Automation of content indexing with machine learning methods and Annif

proven models in productive operation and initial experiments with transformer models

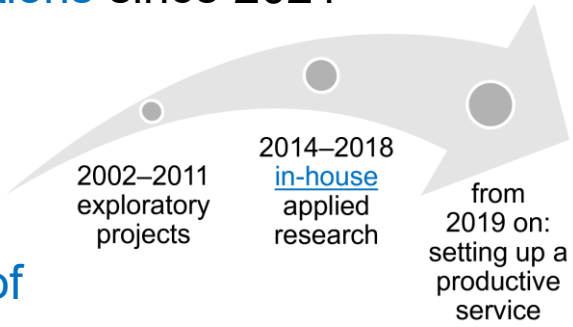
Lakshmi Rajendram Bashyam, Dr. Argie Kasprzik
ZBW – Leibniz Information Centre for Economics
DINI KIM workshop 2026, online, 21.04.2026

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

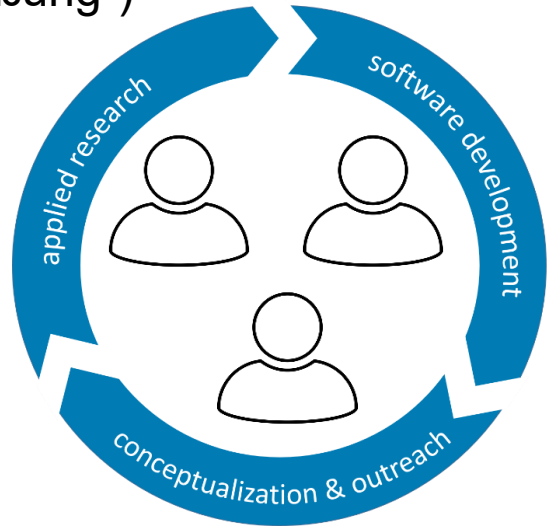
The ZBW is a member of the Leibniz Association.

AutoSE: automated subject indexing as a productive service

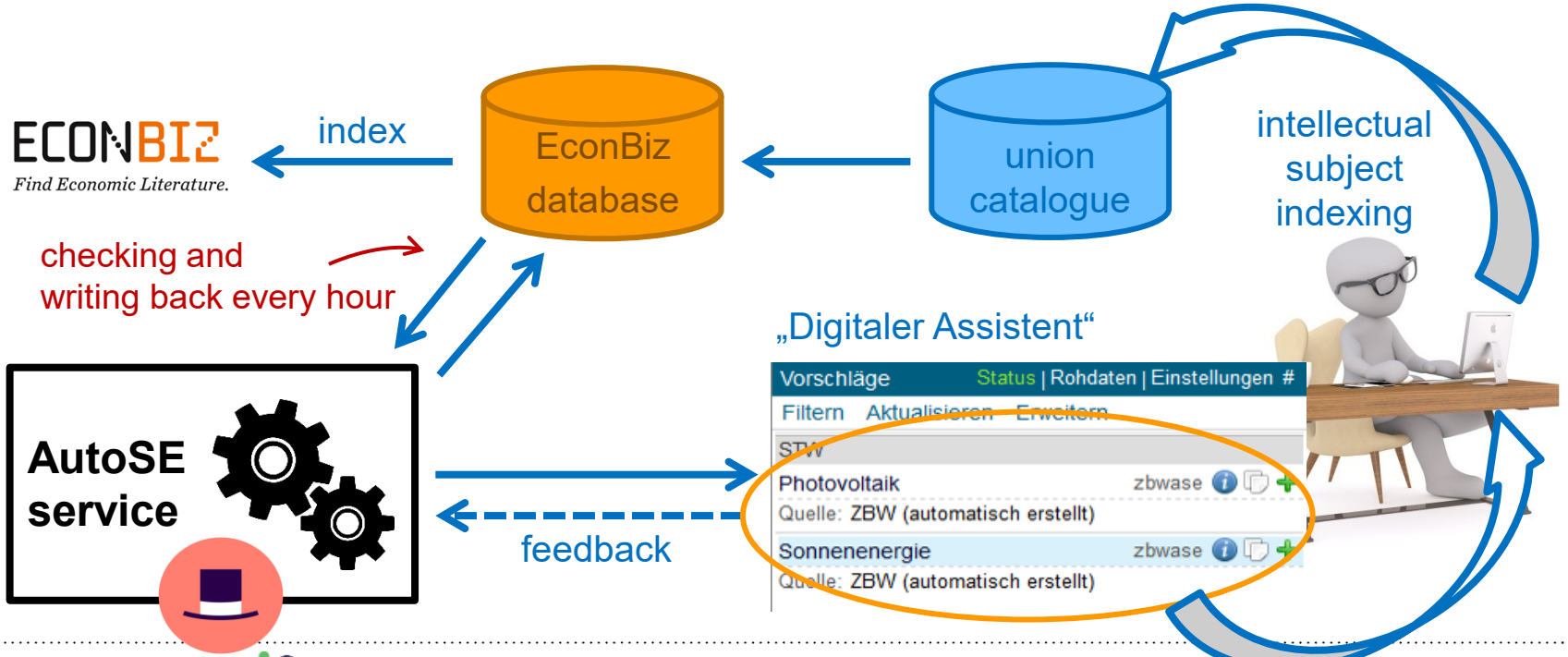
- previously: project AutoIndex (until 2018) – applied research & prototypes
- from 2019: AutoSE („Automatisierung der Sacherschließung“)
- service in productive operations since 2021
- Open Source!



- continuously:
  evaluation and integration of
  new developments from AI research
  (including our own applied research)



Fully automated and machine-assisted subject indexing



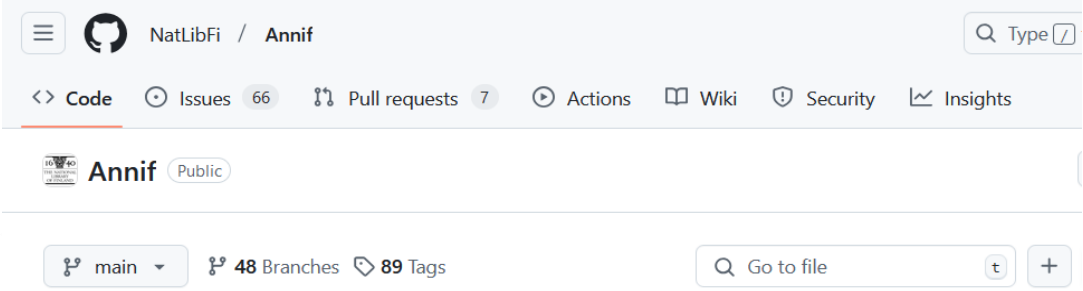
ZBW annif

trained ML models

Open source platform for automated subject indexing: Annif

- developed by the National Library of Finland (NLF):
  a (comparatively) low-threshold toolkit for automated subject indexing,
  offers various models (backends)
- long-standing exchange of ideas between ZBW and NLF,
  ZBW contributes to open source development of Annif
  via GitHub and has provided own backend (*stwfsa*)
- AutoSE uses Annif as a core component

Annif: open source as an opportunity – and a challenge

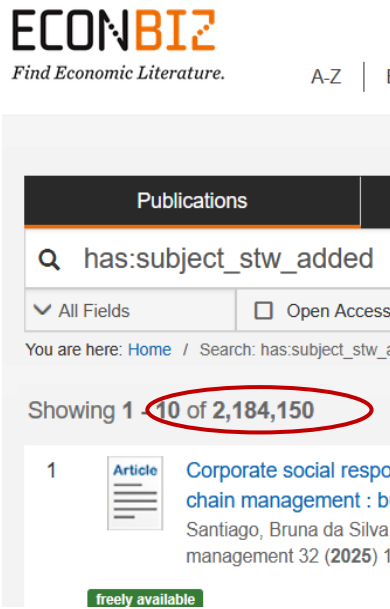


- challenge: to check periodically
  if requirements of productive systems
  such as AutoSE are still compatible with the "easy to use" ambitions of Annif
- Annif is completely open source, including machine learning models, software libraries and other external components – challenge: some projects are not maintained anymore
  - current example: new model X-Transformer → Pecos library
  - but also fastText, omikuji, i.e., "legacy models"!
- vs. long-term productive operations?
- requires community effort!

omikuji
parabel *bonsai*
stwfsa *fastText*

Methods currently used in production

- we combine machine learning algorithms incl. a custom model developed at ZBW (stwfsa *) in a so-called *ensemble*
- complemented by a subsequent application of filters and rules
- separate search for optimal parameters
- inhouse development of an automated quality control (“*qualle*” **)
- data: currently for English publications (more languages planned)
- data: currently titles and author keywords (abstracts: evaluating)
- by March 2026: ~2,2 million ZBW metadata records enriched with AutoSE



F1 score: ~0,6

* https://github.com/zbw/stwfsapy

** https://github.com/zbw/qualle

Backends used for AutoSE, available via Annif

https://github.com/NatLibFi/Annif/wiki/Backends

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Name	Type	Requires extra dependencies	Description	# train documents	Train documents length	Supports hyperopt
TF-IDF	Associative	No	A baseline algorithm, *only for setup testing*.	?	short/long	No
fastText	Associative	Yes	Allows to use word and character level n-grams (i.e. words that appear together and subwords).	10,000+	short/long	No
Omikuji	Associative	Yes	A tree-based algorithm for extreme multilabel classification.	10,000+	short/long	No
SVC	Associative	No	Linear Support Vector Classification for multiclass (*not multilabel*) classification.	?	?	No
MLLM	Lexical	No	Maui-like lexical matching.	100-10,000	long	Yes
Ensemble	Ensemble	No	Combines results from multiple backends with averaging.	NA	NA	Yes
NN-ensemble	Ensemble	Yes	Combines results from multiple backends using a neural network.	1,000-10,000	long	No
PAV	Ensemble	No	A trainable dynamic ensemble that intelligently combines results from multiple projects.	?	?	No
YAKE	Lexical	Yes	An unsupervised keyword extraction method applied to find vocabulary terms. Requires no training.	NA	long	No
STWFSA	Lexical	Yes	Statistical translation weighted finite state	?	short/long	No

Analyzing interaction with our controlled vocabulary (STW)

Tail Gap

~40%

Labels with <100 training examples account for a disproportionate share of all misses

Facet Omission

93.4%

Model catches one facet (e.g. Economics) but entirely misses another (e.g. Geography)

Specific → General

4.1%

Model predicts a broader parent category, failing to pinpoint the precise sub-topic

Silent Failures

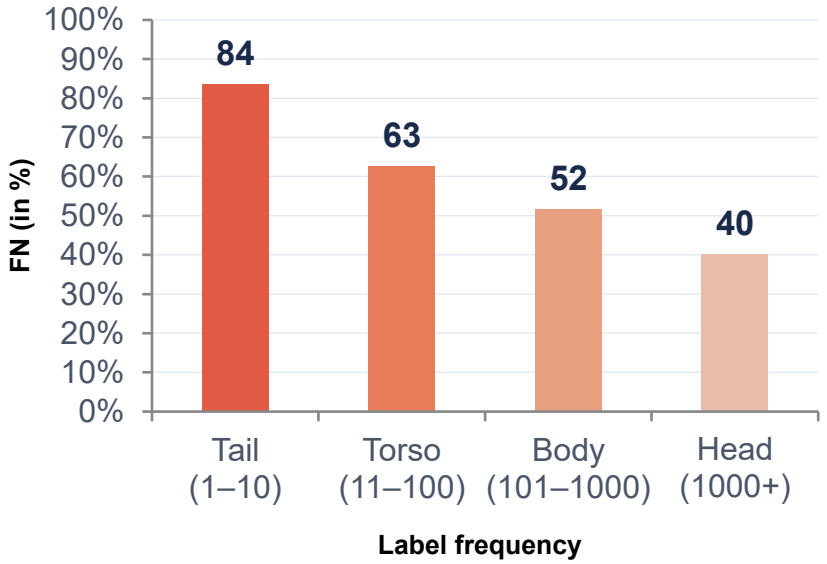
3.1%

Model predicts nothing at all for a valid, labeled document

Analyzing interaction with our controlled vocabulary (STW)

Tail Labels

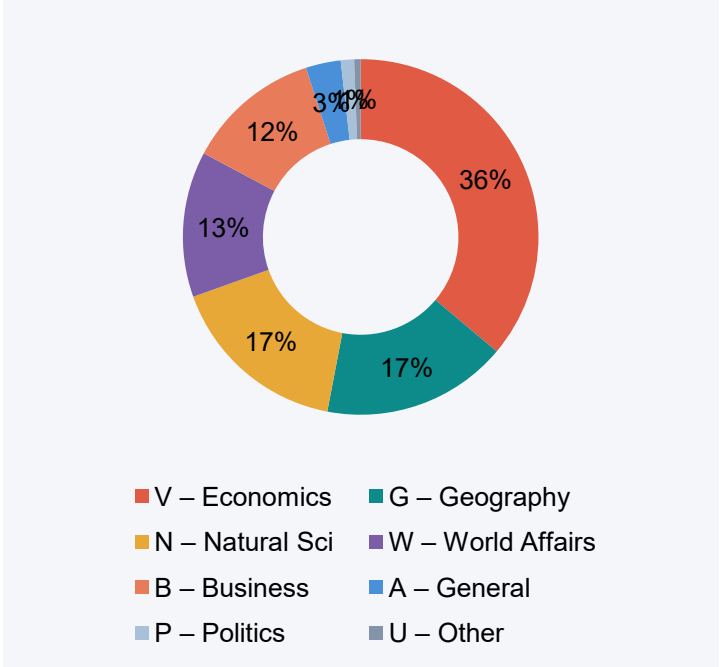
- Model misses 8 out of 10 labels for rare subjects.
- Even for the most frequent labels, the model misses 40%.
- Top missed: United States (2,629), Theory (1,811), World (1,396)



Analyzing interaction with our controlled vocabulary (STW)

Facet omission

Facet	Omissions	%
G.04.02 North America	2,865	3.4%
A.00 General Descriptors	2,521	3.0%
G.00 World	1,454	1.7%
G.01.05 Western Europe	1,352	1.6%
V.07.03 Intl Economic Relations	1,338	1.6%
V.03.07 Macroeconomic Policy	1,298	1.5%



Analyzing interaction with our controlled vocabulary (STW)

Hierarchial near misses

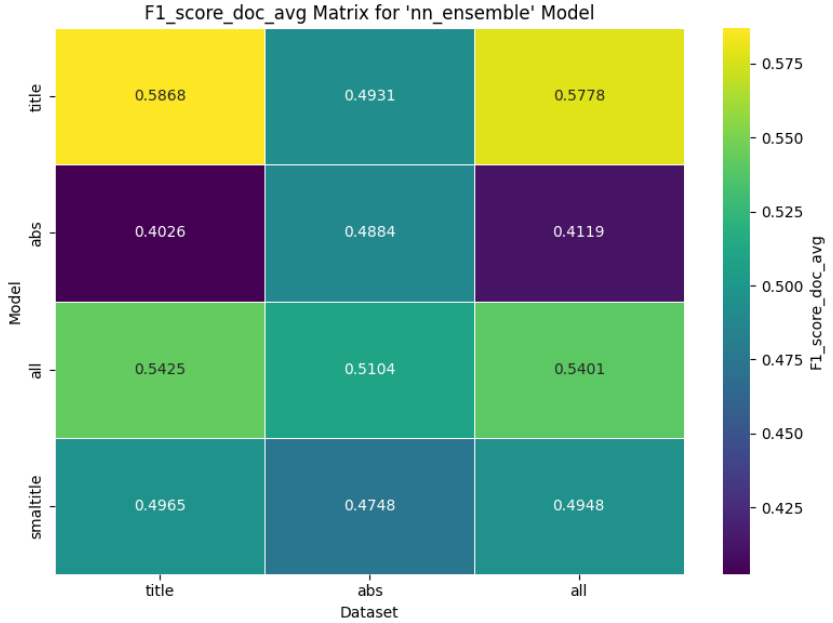
- 12.5% of false positives are hierarchical near-misses (sibling or parent labels predicted instead of the correct one)
- In 4.1% of false negatives, a broader parent category is predicted, failing to pinpoint the precise sub-topic
- High semantic similarity between confused label

Analyzing interaction with input type (→ abstracts)

Task: Automated subject indexing of library publications

Scale: 890K training samples, ~25K test records

Question: Does adding abstracts to training data improve label assignment?



Analyzing interaction with input type (→ abstracts)

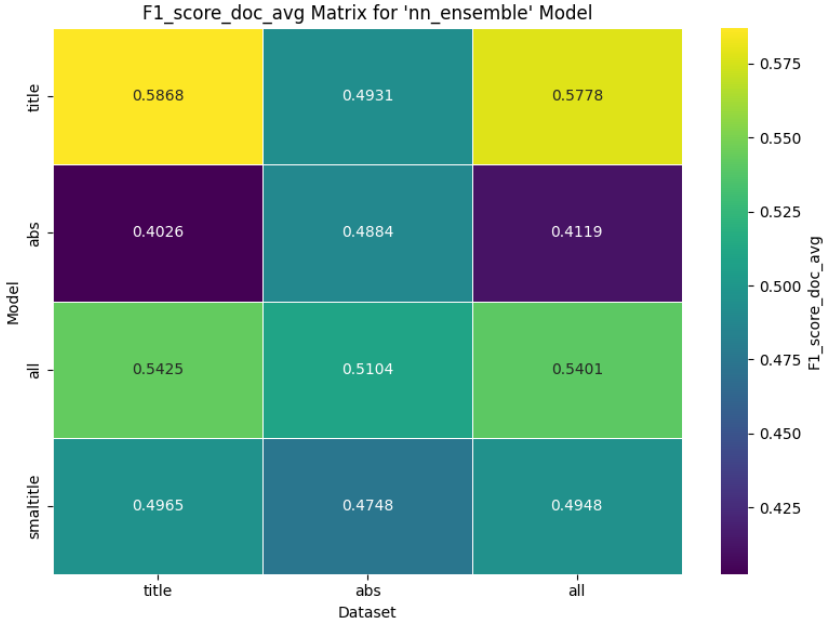
Top Performer: title -> title

F1 = 0.587 | Title-trained model on title test set scores highest across all conditions

Worst performer: abs -> title

F1 = 0.403 | Abstract-trained model collapses completely on title-based queries

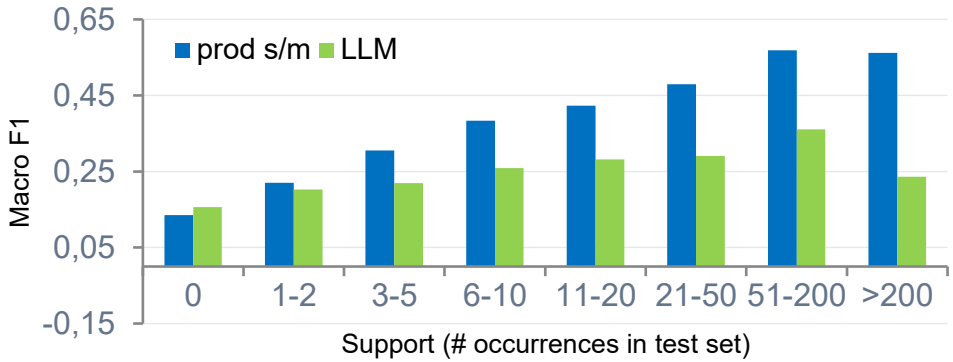
Quantity over Density



Preliminary experiments with LLMs in subject indexing / retrieval

Overall our productive model wins

LLM trades precision for recall — but only on tail labels



Segment	N	Prod s/m F1	LLM F1	ΔF1	Prod s/m Prec	LLM Prec	Prod s/m Rec	LLM Rec
Tail ≤5	2475	0.2114	0.1875	-0.0240	0.299	0.217	0.1799	0.2202
Tail ≤2	1650	0.1644	0.1716	+0.0072	0.19	0.1737	0.1527	0.2145
Support = 0	11	0	0	0.0000	0	0	0	0
Head >50	110	0.5678	0.3505	-0.2173	0.8779	0.5963	0.4387	0.3085

in cooperation with Arben Hajra, ZBW

Preliminary experiments with LLMs in subject indexing / retrieval

Coverage

- Our productive model makes zero predictions on **46.7%** of labels
- LLM only abstains on **29.7%**.

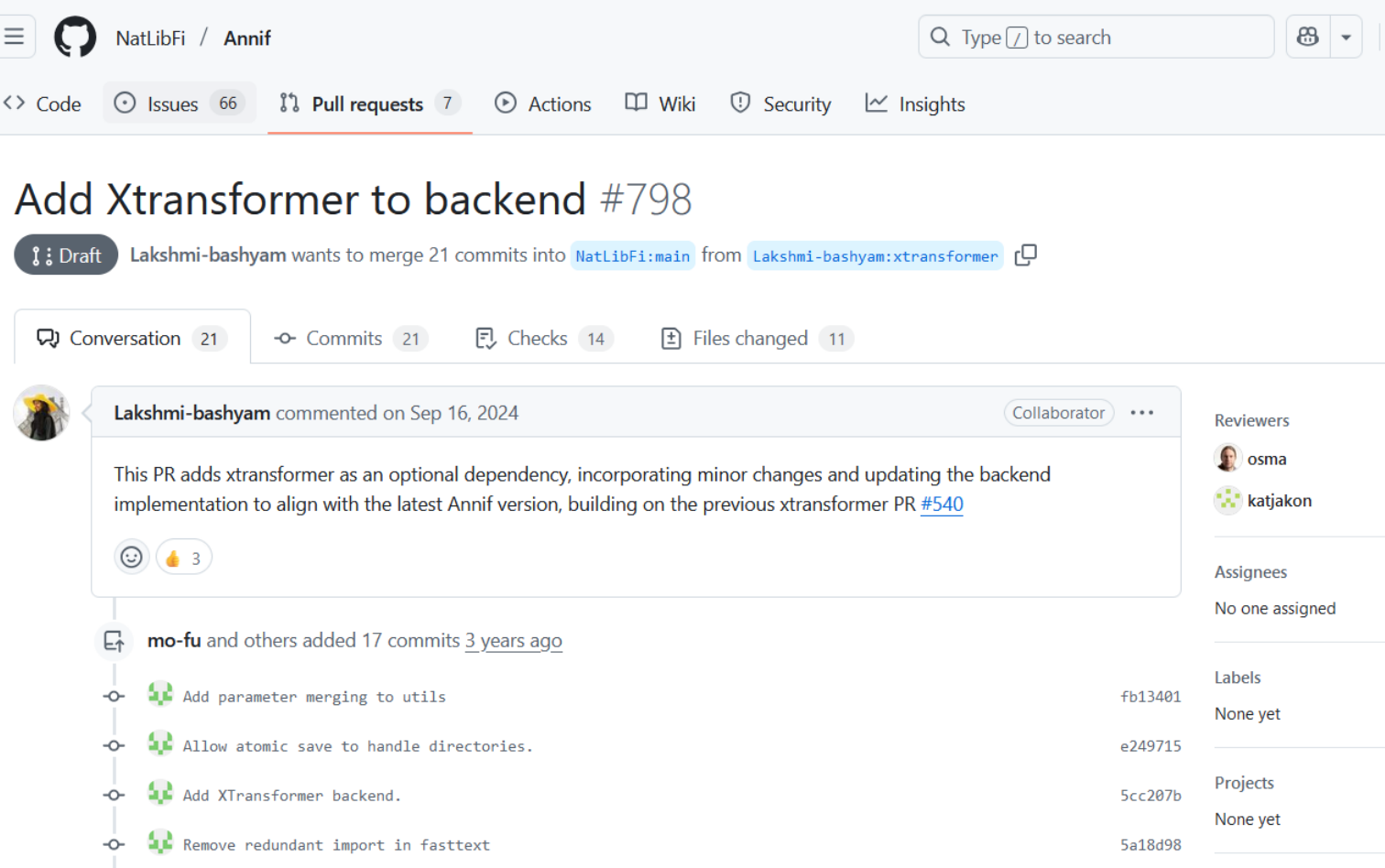
- **LLM is a broader but noisier predictor**

Hallucination

- 698 labels appear *only* in LLM results and not in the autose prod model. Each of these labels one has support=0 and F1=0

in cooperation with Arben Hajra, ZBW

pull request created by AutoSE team to add a transformer model to Annif



Methods we are looking at short-term / mid-term

- Evaluate X-transformer models
- LLMs for Tail labels
  - Leverage LLMs to classify rare labels with insufficient training data (tail: 1–10 examples)
  - Test Evidence-Based Models designed for tail-label scenarios; measure recall gap vs. current ensemble
- Per-Category Specialist Models
  - Train separate models per thesaurus category (V, G, N..) to reduce facet omission errors
- Hierarchy aware prediction
- Human in the loop

Thank you!

AutoSE (incl. some more slide decks and publications)**:**
https://www.zbw.eu/en/about-us/knowledge-organisation/automation-of-subject-indexing-using-methods-from-artificial-intelligence

contact: a.kasprzik@zbw.eu ; l.rajendram-bashyam@zbw.eu