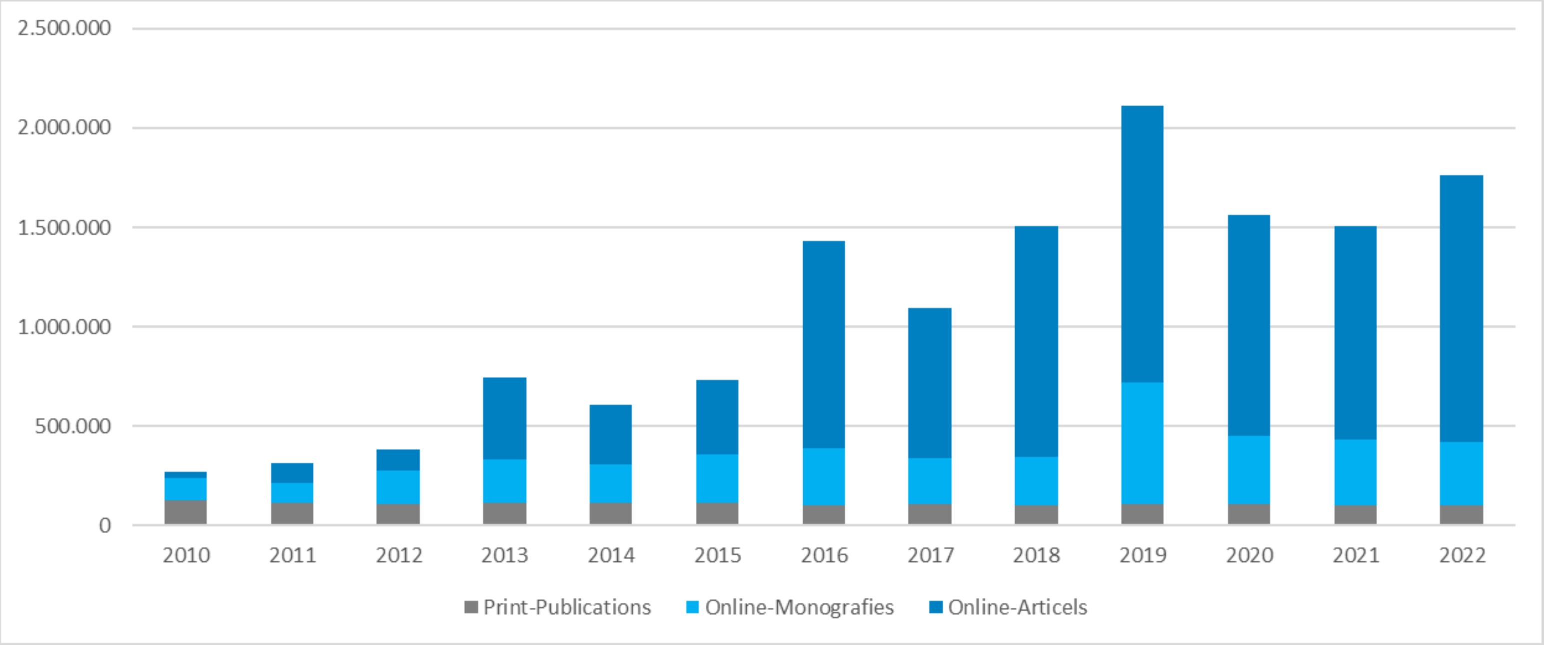
DEUTSCHE NATIONAL BIBLIOTHEK

Maximilian Kähler, German National Library (DNB)

Machine-based Subject Indexing

DEUTSCHE NATIONAL BIBLIOTHEK

Annual numbers of print- and online-publications collected by the DNB since 2010



DEUTSCHE NATIONAL BIBLIOTHEK

Machine-based subject indexing@DNB

- In production since 2014
- Complete refactorization of software architecture in 04/2022[1,2]
- Core component is now the open source library annif[3]
- Annual throughput of 170.000 publications per year
- Our target vocabulary is the **Integrated Authority File**[4] (GND) containing ~1.3M potential concepts

[1] https://blog.dnb.de/erschliessungsmaschine-gestartet/
[2] https://blog.dnb.de/in-der-dnb-lesen-jede-nacht-die-maschinen/
[3] Suominen, Osma; Inkinen, Juho; Lehtinen, Mona: Annif and Finto AI: Developing and Implementing Automated Subject Indexing. JLIS.It, 13(1), 265–282. https://doi.org/10.4403/jlis.it-12740
[4] https://gnd.network/#

Our Project: AI for automated indexing

https://www.dnb.de/ki-projekt

- Funded by German national AI-Strategy through BKM[1]
- Duration: 3 1/2 Jahre (October 2021 – March 2025)

Project Goals:

- Accelerate knowledge transfer of NLP, Data Science and Machine Learning-Methods into the Organization
- Provide new methodological directions to machine-based subject indexing

[1]BKM: Federal Government Commissioner for Culture and the Media https://www.kulturstaatsministerin.de/

DEUTSCHE NATIONAL BIBLIOTHEK

Subject Indexing as XMLC-Problem

- Subject Indexing with a large target vocabulary constitutes an E**x**treme **M**ulti-**L**abel **C**lassification Problem
  - Text documents are assigned with labels from an **a-priori known vocabulary**
  - Multi-Label problem: The number of labels per document may vary and is not a-priori bounded
- Why **extreme**?
  - Large Label-Set: 10^5 – 10^6 labels
  - Long-Tail Characteristic: The majority of labels has few/zero training material

DEUTSCHE NATIONAL BIBLIOTHEK

Interpreting subject indexing as XMLC-Problem opens up a new world of methods!

List of methods registered in the benchmark-initiative: [The Extreme Classification Repository (manikvarma.org)](http://manikvarma.org), weighted by Google-Scholar-Citations per Year, 02/2023

DEUTSCHE NATIONAL BIBLIOTHEK

Core work of our project:

Systematically benchmark XMLC-methods for German scientific texts

- Identify a suitable subset of methods to test
- Adjust methods to work with our data and hardware
- Evaluate and compare

DEUTSCHE NATIONAL BIBLIOTHEK

Our Evaluation setup:

- Methods are benchmarked in **two evaluation tasks**
  - Book Titles[1]
  - Full Text[2]
- Machine-based predictions are compared with **intellectually assigned gold-standard** on an a Test-Set and we look at:
  - Overall performance
  - Performance in various evaluation dimensions
- Promising methods will undergo **qualitative rating** by professional subject indexers to rate idexates of previously unseen material

[1]~950K training titles with intellectually assigned keywords
[2]~167K digital training documents with intellectually assigned keywords

Results

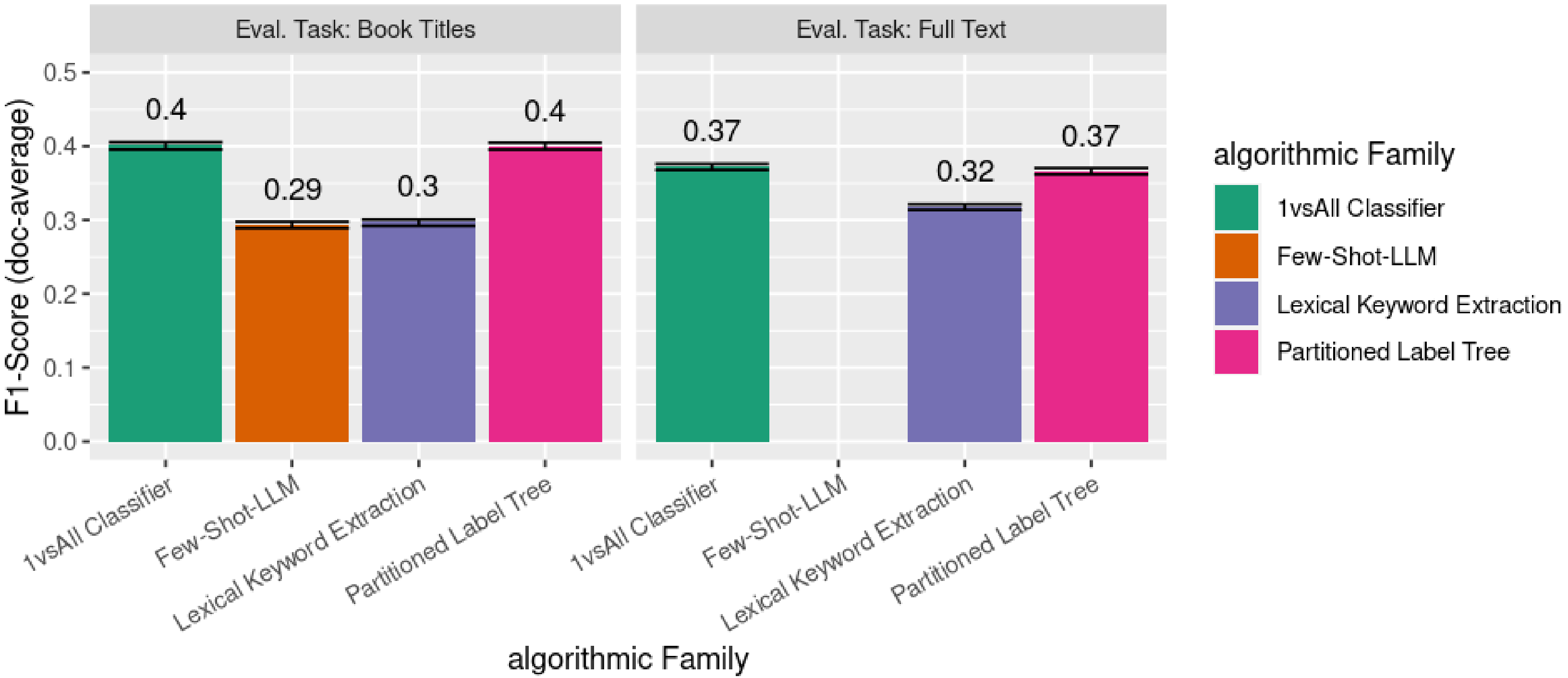
Methods analysed until now…

- 1vsAll-Classifier: Dismec++[1]
- Partitioned Label Tree: Omikuji Rust-Library[2,3,4]
- Lexical Indexing: MLLM[5,6]
- LLM with few-shot instructions: Luminous by Aleph Alpha[7]

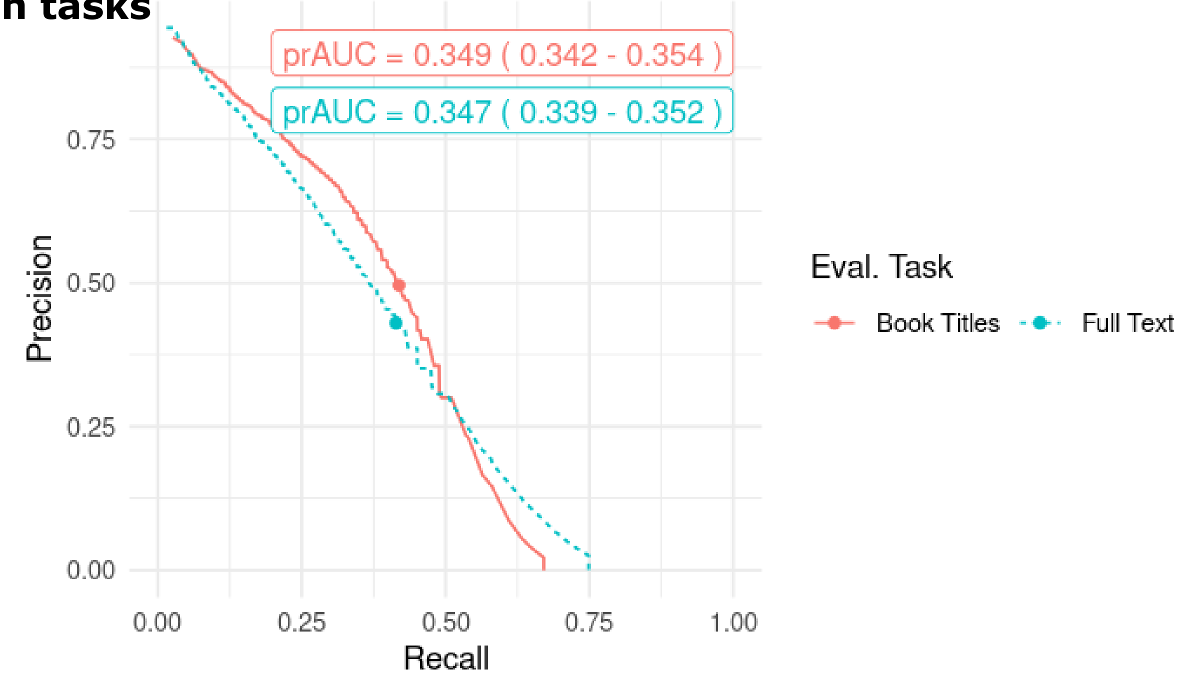
[1]Schultheis, E., & Babbar, R. (2021). *Speeding-up One-vs-All Training for Extreme Classification via Smart Initialization*. https://doi.org/10.48550/arxiv.2109.13122
[2]Khandagale, S., Xiao, H., & Babbar, R. (2020). Bonsai: diverse and shallow trees for extreme multi-label classification. https://doi.org/10.1007/s10994-020-05888-2
[3]Prabhu, Y., Kag, A., Harsola, S., Agrawal, R., & Varma, M. (2018). Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising. https://doi.org/10.1145/3178876.3185998
[4]https://github.com/tomtung/omikuji
[5]Medelyan, O., Frank, E., & Witten, I. H. (2009). Human-competitive tagging using automatic keyphrase extraction. https://doi.org/10.5555/3454287.3454810
[6]https://annif.org/
[7]https://docs.aleph-alpha.com/docs/introduction/luminous/

DEUTSCHE NATIONAL BIBLIOTHEK

F1-Score on Test-Set (Preliminary)

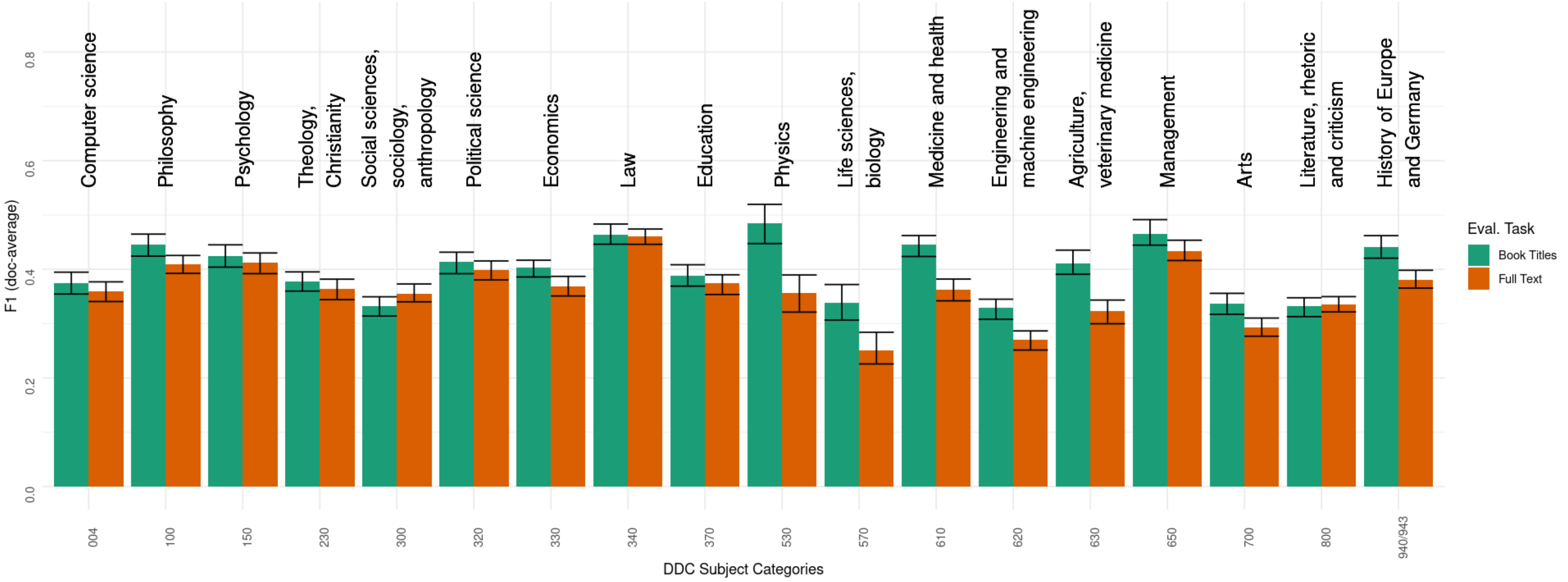


Example 1: Precision-Recall-Curve on Test-Set for Partitioned Label Trees in both evaluation tasks



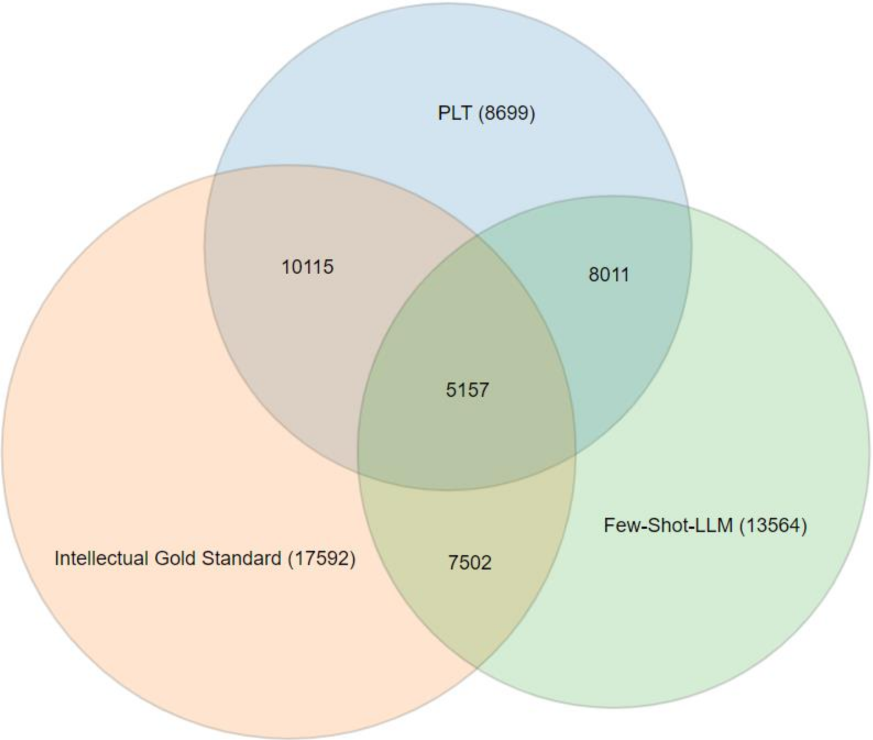
DEUTSCHE NATIONAL BIBLIOTHEK

Example 2: Results for Partitioned Label Trees in selected subject categories



DEUTSCHE NATIONAL BIBLIOTHEK

Example 3: Comparing methods by their mutual overlap



Agreement in number of document-label-pairs for predictions on book titles with two methods:

	Total number of doc-label pairs[1]
Partitionel-Label-Tree (PLT)	21.668
Few-Shot-LLM	23.920
Intellectually assigned gold standard	30.052

[1]Based on Test-Set with 8.415 documents

DEUTSCHE NATIONAL BIBLIOTHEK

Project Challenges

- End-to-end pipeline of experiments has enormous complexity
  - data selection -> pre-processing -> training -> evaluation
- Setting up a fair benchmarking process for evaluation
  - Optimizing individual methods vs. producing valid comparison
- Adaption of NLP methods from English text to German text
- Hardware
- Legal challenges

DEUTSCHE NATIONAL BIBLIOTHEK

Future Directions:

- Bring in more algorithmic families into benchmark:
  - Fine-Tuned Transformer-Architectures (e.g. XR-Transformer[1])
  - Embedding-Approaches (e.g. SLEEC[2)]
- Ensemble methods
- Qualitative evaluation by professional subject indexers

[1]Zhang, J., Chang, W., Yu, H., & Dhillon, I. S. (2021).
Fast Multi-Resolution Transformer Fine-tuning for Extreme Multi-label Text Classification. https://arxiv.org/abs/2110.00685v2
[2]Bhatia, K., Jain, H., Kar, P., & Varma, M. (2015).
Sparse local embeddings for extreme multi-label classification. https://doi.org/https://dl.acm.org/doi/10.5555/2969239.2969321

DEUTSCHE NATIONAL BIBLIOTHEK

Thank you!

Project-Team: Lisa Kluge, Katja Konermann, Nico Wagner, Moritz Kulessa, Markus Schumacher

Please get in touch for further questions and discussion:

Maximilian Kähler

m.kaehler@dnb.de

Our Project@DNB:

https://www.dnb.de/ki-projekt
https://blog.dnb.de/texte-erschliessen-mit-ki/