2 Documents dataset

In this experiment, we utilize CELLAR as the document dataset. CELLAR[2] is the semantic repository of the European Union (EU) official publications, managed by the EU Publications Office(Francesconi et al., 2015). Documents in CELLAR are manually annotated by human agents. There are many metadata attributes assigned to documents, including publication date, document type, EuroVoc thesaurus concepts, and more. In this study, we focus on recommending properties that provide classifications, such as EuroVoc concepts.

Accordingly, we define our document dataset as:

$$DT = \{\text{CELLAR}\}$$

All documents, along with their embedding vectors, are stored in an Elasticsearch database. Metadata is retrieved directly from CELLAR as needed using its SPARQL endpoint.[3]

5.3 Metadata

Various controlled vocabularies are used to label documents in CELLAR (see example in Table 1). The Common Data Model (CDM)[4] provides a variety of properties (predicates) for describing bibliographic resources (documents, agents, events, etc.). In our study we focus on the properties of CDM that are more likely to be related to the topic or the theme of documents. For this purpose, we identified a set of properties that fulfil our objectives. We selected the following metadata properties:

- *EuroVoc concepts*: EuroVoc[5], a multilingual interdisciplinary thesaurus, that allows assigning specific topics to the description of resources. With more than 8000 terms in EuroVoc thesaurus, selecting the correct values to annotate documents with an acceptable accuracy is a time consuming task, even for experts with knowledge about the content of EuroVoc and the documents to annotate.
- *rdf type*: generic document type. There are 505 document types to describe any document in Cellar. For instance, thematic domain, EuroVoc concept, etc.
- *Theme*: the subject of the publication
- *Resource type*: the resource type of a work.
- *Subject matter*: a legal document is about a concept expressed as a subject matter. Very often this property is similar to EuroVoc concepts but is used for different purposes.

Therefore, our classification properties are defined as follows:

$$P = \{\text{EuroVoc, RDF-Type, Theme, Subject,}\\ \text{Resource-Type}\}$$

The example of various document properties can be found in Table 1.

[2] https://op.europa.eu/en/web/eu-vocabularies/news/-/blogs/new-brochure-about-cellar

[3] https://publications.europa.eu/webapi/rdf/sparql

[4] https://op.europa.eu/it/web/eu-vocabularies/cdm

[5] https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/eurovoc

Actual labels	Proposed labels	Frequency
EuroVoc descriptors		
EU financial instrument, **investment**, **structural policy**, **transmission network**, **transport network**, **EU programme**, **sustainable development**, **project of common interest**, **energy grid**, **trans-European network**	**investment**, **project of common interest**, **energy grid**, **trans-European network**	3
	reduction of gas emissions, renewable energy, **EU financial instrument**, **structural policy**, **transmission network**, **transport network**, **EU programme**, **sustainable development**	2
	energy cooperation, consumer information, financial occupation, insurance, investment company, disclosure of information, financial legislation, financial services, risk management, financial risk, energy policy, investment promotion, emission trading, climate change, greenhouse gas, transition economy, climate change policy, EU energy policy, security of supply, electricity supply, gas supply, electrical energy	1
Subject matter descriptors		
Trans-European network	**Trans-European network**, Energy	2
	Investments, Free movement of capital, Environment, Economic policy, Trans-European networks	1
rdf type descriptors		
Work, **Legal resource**	**Work**	10
	Legal resource	4
	Secondary legislation, Consolidated act	2
	Other act of the Council	1
Resource type descriptors		
Regulation	Proposal for a regulation, Consolidated text	2
	Regulation, Communication, Legislative resolution, Roadmap, Proposal for an act, Note[a]	1

[a] Some proposed concepts might be irrelevant to a document because the search space of similar documents is not adjusted for a specific document but all available documents are reused.

Table 1: Labels of various properties of CELEX:32021R1153 document

5.4 Embedding model and metadata properties

To compute the embedding for a CELLAR Document, we use the *all-distilroberta-v1* model. This model, a Sentence Transformer model, maps sentences and paragraphs to a 768-dimensional dense vector space. It is effectively utilized for tasks such as clustering or semantic search.

To establish the context, we utilized the first 5,000 characters from the beginning of each document.

$$C(d) = \text{first}(5000, d)$$

Therefore, our embedding space is defined as follows:

$$E_{all-distilroberta-v1} : C \rightarrow \mathbb{R}^{768}$$

5.5 Metadata inference from similar documents

To validate our hypothesis, we conducted a series of experiments based on English documents from CELLAR. We utilized a snapshot of documents up to the year 2019, which includes more than 500,000 documents along with their associated metadata.

In our initial hypothesis validation experiment, we aim to evaluate recall only which is more important than precision in cases such as automated annotation process. A more detailed results using F1-score is presented in Section 5.6.4.

– We randomly selected a set of 1,000 documents from the CELLAR repository, all of which already have their associated metadata.
– For each document, we identified the first 10 most similar documents (for K-NN, $k = 10$), the 10 least similar documents (ranked 91-100) returned by the metadata recommender, and 10 randomly chosen documents for comparison.
– We then verified the presence of any metadata in the selected documents. To ensure accuracy, the original document was always excluded from the list of similar documents and never appeared in the selection of 10 similar documents. We introduced a hyper-parameter, L, as an experimental parameter to filter metadata values based on the frequency of their occurrence in similar documents. For instance, if $L = 1$, the metadata value must appear at least once among the metadata of similar documents; if $L = 10$, the metadata value must appear in all 10 similar documents.

This experiment was iterated three times, with results averaged to assess the overall efficacy of the the metadata recommendation process.

The results of the experiments show (see Figure 2) that we were able to retrieve significant amount of related metadata for various metadata properties such as *EuroVoc* (Fig. 2a) with 60%, *Theme* (Fig. 2b) with 70%, and *Subject-matter* (Fig. 2c) with 25% recall. However, for *rdf:type* (Fig. 2d) and *resource-type* (Fig. 2e) results are similar in all selected subsets (most/less/random). This is due to the fact that the distribution of these property values are not even. The L parameter determines how many concepts are selected. The lower is L value the more concepts are selected. It results in higher recall which is important in use cases of automated annotation when human annotators are selecting from the narrow list of candidates instead of using full list of concepts. Nevertheless, the human annotator can adjust L value at any time which brings high flexibility for annotators.

5.6 Use-Case: Document Annotation

After validating the hypothesis that metadata could be inferred from similar documents, we conducted a second experiment.

5.6.1 Deployed prototype and collected feedback

We deployed a prototype that implements our approach in a real-world annotation system scenario. Cataloguers have access to a prototype application where they can upload documents and receive metadata recommendations. We have collected usage feedback to establish a benchmark for comparison. This has enabled us to directly compare the performance and effectiveness of our approach with existing annotation systems. Figure 3 displays the user interface of our annotation application following the submission of a document. Annotation candidates are displayed in a table, allowing users to filter the results and select the most relevant ones. Users can also provide feedback to assess the quality of the returned results, which is subsequently used for comparative analysis.

In total, 967 documents were submitted for annotation and feedback was collected.

5.6.2 Evaluation Metrics

Our evaluation framework employed several metrics to measure the performance of our document annotator in comparison to the aforementioned tools. Here is a summary of the metrics:

- **Precision (Average):** Measures the accuracy of the selected annotations, indicating how many are relevant.
- **Recall (Average):** Assesses the tool's ability to identify all relevant annotations within the documents.
- **F1 Score (Average):** Provides a balance between precision and recall, offering a single score that measures overall accuracy.
- **Micro F1 Score:** Aggregates the contributions of all classes to compute the average F1 score, reflecting overall classification performance.
- **NDCG Score (Average):** Evaluates the ranking quality of the annotations by measuring the grading consistency of recommended tags. The value of NDCG is determined by comparing the relevance of the items returned by the search engine to the relevance of the item that a hypothetical "ideal" search engine would return(Järvelin and Kekäläinen, 2002).

5.6.3 Other annotation tools

This evaluation focuses on comparing our document annotator with two other tools.

- **Annif** (Suominen, 2019): Annif is an open-source toolkit designed for automated subject indexing using a variety of machine learning and AI-based algorithms for efficient text classification. Our approach is compared to an existing deployment of Annif, available at the Open Data Portal of the Publications Office of the European Commission[6].
- Eurovoc classifier based on EUBERT[7]: EUBERT is a pretrained BERT model that utilizes the vast corpus of documents from the European Publications Office. It is specifically tailored for tasks like text classification, question answering, and language understanding. The classification model is built on top of EUBERT with 7331 Eurovoc labels.

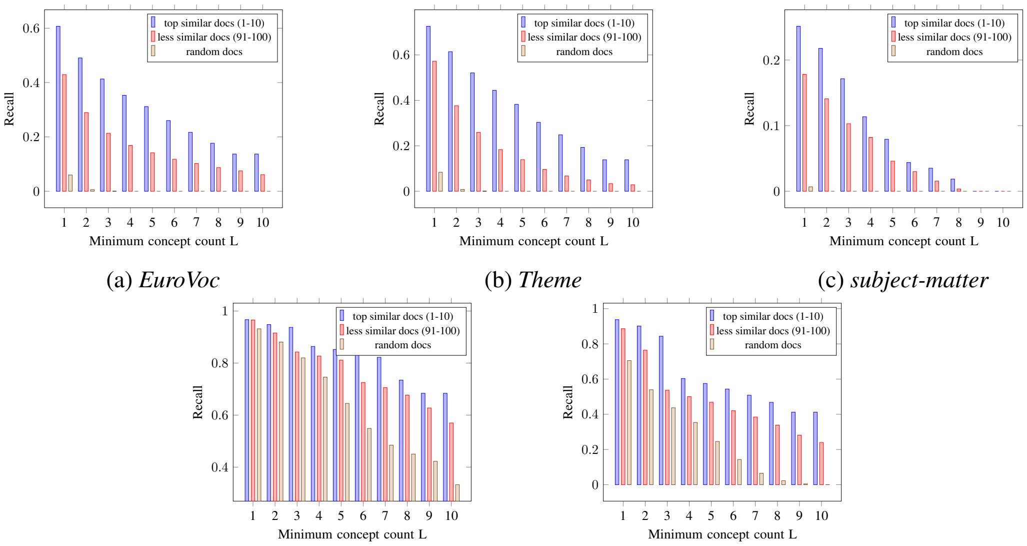
Since the compared tools recommend only Eurovoc metadata, we limit evaluated properties to only Eurovoc thesaurus.

5.6.4 Comparison and discussion

The feedback and quantitative metrics indicate that our prototype (CELLAR Annotator) surpasses both Annif and the Eurovoc classifier in terms of the F1 score across various top k values. The selection of top k values for the CELLAR annotator is based on concept score evaluation described in Section 4.4.3 and using value L=1.

[6] https://data.europa.eu/annif

[7] github.com/racai-ai/pyeurovoc



(a) *EuroVoc*

(b) *Theme*

(c) *subject-matter*

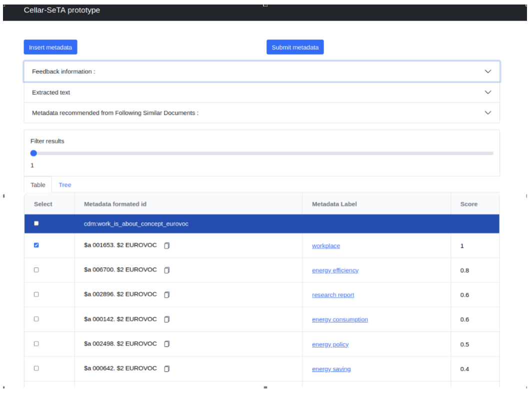
(d) *rdf:type*

(e) *resource-type*

Figure 2: Metadata annotation results. X-axis is minimum concept frequency, y-axis is recall

As shown in Figure 4 (top left), the precision of Cellar Annotator consistently outperformed both Annif and EUBERT across all values of k. This indicates that Cellar Annotator has a higher accuracy in predicting relevant annotations, ensuring that the annotations provided are relevant and accurate. Specifically, the precision for Cellar Annotator remained above 0.9 for all values of $k <= 7$, highlighting its reliability in maintaining high precision even as the number of considered annotations increased.

In terms of recall (Figure 4, top right), Cellar Annotator significantly surpassed both Annif and EUBERT. This suggests that Cellar Annotator is more effective in retrieving all relevant annotations, thereby reducing the number of missed annotations. The recall values for Cellar Annotator consistently stayed above 0.55, while the other methods showed more variability and generally lower recall rates. This demonstrates the robust capability of Cellar Annotator to identify and recall relevant annotations comprehensively.



Select	Metadata formated id	Metadata Label	Score
■	cdm:work_is_about_concept_eurovoc		
☑	$a 001653. $2 EUROVOC	workplace	1
□	$a 006700. $2 EUROVOC	energy efficiency	0.8
□	$a 002896. $2 EUROVOC	research report	0.6
□	$a 000142. $2 EUROVOC	energy consumption	0.6
□	$a 002498. $2 EUROVOC	energy policy	0.5
□	$a 000642. $2 EUROVOC	energy saving	0.4

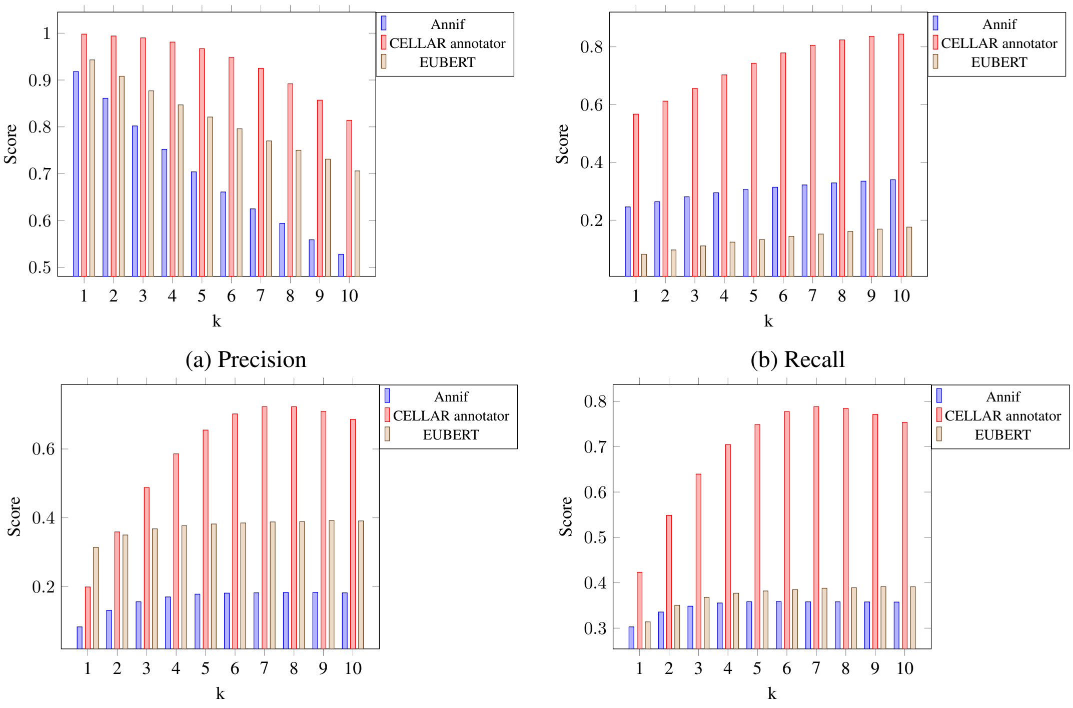
Figure 3: Metadata recommendation prototype

The micro F1 score, which balances precision and recall, further confirmed the superiority of Cellar Annotator (Figure 5, bottom left). The scores for Cellar Annotator were consistently higher, indicating a balanced performance in terms of both precision and recall. The micro F1 scores remained around 0.7 for Cellar Annotator, whereas Annif and EUBERT showed lower and more fluctuating scores. This balanced performance is crucial for applications where both high precision and recall are essential.

Finally, the NDCG scores (Figure 5, bottom right) demonstrated that Cellar Annotator also excels in ranking the most relevant annotations higher. With NDCG scores consistently around 0.75, Cellar Annotator ensures that the most pertinent annotations are prioritized, enhancing the overall utility and effectiveness of the annotation system. This metric is particularly important for user-facing applications where the relevance of top-ranked annotations significantly impacts user experience and satisfaction.

6 Conclusion and perspectives

In this paper, we present a novel method for multi-property multi-label document classification that leverages an encoder embedding-driven approach. Our technique aims to streamline the document an-



(a) Precision

(b) Recall

(c) Micro F1-score

(d) NDCG

Figure 4: Precision, Recall, micro F1-score and NDCG results

notation process by utilizing semantic text similarity and the reuse of annotated data. This approach reduces the complexity associated with deploying multiple models, as it relies on a single model to assess text similarity, which results in enhanced efficiency compared to traditional classification methods.

The practical implementation of our prototype within the European Commission has yielded promising results. Empirical results show that our method surpasses the performance of existing systems, delivering superior accuracy and operational efficiency in practical settings.

For future work, we aim to assess additional state-of-the-art embedding models to further refine our approach. We also plan to expand our methodology by incorporating graph-based semantic similarity measures.

Limitations

One main limitation of the metadata replication approach is the sole focus on the reuse of the metadata data that has been used in the past. This may magnify labels related to the past manual captured metadata. All the possible values from the vocabularies are not necessarily present in the metadata regardless of their usefulness. This also sheds the light on an other limitation regarding the already used metadata: the distribution of the reuse of the values. The human bias induced in the manual metadata annotation could have an impact on the quality of the recommendations and should be further investigated and kept in mind for the industrialisation of this approach.

To address these limitations, we plan to introduce an option for exact matching of classes. This will facilitate the identification of new classes that have not been previously used for annotations, thereby expanding the scope and effectiveness of our metadata recommendations.

One significant constraint of metadata replication lies in its exclusive reliance on previously utilized metadata, potentially perpetuating biases linked to past manual annotations. Such an approach does not guarantee the inclusion of all valuable terms from controlled vocabularies, as not all possible values may be represented within the existing metadata. This limitation underscores another issue concerning the frequency of value reuse in metadata: the influence of human bias during manual annotation could affect the quality of generated recommendations, which merits closer examination and consideration during the process of operationalizing this methodology.

To mitigate these issues, we are proposing the integration of an exact matching feature for class identification. This enhancement aims to uncover novel classes that have not been employed in prior annotations, thus broadening the reach and improving the efficacy of our metadata recommendation system.

References

Andrei-Marius Avram, Vasile Pais, and Dan Ioan Tufis. 2021. PyEuroVoc: A tool for multilingual legal document classification with EuroVoc descriptors. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 92–101, Held Online. INCOMA Ltd.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy. Association for Computational Linguistics.

Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S. Dhillon. 2020. Taming pretrained transformers for extreme multi-label text classification. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3163–3171, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Enrico Francesconi, Marc W. Küster, Patrick Gratz, and Sebastian Thelen. 2015. The ontology-based approach of the publications office of the eu for document accessibility and open data services. In *Electronic Government and the Information Systems Perspective*, pages 29–39, Cham. Springer International Publishing.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20:422–446.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Dezhao Song, Andrew Vold, Kanika Madan, and Frank Schilder. 2022. Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training. *Information Systems*, 106:101718.

Osma Suominen. 2019. Annif: Diy automated subject indexing using multiple algorithms. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, 29(1):1–25.

Lulu Wan, George Papageorgiou, Michael Seddon, and Mirko Bernardoni. 2019. Long-length legal document classification. *CoRR*, abs/1912.06905.