Multilingual Automated Subject Indexing: a Comparative Study of LLMs vs Alternative Approaches in the Context of the EHRI Project

Maria Dermentzi [1], Mike Bryant [1, 2], Fabio Rovigo [3], and Herminio García-González [4]

[1]King's College London, UK
[2]NIOD Institute for War, Holocaust and Genocide Studies, NL
[3]Vienna Wiesenthal Institute for Holocaust Studies, AT
[4]Kazerne Dossin, BE

The European Holocaust Research Infrastructure (EHRI) facilitates transnational Holocaust research by making information about dispersed archival sources accessible through the EHRI Portal. An important aspect of effectively integrating this information is the indexing of collection metadata from institutions worldwide based on a common domain-specific controlled vocabulary (EHRI Terms). However, challenges persist in harmonising the use of subject headings across institutions. Hence, this paper explores approaches for Automated Subject Indexing (ASI), including statistical, lexical, and fusion approaches, as well as the use of Large Language Models (LLMs) for zero-shot classification. We describe how we use archival descriptions ingested into the EHRI Portal as training data for Machine Learning (ML) algorithms, including for fine-tuning an LLM. We evaluate LLM- and non-LLM-based tools quantitatively and qualitatively, with our focus being to explore if and to what extent LLMs can play a role in developing reliable ASI tools. We conclude that although some tools under evaluation score quite highly, they are at present best suited for use as part of semi-ASI workflows with human oversight. However, this study charts a promising course towards developing tools that are reliable enough to facilitate subject indexing processes and enhance metadata according to the FAIR Data Principles.

Keywords: Automated Subject Indexing, Natural Language Processing (NLP), Multilabel Text Classification, Holocaust, Controlled Vocabularies, Archives

1 Introduction

Since its launch in 2010, one of the principal goals of the European Holocaust Research Infrastructure (EHRI) has been to catalyse transnational Holocaust research by making information about dispersed archival materials more interconnected, coherent, and accessible. For researchers into the Holocaust, the landscape of archival sources is a highly complex one, for reasons including the deliberate destruction of evidence, dispersal of materials, and the migration of survivors (Blanke et al., 2014). This has resulted in material being fragmented, duplicated, and separately described by a multitude of institutions with different mandates, cataloguing practices, and approaches to curation. The EHRI Portal is an attempt to integrate archival metadata from these dispersed sources into a single framework, creating a virtual observatory within which fragmented collections can be connected and contextualised (Blanke et al., 2017).[1]

Aggregating and integrating archival descriptions from hundreds of institutions around the world poses considerable institutional, social, and technical challenges. As previously observed (Erez et al., 2020; García-González and Bryant, 2023; Rodriguez et al., 2016), in the context of Holocaust-related archives, different institutions take a wide range of approaches to describing their materials, even putting aside those differences imposed by factors such as their specific collection management systems and institutional environments. The use of index terms — controlled vocabularies of subject headings, people, organisations, and places — is a cornerstone technique in enhancing the discovery and retrieval of archival material. Yet across the hundreds of institutions holding Holocaust-related material, or specialising in the field, there is negligible commonality in the application of index terms or the use of common controlled vocabularies. In most cases where index terms *are* used, an in-house vocabulary is the typical approach. While a small minority of institutions do use general-purpose subject heading thesauri such as the Library of Congress Subject Headings (LCSH)[2], that does not necessarily mean that they interpret and apply the index terms to their material in consistent or interoperable ways (Erez et al., 2020).

This lack of a common vocabulary for Holocaust-related material was one of the problems EHRI set out to address in its first phase (2010-2014) with the creation of the *EHRI Terms* vocabulary, a hierarchically organised, multilingual set of subject headings. While the starting point for EHRI Terms was a controlled vocabulary in use at Yad Vashem (Erez et al., 2020),[3] it has since undergone considerable expansion and restructuring in both the terms and available languages, and at the time of writing consists of 913 terms translated in 12 languages.[4] The vocabulary is used by EHRI's own subject-matter experts when creating new descriptions for as-yet uncatalogued material, and as an integration and interlinking point when ingesting into the EHRI Portal material described by EHRI's many partner institutions. It is also available in the Resource Description Framework (RDF) in Simple Knowledge Organization System (SKOS) format, along with two other controlled vocabularies focusing on World War II *ghettos* and *camps*.[5]

While EHRI has built a robust data integration infrastructure, improving the discoverability and interconnectedness of multilingual collection metadata, sourced from

---

[1] `https://portal.ehri-project.eu/`.

[2] `https://www.loc.gov/aba/publications/FreeLCSH/LCSH44-Main-intro.pdf`

[3] `https://www.yadvashem.org/`

[4] English, Hebrew, Italian, Dutch, Russian, Ukrainian, Czech, Hungarian, French, Polish, Serbo-Croatian, and German. As of 24 January 2025.

[5] `https://portal.ehri-project.eu/vocabularies/`

multiple institutions, remains key objective for the project. The harmonisation of subject headings, using the EHRI Terms vocabulary as an integration point, is one particular area where there is room for improvement, not only because much of the aggregated metadata does not currently align with it but because many archival collections lack index terms entirely. At present, 25% of collection-level descriptions have no index terms at all.[6] Of those that *do* have index terms, only 30% have subject terms aligned with the EHRI Terms vocabulary.[7]

Past work has described how vocabularies in use by a selection of partner institutions were systematically co-referenced to EHRI Terms (Erez et al., 2020). This paper, however, investigates whether archival descriptions in the EHRI Portal could be made more discoverable and interlinked *at scale* through Automated Subject Indexing (ASI), i.e., the use of machine-based methods, such as computational linguistics and statistics, to perform the subject indexing steps typically performed by human indexers (Golub, 2021). The findings of this investigation are applicable beyond EHRI and its partners, as they could help other institutions address similar challenges in their specific contexts.[8] Experimenting with ASI is also part of a larger effort to integrate automated metadata generation techniques into EHRI's workflows. Other techniques we have been exploring include multilingual Named Entity Recognition (Dermentzi and Scheithauer, 2024) and Automatic Speech Recognition (Wynne, 2023). Such efforts fit into the broader context of cultural heritage institutions viewing their collections as data that can be used to experiment with and develop innovative digital tools as described in Candela et al. (2023) and Candela et al. (2024).

The rest of this paper is structured as follows: in Section 2 we provide an overview of ASI and the related work in this area; in Section 3 we describe in more detail the EHRI Terms vocabulary and a corpus of documents derived from metadata in the EHRI Portal; in Section 4 we present our experiments, the different tools under evaluation, the quantitative and qualitative methodology and the metrics used; in Section 5 we present the results obtained from our evaluation, followed in Section 6 by a discussion of various issues identified in this work. Finally, in Section 7 we draw the future lines of work based on the obtained results and in Section 8 we present our conclusions.

2 Automated Subject Indexing

Approaches to ASI vary based on the purpose of the application and the field from which it originates (Golub, 2021). The two prevailing approaches are the statistical/associative and lexical techniques respectively (Suominen and Koskenniemi, 2022; Toepfer and Seifert, 2020).[9] The statistical associative approach involves Multi-label Text Classification (MTC) methods, where a supervised Machine Learning (ML) model is trained on the already indexed texts of a collection. The model learns weights that are used to predict the correct set of terms given the text of a document. Lexical approaches, on the other hand, employ string-matching to match terms in the controlled vocabulary with words in the text of the record's description using similarity measures (Golub, 2021). Recent research has also suggested fusion approaches that

6 Accessed 29th Jan 2024: `https://portal.ehri-project.eu/api/datasets/E136TY2zwL`

7 Accessed 29th Jan 2024: `https://portal.ehri-project.eu/api/datasets/CtaJXtPuZA`

8 This is especially the case given EHRI's use of archival standards, e.g., General International Standard Archival Description (ISAD(G)), and the fact that the tools used by the authors are open-source, meaning that others can reuse and adapt them for their use cases.

9 Lexical approaches are also known as string-matching or rule-based approaches (Golub, 2021).

combine statistical and lexical methods using ensemble techniques (Suominen and Koskenniemi, 2022; Toepfer and Seifert, 2020).

More recently, the emergence of Large Language Models (LLMs) has enabled a novel approach: transfer learning using zero-shot or few-shot classification (Chow et al., 2024). A general-purpose pre-trained LLM can be used to predict suitable labels from a list of candidate terms (derived from a controlled vocabulary) without needing fine-tuning on domain-specific data (Zhang et al., 2023, p. 3). Presently, the zero-shot classification approach (where the model is made to return predictions without having been provided with any examples of its task) is commonly used as a data augmentation method to overcome class imbalance or the lack of training examples (Møller et al., 2023; Van Nooten and Daelemans, 2023). Zhang et al. (2023) and Chow et al. (2024) have experimented with subject indexing using few-shot learning (meaning the instructions provided to the model included a few examples of the task along with the list of the candidate subject terms).

While EHRI has not so far deployed ASI, it has used lexical techniques in its Entity Matching Tool (EMT) for co-referencing index terms from third-party vocabularies to EHRI's terms (or to GeoNames in the case of place names).[10] If the third-party vocabulary used to index an archival description uses terms that are morphologically very similar to the EHRI Terms, and in a language included within the EHRI Terms vocabulary, this method can return accurate results. This tool, however, was intended for assisting keyword-to-keyword alignment or aligning a set of discrete textual references, and not for generating keywords from the running text of an archival description. It does not consider the semantics of specific terms, and cannot assist with terms in languages other than those included in the target vocabulary.

3 Dataset

This section characterises our dataset — derived from archival descriptions within the EHRI Portal — and the EHRI Terms vocabulary that comprises our label set. As an aggregator of archival metadata, EHRI integrates descriptions of Holocaust-related archival material from a wide range of collection holders around the world. These range from small local archives and museums with very limited digital capabilities, to large and comparatively well-resourced institutions dedicated to memorialising the Holocaust. Among the latter include prodigious aggregators of physical archival material, such as Yad Vashem and the United States Holocaust Memorial Museum (USHMM)[11], who have amassed extensive collections of material sourced from other institutions and described using their own in-house standards and styles. A more in-depth explanation on how data integration takes place within EHRI's current phase can be consulted in (Vanden Daelen et al., 2024).

As noted in the introduction, diversity and complex provenance are notable characteristics of the archival landscape relating to the Holocaust. One result of this is that some important collections do not yet have any metadata available in digital form. Since 2010 EHRI has sought to mitigate this by providing its own English-language descriptions of such collections in the EHRI Portal. Thus, the Portal contains not just metadata sourced from third parties, but original descriptions authored by subject matter experts following EHRI's ISAD(G)-based descriptive standard and indexed against EHRI's vocabularies, including EHRI Terms.

10 `https://emt.ehri-project.eu/`

11 `https://www.ushmm.org/`

EHRI Terms[12] is structured as a Directional Acyclic Graph (DAG) with 15 top-level terms ("top concepts" in SKOS parlance[13]). The hierarchical structure has a maximum depth of 10 from broadest to narrowest, but the median depth is shallower at just 3 terms. The vocabulary is, additionally, not strictly hierarchical; many terms have multiple "broader" concepts, for example, the term "Healthcare" has three broader terms ("State Functions", "Aid, Welfare, Rescue", and "Daily Life"). Overall, 291 of 913 subjects fit within more than one higher-level category.

To construct a training and test set from structured metadata in the EHRI Portal we were restricted to the subset of archival descriptions which had already been assigned subject headings, either manually (by EHRI's cataloguers) or via the co-referencing process described in Erez et al. (2020). This resulted in 48,854 individual archival units at all levels of description (from collection to item level). We then transformed each description into a single text by concatenating the fields considered most likely to capture the topical content of the material, as opposed to administrative details. The fields selected were ISAD(G) elements 3.1.2. Title; 3.2.2. Administrative / Biographical History; 3.2.3. Archival History; and 3.3.1. Scope and Content. We acknowledge that there is inevitably some imprecision in the selection of these fields, in terms of relevancy.

This dataset contained archival descriptions in 17 different languages. Polish and Finnish were represented in the dataset by only one example each. To retain these two descriptions in the dataset before splitting it into training and test sets would mean that descriptions in Polish and Finnish could only ever be present either in the training or in the test set. To avoid training on only one example of these two languages, or testing on them without having seen any in the same language during training, we decided to remove these descriptions from our dataset, leaving us with 48,852 entries. After de-duplicating our dataset to remove descriptions of the same archival units in different languages, we were left with 36,801 archival units.

In common with many multi-label classification tasks, particularly those featuring a large number of labels, class imbalance was a significant feature of our dataset. After some exploratory data analysis, we noticed that the most populous term, "Photographs", was represented by 4,819 positive examples, while some of the least popular terms, like "Jewish religious leaders and functionaries", were represented by only one example. Given the hierarchical nature of the label set, it is unsurprising that certain broader terms, especially terms pertaining to the genre of the material, such as "Photographs" or "Letters" were more frequently used than highly specific ones (e.g., "Activists in justice against criminals"). In fact, 215 terms had fewer than five positive examples in the dataset. To ensure better label distribution across training and test datasets, we removed terms with fewer than five examples. As a result, some descriptions matched to only one subject term lost their positive label and were also removed. The final dataset included 36,759 archival descriptions. Of the 913 subject terms in the EHRI vocabulary, only 554 met our threshold of at least five examples per term.

As in related work (Skenderi et al., 2021), to obtain a training and a test dataset we used the iterative multi-label stratification method proposed by Sechidis et al. (2011) as implemented in the Python package `iterative-stratification` (Bradberry, 2021). We opted for this method of stratification to increase the chances that all labels (as well as all languages) – no matter how rare in the dataset – would still have positive

12 https://portal.ehri-project.eu/vocabularies/ehri_terms
13 https://www.w3.org/TR/skos-reference/#L2446

examples in both the training and test set. This was crucial to ensure that the evaluation of our model on the test set would be as complete as possible, covering the entire label space. It would also reduce the probability that certain multi-label classification metrics would be impossible to calculate due to the lack of positive examples as explained in Sechidis et al. (2011).

Specifically, we performed an initial multi-label stratified shuffle split of the dataset to obtain one split, where the test set size would include 30 per cent of the dataset (11,027 archival descriptions) and the train set would include the remaining 70 per cent of the dataset, i.e. 25,732 archival descriptions. The test split obtained through this process was further stratified/split once more in the same manner to obtain a third, considerably smaller, test set for evaluation purposes, which included 167 archival descriptions out of the 11,027 descriptions obtained through the first split. Since we wanted to perform both quantitative and qualitative evaluations to compare the different models used in our experiments, it was essential to have a manageable yet representative final test set that would be feasible for human evaluators to process in only a few working hours. Moreover, since we wanted to experiment with at least one LLM for zero-shot classification, obtaining predictions for a very large number of archival descriptions using 554 candidate labels would have been very costly and time-consuming to run on locally available hardware. While the larger test set obtained through the second splitting process (comprising 10,860 archival descriptions) was used to estimate model performance while training and fine-tuning new models, the smaller test set (167 archival descriptions), henceforth called the "evaluation" set to avoid confusion, was reserved for the final quantitative and qualitative evaluation and comparison of all the models in our experiment and was not used during training or fine-tuning. However, due to the very small size of the evaluation set and despite following the iterative stratification method for obtaining it, only the 134 most frequent labels had positive examples within it, while the remaining 420 did not. Similarly, although the test set used during training included examples in all 15 languages present in the training set, the evaluation set included examples only in the nine most frequent languages of our dataset,[14] represented proportionally based on their distribution within our dataset. While performing these splits, a random state seed was used to make our experiment reproducible.

The dataset has been made openly available following best practices on publishing cultural heritage datasets (Alkemade et al., 2023; Candela et al., 2023).[15]

4 Experiments

This section describes how we selected the tools we put to the test: a transformer-based zero-shot classifier, a transformer-based model for fine-tuning, and other classical and non-LLM approaches. Key considerations in selecting the transformer-based models were support for multilingual text in languages relevant to EHRI's material, being open source and easily accessible, having well-documented training data, and giving us the ability to quantitatively evaluate their performance using our in-house compute infrastructure.

14 English, Hebrew, Dutch, Czech, German, Multiple (used for descriptions containing text in more languages than one), French, Russian, Italian.

15 `https://doi.org/10.5281/zenodo.14697253`

4.1 Zero-shot Classification Approach

The first approach, zero-shot classification, is the labelling of input data without the use of any prior context, e.g. labelled examples. In the context of LLMs it relies on the model having sufficient training data to have encountered related examples in its pre-training phase and to be able to make suitable inferences when given new data to classify. The purpose of this approach was to help us answer the question of whether using an LLM out of the box for classification purposes could be sufficiently reliable to be used as part of an automated or semi-ASI process or, alternatively, whether it could be used to augment our labelled dataset in order to reduce our class imbalance problem.

While at the time of writing the most advanced generally available LLM is OpenAI's GPT4 (OpenAI, 2023), we decided to omit this from the comparison at this stage due to the fact that it is not open source and due to the lack of transparency surrounding its development and training data. For practical reasons, we were also unable to include some newer LLMs capable of being self-hosted, such as TII's Falcon [16] and Meta's Llama2[17], both of which have exceedingly demanding inference requirements.

The model we selected for the zero-shot approach was mDeBERTa-v3-base-mnli-xnli (Laurer et al., 2023) available through Hugging Face (HF) which uses mDeBERTa v3 (He et al., 2023) as base model. mDeBERTa v3 is a multilingual model pre-trained on the CC100 multilingual dataset covering 100 languages (including all of the languages in our dataset) (He et al., 2023). mDeBERTa-v3-base-mnli-xnli is a fine-tuned version of mDeBERTa v3 for Natural Language Inference (NLI), trained using the XNLI (Conneau et al., 2018) and MNLI (Williams et al., 2018) datasets. These datasets include texts in 15 languages, but some of the languages that are present in our EHRI MTC dataset were not included in the datasets used to fine-tune mDeBERTa v3. However, since multilingual models for NLI are known for their cross-lingual transfer learning capabilities (Conneau et al., 2020; Zhang et al., 2023), the fine-tuned model can still classify texts in the rest of the languages covered in the dataset that was used to train the base model despite not having seen any examples of them during fine-tuning.

4.2 Fine-tuning Approach

The second approach was to fine-tune a pre-trained LLM on EHRI's data, thus giving us some insight into whether the already-indexed archival descriptions can be harnessed to train models capable of performing effective classification of the remaining material. Although there is no shortage of pre-trained multilingual Transformer models to fine-tune, we decided to select BERT-base, Multilingual Cased (Devlin et al., 2019). In addition to being multilingual and open-source, we picked this model over more state-of-the-art (SOTA) alternatives for other important reasons. First, compared to more recent models it is much lighter and does not require significant resources to fine-tune. Second, it was trained on Wikipedia data rather than a CommonCrawl dataset, eliminating the chance of the model having already seen data in our training or test set (derived from either the EHRI Portal itself or EHRI-affiliated institutions). Third, as the original transformer implementation, it has gained a lot of popularity and is still widely used for tasks like text classification. It is also often used as a baseline with which to compare more recent models, helping us determine whether spending

16 `https://huggingface.co/tiiuae/falcon-180B`

17 `https://ai.meta.com/llama/`

more resources to fine-tune SOTA models would be something worth pursuing.

As suggested by the authors in the instructions accompanying the GitHub repository of the BERT family of models (Devlin et al., 2023),[18] the "cased" version of the model is suggested for use in multilingual settings, especially with non-Latin alphabets, since it does less normalisation of the input. According to the authors, the data that was used to train the model comprised the entire Wikipedia dump for each of the top 100 languages with the largest Wikipedia instances.[19] This list of languages includes all of those present in our train, test and evaluation sets. The transformers library (Wolf et al., 2020) developed by HF, as well as tutorials developed by the HF community (Rogge, 2021), provided an accessible way to proceed with fine-tuning Transformer models for multi-label classification. Underneath the hood, fine-tuning the pre-trained base BERT model on our EHRI dataset for MTC adds an extra linear layer on top of BERT's final layer, which is trained to output scores that indicate how likely it is for each label to be the correct label for each text in the input.

To fine-tune BERT, we used a batch size of 32 and trained for 25 epochs, selecting the model with the best micro-averaged F1 score. We set the learning rate to 2e-5. To avoid overfitting on the training set, we set the hidden dropout probability hyperparameter to 0.2 when instantiating the model and the weight decay hyperparameter to 0.01 when specifying the training arguments. The best micro-averaged F1 score on the large test set (≈ 0.56) was achieved by the model obtained in the final epoch.[20]

4.3 Other Approaches

As a framework for quantitative evaluation we chose to employ Annif (Suominen, 2019; Suominen and Koskenniemi, 2022; Suominen et al., 2022, 2023), a tool for multilingual subject indexing already used by a number of cultural heritage institutions. Annif has been designed to enable the use of different multi-label classification backends tailored for specific types of material. It includes implementations of classical indexing algorithms like TF-IDF, as well as state-of-the-art approaches such as fastText (Joulin et al., 2017), and Parabel (Prabhu et al., 2018). Annif is written in Python and can be extended with additional backends by implementing specific subclasses. These new backends can then leverage Annif's evaluation tools for generating metrics such as precision, recall, and F1 score (per-document, as well as using macro or micro averaging) along with ranked metrics such as non-discounted cumulative gain (NDCG).

In addition to the LLM-based approaches detailed in Section 4.1 and 4.2, we also evaluated a selection of Annif backends in order to provide a performance comparison.[21] These included:

TF-IDF A baseline method capable of fast, resource-efficient labelling.

MLLM Maui-like Lexical Matching, inspired by Medelyan (2009).

fastText A fastText classifier.

---

18 `https://github.com/google-research/bert/blob/master/multilingual.md`

19 See `https://github.com/google-research/bert/blob/master/multilingual.md#data-source-and-sampling`

20 EHRI's fine-tuned BERT-based model is available on Hugging Face (`https://huggingface.co/mdermentzi/finetuned-bert-base-multilingual-cased-ehri-terms`) and Zenodo (`https://doi.org/10.5281/zenodo.14680007`)

21 More details about these backends can be found in the Annif wiki page: `https://github.com/NatLibFi/Annif/wiki`

Omikuji A Parabel implementation as an example of a state-of-the-art non-LLM method.

NN Ensemble A neural-network-based method of dynamically combining the outputs of several other discrete classification tools, in this case TF-IDF, MLLM, fastText, and Omikuji.

While it is capable of being configured for use in a multilingual setting, Annif does not currently have explicit support for heterogeneous corpora, e.g. those containing documents in a variety of languages. Since testing on our predominantly English evaluation dataset showed minimal impact on performance, however, we did not seek to address this limitation at this stage, noting that it could be worthwhile to explore as future work.

We created custom Annif backends for our fine-tuned BERT model and our zero-shot model of choice run via the HF transformers library, allowing us to deploy them and evaluate their performance in a consistent manner alongside Annif's native backends. For the purpose of this experiment we only implemented the inference/suggestion functionality, however as future work the fine-tuning process could also be reimplemented to fit within Annif's backend framework, including training and hyper-parameter optimisation.[22]

4.4 Evaluation Metrics

We decided to evaluate the results of our experiments both quantitatively and qualitatively. Properly evaluating subject terms and consequently subject indexing tools is a very challenging task for a plethora of reasons (Golub et al., 2016). First and foremost, assigning and evaluating subject terms is a very subjective process and a term that may be considered pertinent by some may not be perceived as such by others. In fact, several inter-indexer consistency studies have proven that consistency levels between different indexers are typically very low (Tonta, 1991; Vaughan and Rafferty, 2011). Moreover, depending on the target audience of an institution, narrower and more specific subject terms may be preferred over broader ones, while other institutions may prioritise indexing records with as many terms as potentially useful rather than only a few specific ones (Golub et al., 2016). Ultimately, people browsing a catalogue to search for information on a specific topic will perceive a subject term to be correctly assigned only if it leads them to records relevant to their specific needs. Ideally, the evaluation of ASI tools should be based on how successful they are at helping users find answers to their questions. However, performing this kind of evaluation is extremely difficult and due to the subjective nature of the task may still lead to inconclusive results.

Usually, already indexed records are used as gold standard datasets based on which different tools can be evaluated quantitatively (Ahmed et al., 2023; Golub, 2021; Golub et al., 2016; Suominen et al., 2022). However, since subject indexing is a subjective task that is influenced by many factors (level of indexer's expertise, institutional priorities, length of the vocabulary, etc.) (Golub, 2021; Golub et al., 2016), it is very hard to develop a truly comprehensive gold standard dataset on any real-life collection indexed with a sizeable enough controlled vocabulary. While we acknowledge this limitation, for the purposes of the quantitative evaluation herein we assume that the terms

[22] A branch containing modified Annif code is available on Github (`https://github.com/EHRI/Annif/tree/ehri_masi`) and Zenodo (`https://doi.org/10.5281/zenodo.14697325`).

assigned to EHRI-ingested records through the methods described in the previous sections are complete and correct (Golub et al., 2016). As explained by Suominen et al. (2022), treating parts of already indexed collections as gold standard collections allows for quick and easy experimentation and evaluation of different algorithms. However, the authors agree with Suominen et al. (2022) that this method is best seen as a "ballpark estimate of quality" and other methods of evaluation should also be considered.

Similar to previous work (Ahmed et al., 2023; Asula et al., 2021; Chou and Chu, 2022; Gárdos et al., 2023; Skenderi et al., 2021; Toepfer and Seifert, 2020), to quantitatively evaluate the different multi-label classification methods we computed precision, recall and F1 scores. We computed these scores in three different ways: micro-averaged (taking into account total true positives, false negatives and false positives), weighted macro-averaged (computing the metrics per label weighted by each label's number of true instances to account for class imbalance) and averaged per document. All of these scores range from zero to one, with one being the best possible score.

We have used as our primary metric the F1 score as provided by the Annif framework (Suominen et al., 2022) which considers the top five suggestions returned by the algorithm, averaged across all text samples. Since labels display some degree of imbalance due to the hierarchical nature of the vocabulary and preponderance of broader terms, other metrics display a degree of divergence from this figure and have also been given in Table 1. The F1 micro average, which considers true positive, false positive, and false negative values summed across all text samples, gives proportionally higher weight to performance on common labels. The F1 weighted macro score, on the other hand, averages individual scores across all labels and gives a better indication of performance on rare labels. Since labels are unordered we have not included NDCG or other ranked measures.

Since only 134 of the 554 subject terms we are considering in this experiment had positive examples in the evaluation set, our evaluation of these models only covers around 24 per cent of the labels that the models were trained (or–in the case of zero-shot classification–asked) to predict. We acknowledge that this is an important limitation but at this stage of our research, this evaluation is regarded more like a first soundness check before proceeding with a more thorough evaluation. Nevertheless, thanks to the stratification method that we followed, the evaluation set still contained sufficient labels with positive examples which helped us understand the overall strengths and weaknesses of the different models under evaluation.

5 Results

5.1 Quantitative evaluation

The results of the quantitative evaluation, in terms of F1 document average, weighted subject average (weighted macro average), and micro average are shown in Table 1 and Figure 1. Considering the F1 document average and F1 subject average scores, Annif's Neural Network ensemble classifier gave us the best results with 0.5884 and 0.5568 respectively. However, the fine-tuned BERT model achieved the highest micro average F1 score with 0.5916, and beat all other non-ensemble methods across all metrics, including Omikuji/Parabel, the best-performing non-ensemble classifier. This might be explained by the fact that during training the fine-tuned BERT model was optimised for better micro-averaged F1 scores.

The quantitative evaluation scores achieved by our zero-shot LLM model, mDeBERTA v3, were generally poor compared with other methods, with a document average F1 of just 0.1099, narrowly ahead of MLLM but behind other classification methods.

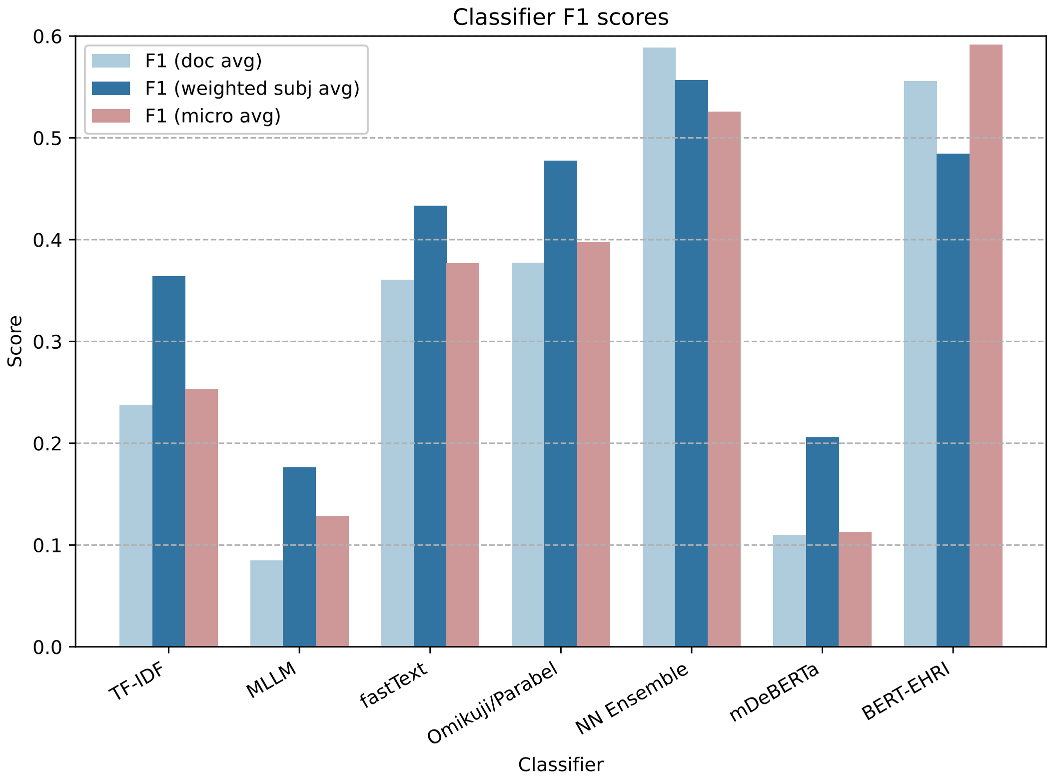

Figure 1: F1 scores for 7 classifiers under test

	F1 (doc avg)	F1 (weighted subj avg)	F1 (micro avg)
TF-IDF	0.2373	0.3640	0.2533
MLLM	0.0849	0.1762	0.1284
fastText	0.3606	0.4331	0.3768
Omikuji/Parabel	0.3771	0.4776	0.3975
NN Ensemble	**0.5884**	**0.5568**	0.5256
mDeBERTa	0.1099	0.2055	0.1126
BERT-EHRI	0.5556	0.4844	**0.5916**

Table 1: F1 document average, weighted subject average, and micro average scores for different models. The results in bold represent the best metrics per type of measure. The two horizontal grey lines delimit the three categories of tools under evaluation, namely the first group is composed of the Annif originally supported multi-label classification backends, the second group comprises the zero-shot LLM and the final group contains the fine-tuned LLM.

5.2 Qualitative Evaluation

Given the complexities of evaluating subject indexing tools as described in Section 4.4, we decided to complement our quantitative evaluation with a small-scale human evaluation. Complementing quantitative with qualitative evaluation is typical in similar studies (Chou and Chu, 2022; Suominen et al., 2022). Due to resource restrictions, we decided to limit the human evaluation to the best-performing Annif model and the two LLM-based models, since the focus of this study is on comparing LLMs against older methods that are less computationally intensive. The human evaluation was conducted by three of the authors of this paper who specialise in Digital Humanities and are familiar with EHRI's vocabularies and metadata under their capacity as research software developers or digital archivists working for EHRI.

We performed a subjective qualitative evaluation of four distinct label sets:

- the original labels assigned by EHRI staff or co-referenced with data from EHRI partner institutions,
- the NN Ensemble classifier, which performed best out of the Annif backends in our quantitative evaluation,
- the BERT-based model fine-tuned as described above,
- the mDeBERTa-based zero-shot model.

The evaluation was performed using a programmatically-generated spreadsheet (see Figure 2 for illustrative purposes) that displayed input texts alongside columns containing labels assigned by each tool, anonymised as Tool 1, 2, 3 and 4 and randomly ordered.[23] Beside each individual label were checkboxes to denote whether it was a favourable or unfavourable label in the opinion of the adjudicator, with the default state (both boxes unchecked) indicating that there was insufficient evidence on the basis of the input text to make a judgement. The undecided (neutral) state was particularly relevant in cases where the labels were assigned on the basis of external knowledge, e.g. direct knowledge of the underlying archival sources, rather than being an alternative view of descriptive text, a very common case in our dataset and an inherent limitation of deriving keywords from metadata that may be incomplete or otherwise brief. When performing the evaluation, the three judges were asked to rate the subject terms based on the following question: "If you used the subject term under consideration as an access point (either as a search term or by clicking on it to browse linked archival units), would you expect to see the archival unit that contains the corresponding input text in the list of results returned by the EHRI Portal?". For texts in languages that were not amongst the ones spoken by the evaluators, Google Translate was used.

[23] A Python script for generating this spreadsheet on Google Sheets can be found in the Annif repository containing our custom backends, named `annif-to-spreadsheet.py`.

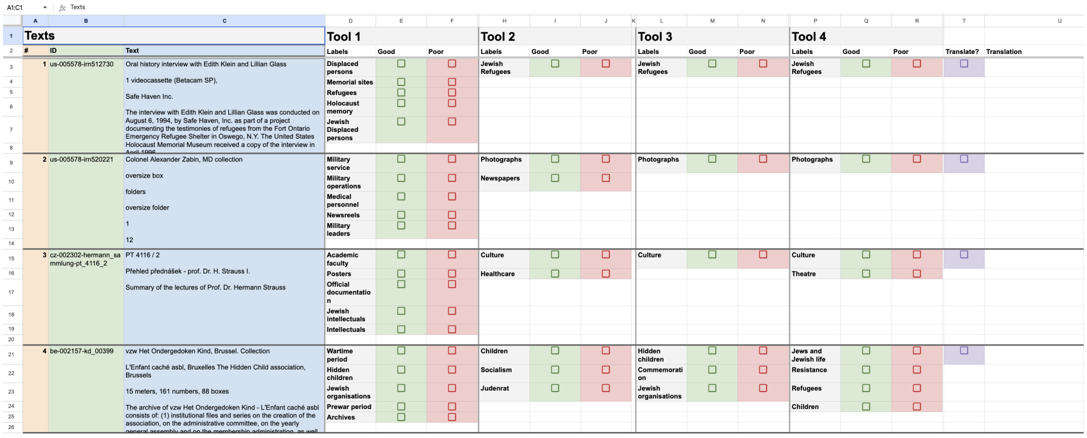

Figure 2: Spreadsheet for performing a subjective qualitative evaluation of three MTC tools along with the anonymised gold standard labels.

Scoring the subjective evaluation was done by summing each of the good, neutral, and poor labels assigned by three separate judges to the labels produced by each of the four methods, with one method including the ground truth data. The inclusion of the ground truth data, the original labels assigned either by EHRI specialists or co-references from the index terms assigned by EHRI partner institutions, was intended to serve as a control.

The result of the manual evaluation is shown in Table 2 and demonstrates the subjectivity of the multi-label classification task. Unsurprisingly, in terms of percentages, the gold standard label set achieved the highest scores of labels deemed favourable and the smallest scores of unfavourable labels. The tool with the second-highest favourable scores as percentages was the fine-tuned multilingual BERT model (favourable judgements ranging from 45 to 53 per cent), with the NN Ensemble classifier performing similarly (favourable judgements ranging from 43 to 48 per cent).

A key outcome, however, is that the predictions made by the zero-shot mDeBERTa classifier were judged qualitatively to be substantially better than suggested by our quantitative evaluation. In terms of absolute numbers, human judges awarded it the highest number of favourable labels. In a case where the institutional priority would be to have an exhaustive index, this tool could lead to better recall, potentially increasing the number of helpful index terms assigned to our descriptions. However, it was also the model that returned the largest number of suggestions in general, out of which many were deemed unfavourable (though mostly by one reviewer). That said, most importantly, "neutral" scores ranged between 37-55 per cent across all ASI tools, which means that for many of the labels included in these label sets, we simply did not have enough text to determine whether they were good or poor suggestions.

6 Discussion

The EHRI Portal includes metadata describing both physical and digital materials that may or may not have been made publicly accessible online. In our experiment we used all metadata as described in Section 3 without distinguishing between the medium of the archival material or our ability to access it ourselves. This was a fundamental limitation of the qualitative evaluation. Practising archivists will often choose index terms on the basis of their specialised knowledge of the source material to which they

	GS Labels			NN Ensemble			BERT-FT			mDeBERTa		
J	F	N	U	F	N	U	F	N	U	F	N	U
J1	221	69	4	267	252	37	233	179	28	340	312	183
	75%	23%	1%	48%	45%	7%	53%	41%	6%	41%	37%	22%
J2	210	83	1	239	288	29	197	213	30	342	444	49
	71%	28%	0%	43%	52%	5%	45%	48%	7%	41%	53%	6%
J3	191	99	4	238	303	15	217	209	14	344	460	31
	65%	34%	1%	43%	55%	3%	49%	48%	3%	41%	55%	4%

Table 2: Results of the manual evaluation showing total favourable (F), neutral (N), and unfavourable (U) assessments of gold standard labels and ASI tools. For each judge (J), the first row shows absolute numbers and the second row shows percentages rounded to the nearest integer. Due to rounding, percentages may not sum to exactly 100%.

have direct access, with textual description serving a disjoint purpose. This might have influenced the results of both the qualitative and quantitative evaluation.

When it comes to the qualitative evaluation, without access to the source material the input text was frequently not sufficient to help us conclude on whether the subject terms predicted by the different tools under evaluation were appropriate or not. This might explain why the neutral assessment is in many cases the prevailing one for many suggested labels. Considering quantitative evaluation, the fact that the texts in our dataset often contained little clue as to what the correct classification may be could also account for the relatively low performance of all the classifiers we have tested. An offsetting factor is the prevalence of broader terms in the dataset that is the result of co-referencing partner institutions' in-house vocabularies with EHRI Terms; if labels assigned in the gold standard dataset had been generally narrower (more specific) then performance in the quantitative evaluation would likely have been lower still.

For example, in many cases the concatenated text used in the first column of the evaluation spreadsheet would only include information similar to the following input text:

> "Oral history interview with Olga Horak
> 1 sound cassette,
> Australian Institute for Holocaust and Genocide Studies
> The Australian Institute of Holocaust Studies conducted the interview with Olga Horak on March 12, 1990 for the Twelfth Hour Project. The United States Holocaust Memorial Museum received a copy of the interview in April 2006 from the State Library of New South Wales where the collection is housed."[24]

Among the subject terms predicted for the description above by one of our tools under evaluation were the following: Forced labour, Death marches, Escape, Hiding places, Camp inmates. Without having listened to the oral interview this text is concerned with, we lack adequate context to rate the terms listed above based on this text alone. However, finding the metadata associated with this Oral History on the website of the source institution[25], the United States Holocaust Memorial Museum (USHMM), one may find a more insightful description that is not part of the metadata ingested by the EHRI Portal for this description and see that some of the predicted

24 The text is derived from the following description: `https://portal.ehri-project.eu/units/us-005578-irn523956-irn43072`

25 `https://collections.ushmm.org/search/catalog/irn43072`

labels (Death Marches, Hiding Places, Camp Inmates) were actually included in the original set of terms as indexed by the USHMM cataloguers themselves; while additionally the term "Escape" could also have been considered favourably.

Indeed, there were many examples where — despite the lack of sufficient salient information in the input text — some tools were suggesting terms that would have been considered favourably by human cataloguers if they had access to more context or the source material. Inspecting the reasons why this may be happening was outside the scope of this experiment. However, there are at least three potential explanations. First, this was a typical occurrence when it came to texts deriving from USHMM metadata concerning oral histories where the descriptions that EHRI had ingested contained almost no salient information about the contents of the interview (another example is the description with ID `us-005578-irn558093`[26]). After inspecting some of these examples, we realised that the interviewees mentioned in their titles are people who have given multiple interviews throughout their lives. This means that is possible that the metadata of other interviews they have given were present in the dataset, despite our deduplication process. Consequently, the models may have correlated the topics that certain persons tend to talk about in their interviews across multiple distinct archival descriptions. Another reason could be that some of the models might have learned to associate certain series of interviews with certain topics. For example, the description associated with the item with ID `us-005578-irn508794`[27] includes text that reads "(...) The United States Holocaust Memorial Museum, Oral History Branch, in cooperation with Watchtower Bible and Tract Society, Inc. (...)". In the USHMM catalogue, items that include this text in their metadata tend to carry the label "Jehovah's Witnesses" that was correctly predicted by at least one of our tools. A third reason which is relevant specifically to the predictions made by the zero-shot tool is that the base model was trained on a CommonCrawl dataset which might have already "seen" some of the descriptions in our dataset. It is also entirely possible that a combination of these explanations might be valid at the same time.

In the subjective qualitative evaluation, all three judges gave the zero-shot LLM model high scores for the number of labels considered favourable. This is justified by the fact that its predictions were usually much more specific and informative compared to the predictions returned by the other tools, which tended to be broader and more generic. A couple of examples that attest to this assessment are the label sets predicted for text with ID `us-005578-irn515693`[28], where mDeBERTa returned "Identification papers", "Displaced persons", and "Allied Forces" among the top five predictions, which can all be deemed favourable and extend the golden standard "Children" and "Photographs" labels. Another such example is the description with ID `us-005578-irn36408`[29], which included only one label in the gold standard dataset, "Photographs", while mDeBERTa additionally returned "Prewar period", "Jewish population", and "Jewish culture" among the top 5 predictions. However, given the fact that this model was not trained using EHRI's dataset but was instead used out-of-the-box, the creativity of this model was not surprising to the authors because as previously explained, the majority of the true subject terms in our dataset consisted of mainly broader terms and genre terms which may be perceived as having low information value.

26 `https://portal.ehri-project.eu/units/us-005578-irn533934-irn558093`
27 `https://portal.ehri-project.eu/units/us-005578-irn39098-irn508794`
28 `https://portal.ehri-project.eu/units/us-005578-irn515693`
29 `https://portal.ehri-project.eu/units/us-005578-irn36408`

However, there were also a number of cases where a lack of context about the overall task — labelling Holocaust-related archival descriptions — resulted in incorrect or inappropriate labels predicted by the zero-shot model as a result of text describing the history of the collection or chain of custody. As an example, the text with ID `us-005578-irn508794` reading "[USHMM], Oral History Branch, in cooperation with Watchtower Bible and Tract Society, Inc. produced the interview with…" was labelled with the term "Collaborators", a word which has obvious semantic connection with the metadata due to the existence of the phrase "in cooperation" within the text but could be inappropriate to the content of the material (as it is for this case). A more benign example of potential mislabelling is the suggestion of terms such as "Memorial sites" or "Holocaust Memory" being suggested from text containing United States Holocaust Memorial Museum as the donor institution. A common thread to these cases is that the text used as input for the classifier is imprecise and not properly descriptive of the contents of the material (as opposed to its provenance), but as future work we will investigate whether more tailored prompting can improve the precision of the zero-shot classifier here. Another example where the zero-shot model predicted inappropriate terms was the case of the text with ID `nl-002896-nl_asdniod_imagecollection-212545`[30], which concerned a family that has been recognised as righteous amongst the nations[31] and yet, amongst the predicted terms, one will find the inappropriate term "Jewish collaborators".

Another remark that needs to be made with regard to (mainly) the zero-shot model but also to a lesser extent the fine-tuned BERT model is that our decision to conduct this experiment without taking the hierarchical nature of our vocabulary into account (but rather assuming a flat hierarchy) led these two models to sometimes predict hierarchically relevant and overlapping terms. For example, in ID `us-005578-irn49797`[32], amongst the top five labels predicted by the zero-shot model are "Photographs", "Photographers" and "Photography" as separate predictions. In `us-005578-irn1002858`[33], amongst the top five labels one will find "Political activities", "Politics", "Political organisations" and "Political movements". Future work should account for the different levels in our vocabulary's hierarchy and seek to avoid such redundant classification.

Overall, despite these limitations, we found that in many cases the tools were not only predicting the true labels found in the gold standard dataset (either partially or completely) but they were suggesting additional labels that were highly relevant and informative. Although a fully automated solution is not yet feasible for our use-case, therefore, the tools discussed here all show considerable promise as part of a semi-automated indexing system, with a subject matter expert able to verify labels. Such a system, complementing EHRI's existing co-referencing tools such as the EMT described in Section 2, could enrich our dataset with sufficient labels to, in turn, train better classification models.

7 Future Work

LLMs are a fast-developing technology and, as noted above, the models covered in this paper for fine-tuned and zero-shot MTC respectively are far from the state-of-the-art.

---

30 `https://portal.ehri-project.eu/units/nl-002896-nl_asdniod_imagecollection-212545`

31 `https://www.yadvashem.org/odot_pdf/Microsoft%20Word%20-%206069.pdf`

32 `https://portal.ehri-project.eu/units/us-005578-irn49797`

33 `https://portal.ehri-project.eu/units/us-005578-irn1002858`

As new "open source" models become available, with well-documented training data sets and inference characteristics that can be reasonably well understood, we look forward to seeing how they compare in terms of the qualitative evaluation discussed in Section 5.2. We are also keen to understand how different prompting techniques can better align the labels produced by LLMs for given texts with cataloguers' expectations, by incorporating more domain-specific information into the LLM's context window. For archival descriptions which are structured hierarchically, more "archival context" is typically available from parent and/or sibling descriptions, as well as information about the holding institution, and this can be incorporated to give a classifier more data with which to work.

We are also keen to expand this experiment to incorporate more of the expertise of professional archivists in developing our dataset and understanding the characteristics of (semi-)automated subject indexing tools, not limited solely to LLMs. Integrating a semi-automated labelling pipeline into EHRI's cataloguing environment will allow us to gain more experience in the practical application of multilingual ASI with the EHRI Terms vocabulary, expanding our overall dataset and improving its overall balance with regard to less represented languages and narrower, more specific index terms.

8 Conclusion

This paper has presented an experiment where two different approaches to the use of LLMs for text classification, zero-shot and fine-tuning, were compared with alternative non-LLM-based MTC tools, using Annif as a framework, metadata from the EHRI Portal as a dataset, and the EHRI Terms vocabulary as a label set.

As described in Section 3, our dataset, drawn from metadata in the EHRI Portal, is not ideally suited for training text classifiers, with a degree of class imbalance favouring broader, more generic terms, and a lack of examples featuring narrower, more specific terms. Because of this, we have reduced the size of the label set used in this experiment from 913 to 554 labels to ensure sufficient representation. We have also had to carefully stratify the dataset in order to produce training and test sets, along with a small 167-example evaluation set, that each contains sufficient numbers of labelled examples and is balanced in terms of languages used. A small evaluation set was necessary to make both manual (human) and computationally intensive processing more feasible.

Using Annif as a framework, we created backends for our zero-shot and fine-tuned models so that they could be run in the same manner as Annif's own classification backends, and evaluated using the same metrics. Due to class imbalance, different F1 metrics produce slightly different results across the various classifiers, although the difference is not large. With fine-tuning, our BERT-based model performed in a manner very comparable to the best-performing Annif backends, exceeding the performance on our data of the best non-ensemble classifier (Omikuji/Parabel) and closely matching the NN Ensemble. Given that this is a fairly lightweight 109M parameter LLM and trainable on relatively modest resources, this is a promising result. According to the quantitative evaluation, the zero-shot mDeBERTa classifier was one of the worst-performing ASI backends.

Along with this quantitative evaluation, we also undertook a blind qualitative evaluation on our small evaluation set, with output from the best-performing out-of-the-box Annif backend (NN Ensemble), our two LLM-based classifiers, and the original labelled data as a control. The results of the qualitative evaluation suggest that the potential of the zero-shot model is not fully captured by the quantitative evaluation. Quite consistently it output labels that were considered good by all three human judges, along with a relatively large quantity of labels considered poor by one judge. As an assistive tool, therefore, it could be considered more useful than the trained models, which — consistent with the label distribution in the dataset they were trained on — tend towards broader, less specific terms. Potentially offsetting this, however, are the computational requirements of the zero-shot model, which was several orders of magnitude slower than alternative methods on the same hardware.

As highlighted in Section 6, based on our experiments the ASI tools described in this paper can realistically be considered only as part of a semi-ASI workflow. The fine-tuned multilingual BERT model can return good label sets, but the NN Ensemble model shows very similar performance while being considerably less resource-intensive to run. Given its tendency to result in more specific terms, zero-shot classification using LLMs could be potentially useful as part of data augmentation efforts to improve the representation of more specific labels, resulting in a richer and more balanced dataset that could, in turn, contribute to training more accurate domain-specific models.

Funding

This work has been carried out in the context of the EHRI-3 project funded by the European Commission under the call H2020-INFRAIA-2018-2020, with grant agreement ID 871111 and DOI 10.3030/871111.

References

Mustak Ahmed, Mondrita Mukhopadhyay, and Parthasarathi Mukhopadhyay. Automated Knowledge Organization AI ML based Subject Indexing System for Libraries. *DESIDOC Journal of Library & Information Technology*, 43:45–54, March 2023. doi: 10.14429/djlit.43.01.18619.

Henk Alkemade, Steven Claeyssens, Giovanni Colavizza, Nuno Freire, Jörg Lehmann, Clemens Neudeker, Giulia Osti, Daniel van Strien, et al. Datasheets for digital cultural heritage datasets. *Journal of open humanities data*, 9(17):1–11, 2023. doi: 10.5334/johd.124. URL `https://doi.org/10.5334/johd.124`.

Marit Asula, Jane Makke, Linda Freienthal, Hele-Andra Kuulmets, and Raul Sirel. Kratt: Developing an Automatic Subject Indexing Tool for the National Library of Estonia. *Cataloging & Classification Quarterly*, 59(8):775–793, November 2021. ISSN 0163-9374. doi: 10.1080/01639374.2021.1998283. URL `https://doi.org/10.1080/01639374.2021.1998283`.

Tobias Blanke, Veerle Vanden Daelen, Michal Frankl, Conny Kristel, Kepa Rodriguez, and Reto Speck. The Past and the Future of Holocaust Research: From Disparate Sources to an Integrated European Holocaust Research Infrastructure. In Andrea Rapp, Norbert Lossau, and Heike Neurot, editors, *Evolution der Informationsinfrastruktur*, pages 157–177. Verlag Werner Hülsbusch, Glückstadt, May 2014. ISBN 978-3-86488-043-8.

Tobias Blanke, Michael Bryant, Michal Frankl, Conny Kristel, Reto Speck, Veerle Vanden Daelen, and René Van Horik. The European Holocaust Research Infrastructure Portal. *Journal on Computing and Cultural Heritage*, 10(1):1:1–1:18, January 2017.

ISSN 1556-4673. doi: 10.1145/3004457. URL https://dl.acm.org/doi/10.1145/3004457.

Trent J. Bradberry. iterative-stratification, 2021. URL https://github.com/trent-b/iterative-stratification. original-date: 2018-02-04T00:32:10Z.

Gustavo Candela, Nele Gabriëls, Sally Chambers, Milena Dobreva, Sarah Ames, Meghan Ferriter, Neil Fitzgerald, Victor Harbo, Katrine Hofmann, Olga Holownia, et al. A checklist to publish collections as data in GLAM institutions. *Global Knowledge, Memory and Communication*, 2023. doi: 10.1108/GKMC-06-2023-0195. URL https://doi.org/10.1108/GKMC-06-2023-0195.

Gustavo Candela, Vicky Dritsou, Sally Chambers, Agiatis Benardou, Alba Irollo, and Toma Tasovac. Introduction to Collections as Data. *DARIAH-Campus*, September 2024. URL https://campus.dariah.eu/en/resource/posts/introduction-to-collections-as-data. Publisher: DARIAH-Campus.

Charlene Chou and Tony Chu. An Analysis of BERT (NLP) for Assisted Subject Indexing for Project Gutenberg. *Cataloging & Classification Quarterly*, 60(8):807–835, November 2022. ISSN 0163-9374. doi: 10.1080/01639374.2022.2138666. URL https://doi.org/10.1080/01639374.2022.2138666.

Eric H. C. Chow, T. J. Kao, and Xiaoli Li. An Experiment with the Use of ChatGPT for LCSH Subject Assignment on Electronic Theses and Dissertations. *Cataloging & Classification Quarterly*, 62(5):574–588, July 2024. ISSN 0163-9374. doi: 10.1080/01639374.2024.2394516. URL https://doi.org/10.1080/01639374.2024.2394516. Publisher: Routledge _eprint: https://doi.org/10.1080/01639374.2024.2394516.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018. Place: Brussels, Belgium.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://www.aclweb.org/anthology/2020.acl-main.747.

Maria Dermentzi and Hugo Scheithauer. Repurposing Holocaust-Related Digital Scholarly Editions to Develop Multilingual Domain-Specific Named Entity Recognition Tools. In Isuri Anuradha, Martin Wynne, Francesca Frontini, and Alistair Plum, editors, *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024*, pages 18–28, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.htres-1.3.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT, October 2023. URL https://github.com/google-research/bert. original-date: 2018-10-25T22:57:34Z.

Sigal Arie Erez, Tobias Blanke, Mike Bryant, Kepa Rodriguez, Reto Speck, and Veerle Vanden Daelen. Record linking in the EHRI portal. *Records Management Journal*, 30(3):363–378, 2020. ISSN 09565698. doi: 10.1108/RMJ-08-2019-0045. URL https://www.proquest.com/docview/2466656464/abstract/620A0D6747804A5CPQ/1.

Herminio García-González and Mike Bryant. The Holocaust Archival Material Knowledge Graph. In Terry R. Payne, Valentina Presutti, Guilin Qi, María Poveda-Villalón, Giorgos Stoilos, Laura Hollink, Zoi Kaoudi, Gong Cheng, and Juanzi Li, editors, *The Semantic Web – ISWC 2023*, Lecture Notes in Computer Science, pages 362–379, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-47243-5. doi: 10.1007/978-3-031-47243-5_20.

Koraljka Golub. Automated Subject Indexing: An Overview. *Cataloging & Classification Quarterly*, 59(8):702–719, November 2021. ISSN 0163-9374. doi: 10.1080/01639374.2021.2012311. URL https://doi.org/10.1080/01639374.2021.2012311.

Koraljka Golub, Dagobert Soergel, George Buchanan, Douglas Tudhope, Marianne Lykke, and Debra Hiom. A framework for evaluating automatic indexing or classification in the context of retrieval. *Journal of the Association for Information Science and Technology*, 67(1):3–16, 2016. ISSN 2330-1643. doi: 10.1002/asi.23600. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.23600. Number: 1 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23600.

Judit Gárdos, Julia Egyed-Gergely, Anna Horváth, Balázs Pataki, Roza Vajda, and András Micsik. Identification of social scientifically relevant topics in an interview repository: a natural language processing experiment. *Journal of Documentation*, 80(2):354–377, January 2023. ISSN 0022-0418. doi: 10.1108/JD-12-2022-0269. URL https://doi.org/10.1108/JD-12-2022-0269. Publisher: Emerald Publishing Limited.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, March 2023. URL http://arxiv.org/abs/2111.09543. arXiv:2111.09543 [cs].

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2068.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI. *Political Analysis*, pages 1–17, June 2023. ISSN 1047-1987, 1476-4989. doi: 10.1017/pan.2023.20. URL https://www.cambridge.org/core/journals/political-analysis/article/

less-annotating-more-classifying-addressing-the-data-scarcity-issue-of-supervised-machine-learning-with-deep-transfer-learning-and-bertnli/05BB05555241762889825B080E097C27.

Olena Medelyan. *Human-competitive automatic topic indexing*. Thesis, The University of Waikato, 2009. URL https://researchcommons.waikato.ac.nz/handle/10289/3513. Accepted: 2009-12-18T00:41:54Z.

Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. Is a prompt and a few samples all you need? Using GPT-4 for data augmentation in low-resource classification tasks, April 2023. URL https://arxiv.org/abs/2304.13861v1.

OpenAI. GPT-4 Technical Report, March 2023. URL http://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs].

Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. Parabel: Partitioned Label Trees for Extreme Classification with Application to Dynamic Search Advertising. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, pages 993–1002, Lyon, France, 2018. ACM Press. ISBN 978-1-4503-5639-8. doi: 10.1145/3178876.3185998. URL http://dl.acm.org/citation.cfm?doid=3178876.3185998.

Kepa J Rodriguez, Vladimir Alexiev, Laura Brazzo, Charles Riondet, Yael Gherman, and Reto Speck. EHRI-2 - D.11.2 Road Map Domain Vocabularies. Deliverable GA no. 654164, 2016. URL https://www.ehri-project.eu/sites/default/files/downloads/ehri_downloads/D%2011.2%20Road%20map%20domain%20vocabularies.pdf. Issue: GA no. 654164.

Niels Rogge. Fine-tuning BERT (and friends) for multi-label text classification, November 2021. URL https://github.com/NielsRogge/Transformers-Tutorials/blob/master/BERT/Fine_tuning_BERT_(and_friends)_for_multi_label_text_classification.ipynb.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. On the Stratification of Multi-label Data. In Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis, editors, *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pages 145–158, Berlin, Heidelberg, 2011. Springer. ISBN 978-3-642-23808-6. doi: 10.1007/978-3-642-23808-6_10.

Erjon Skenderi, Jukka Huhtamäki, and Kostas Stefanidis. Multi-Keyword Classification: A Case Study in Finnish Social Sciences Data Archive. *Information*, 12(12):491, December 2021. ISSN 2078-2489. doi: 10.3390/info12120491. URL https://www.mdpi.com/2078-2489/12/12/491.

Osma Suominen. Annif: DIY automated subject indexing using multiple algorithms. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, 29(1):1–25, July 2019. ISSN 2213-056X. doi: 10.18352/lq.10285. URL https://liberquarterly.eu/article/view/10732.

Osma Suominen and Ilkka Koskenniemi. Annif Analyzer Shootout: Comparing text lemmatization methods for automated subject indexing. *The Code4Lib Journal*, (54), August 2022. ISSN 1940-5758. URL https://journal.code4lib.org/articles/16719.

Osma Suominen, Juho Inkinen, and Mona Lehtinen. Annif and Finto AI: Developing and Implementing Automated Subject Indexing. *JLIS.it*, 13(1):265–282, January 2022. ISSN 2038-1026. doi: 10.4403/jlis.it-12740. URL https://www.jlis.it/index.php/jlis/article/view/437.

Osma Suominen, Juho Inkinen, Tuomo Virolainen, Moritz Fürneisen, Bruno P. Kinoshita, Sara Veldhoen, Mats Sjöberg, Philipp Zumstein, Robin Neatherway, and Mona Lehtinen. Annif, August 2023. URL https://github.com/NatLibFi/Annif.

Martin Toepfer and Christin Seifert. Fusion architectures for automatic subject indexing under concept drift. *International Journal on Digital Libraries*, 21(2):169–189, June 2020. ISSN 1432-1300. doi: 10.1007/s00799-018-0240-3. URL https://doi.org/10.1007/s00799-018-0240-3.

Yaşar Tonta. A Study of Indexing Consistency Between Library of Congress and British Library Catalogers, 1991. URL http://eprints.rclis.org/9464/. Issue: 2 Number: 2 Pages: 177-185 Volume: 35.

Jens Van Nooten and Walter Daelemans. Improving Dutch Vaccine Hesitancy Monitoring via Multi-Label Data Augmentation with GPT-3.5. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 251–270, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wassa-1.23. URL https://aclanthology.org/2023.wassa-1.23.

Veerle Vanden Daelen, Herminio García González, Dorien Styven, Jarno Maertens, Michala Lônčíková, Angel Chorapchiev, Eliot Nidam, Eli Furman, Johannes Meerwald, Florine Miez, Rachel Pistol, Mike Byrant, Martin Posch, Éva Kovács, Fabio Rovigo, László Csősz, Mantas Šikšnianas, Giorgos Antoniou, Laura Brazzo, Rebecca Dillmeier, and Michał Czajka. EHRI-3 Deliverable 9.6: Overview of Data Integration, 2024. URL https://www.ehri-project.eu/wp-content/uploads/2024/12/D9.6-Overview-of-data-integration.pdf.

Hughes Alan Vaughan and Pauline Rafferty. Inter-indexer consistency in graphic materials indexing at the National Library of Wales. *Journal of Documentation*, 67(1): 9–32, January 2011. ISSN 0022-0418. doi: 10.1108/00220411111105434. URL https://doi.org/10.1108/00220411111105434. Publisher: Emerald Group Publishing Limited.

Adina Williams, Nikita Nangia, and Samuel Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/N18-1101. Place: New Orleans, Louisiana.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. HuggingFace's Transformers: State-of-the-art Natural Language Processing, July 2020. URL http://arxiv.org/abs/1910.03771. arXiv:1910.03771 [cs].

Martin Wynne. Using Holocaust Testimonies as Research Data, May 2023. URL https://www.clarin.ac.uk/article/using-holocaust-testimonies-research-data.

Shiwei Zhang, Mingfang Wu, and Xiuzhen Zhang. Utilising a Large Language Model to Annotate Subject Metadata: A Case Study in an Australian National Research Data Catalogue, October 2023. URL http://arxiv.org/abs/2310.11318. arXiv:2310.11318 [cs].