ifis

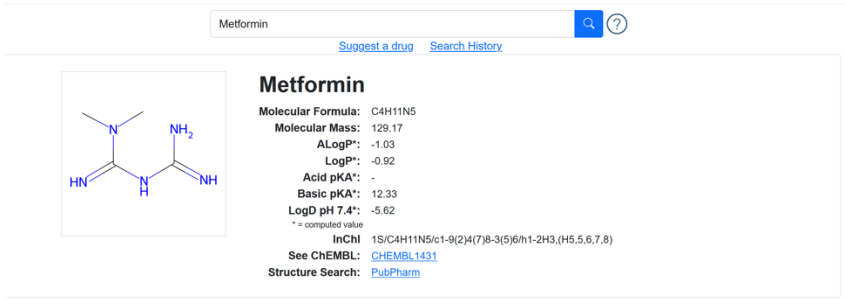
Institut für Informationssysteme
Technische Universität Braunschweig

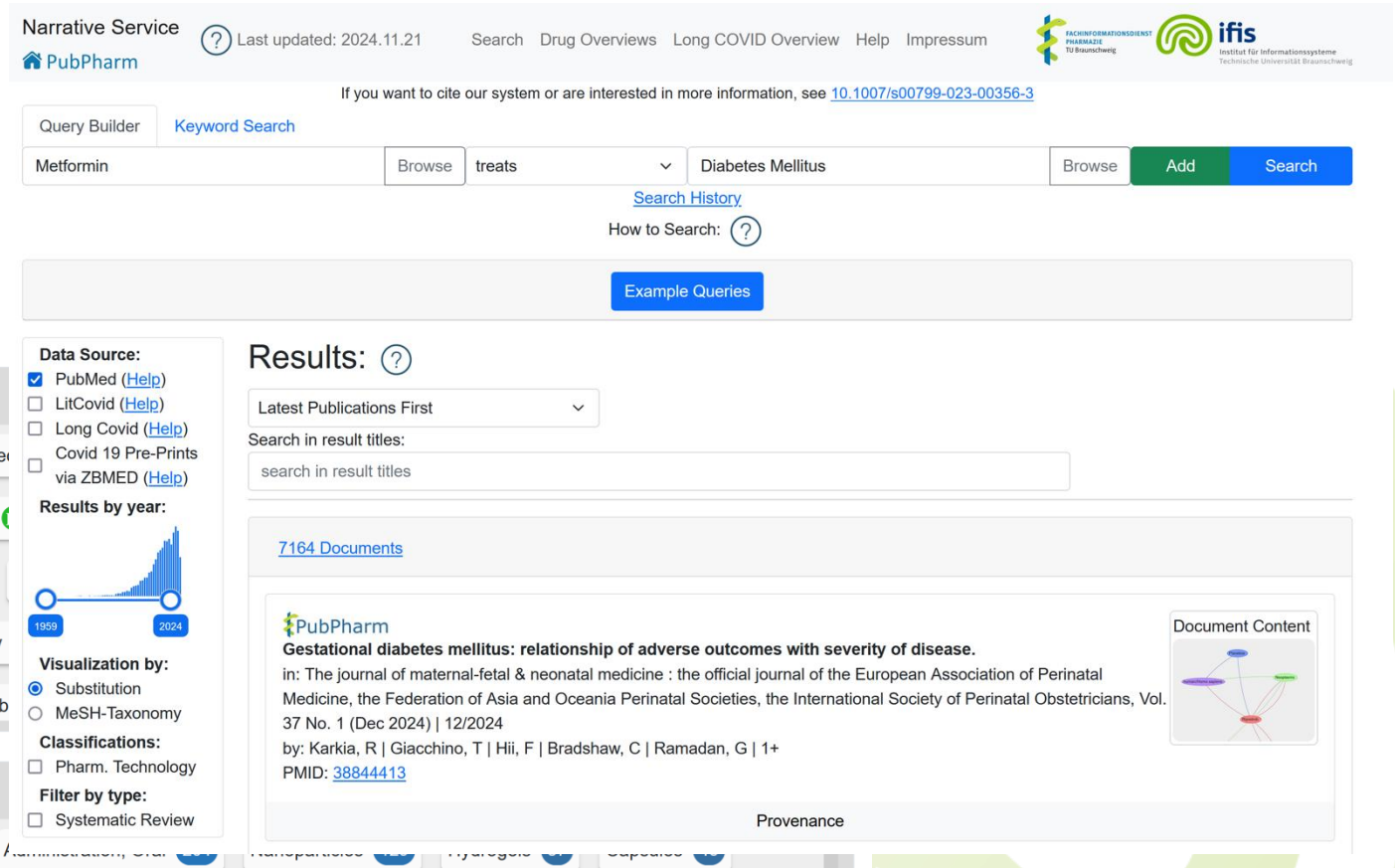
A Library Perspective on Supervised Text Processing in Digital Libraries: An Investigation in the Biomedical Domain

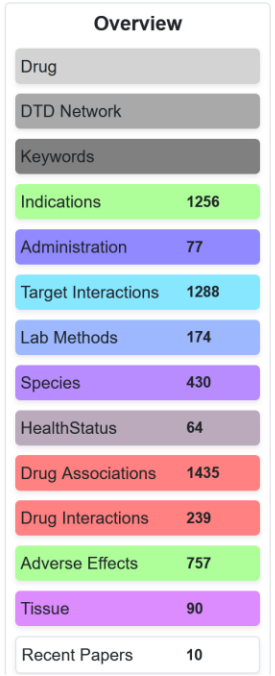
Hermann Kroll, Pascal Sackhoff, Bill Matthias Thang, Maha Ksouri and Wolf-Tilo Balke

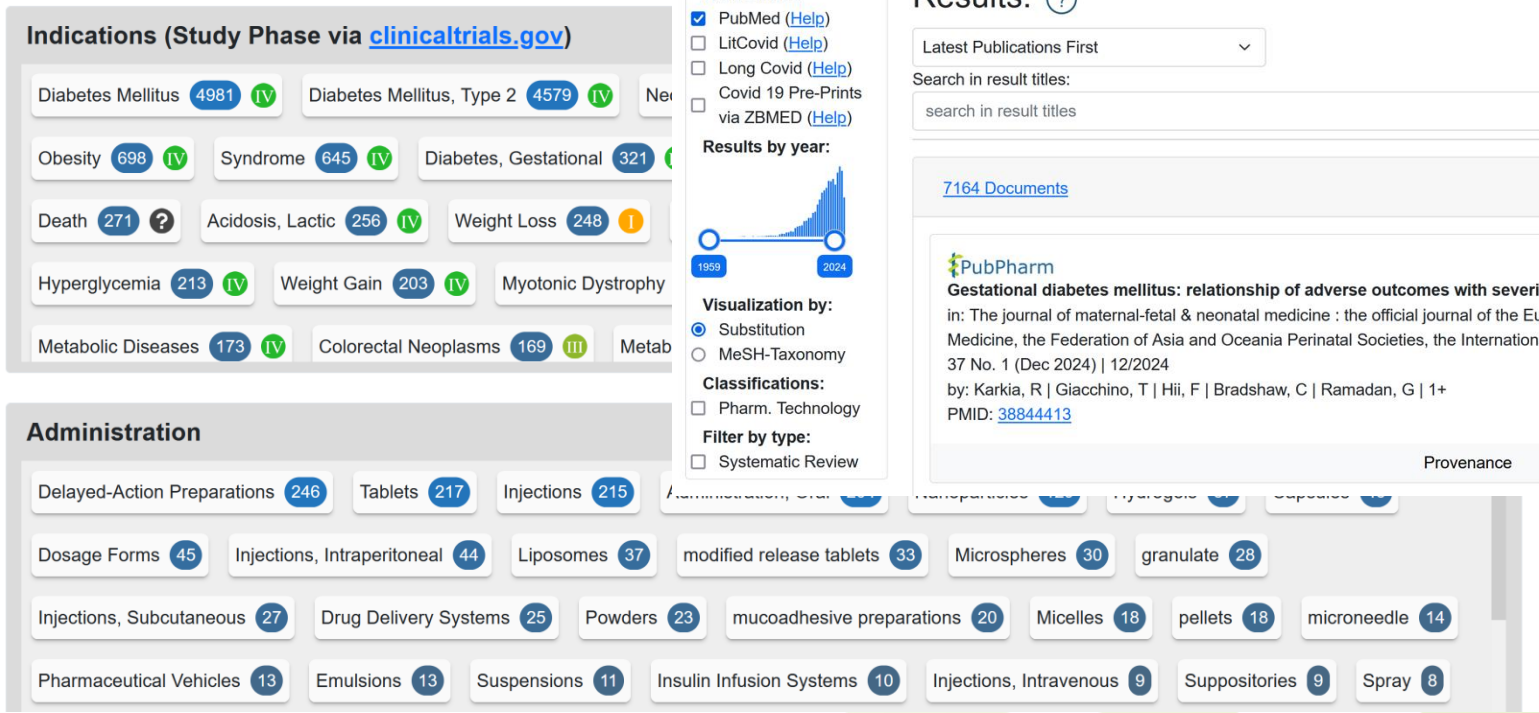
Institute for Information Systems
TU Braunschweig, Germany

Narrative Service



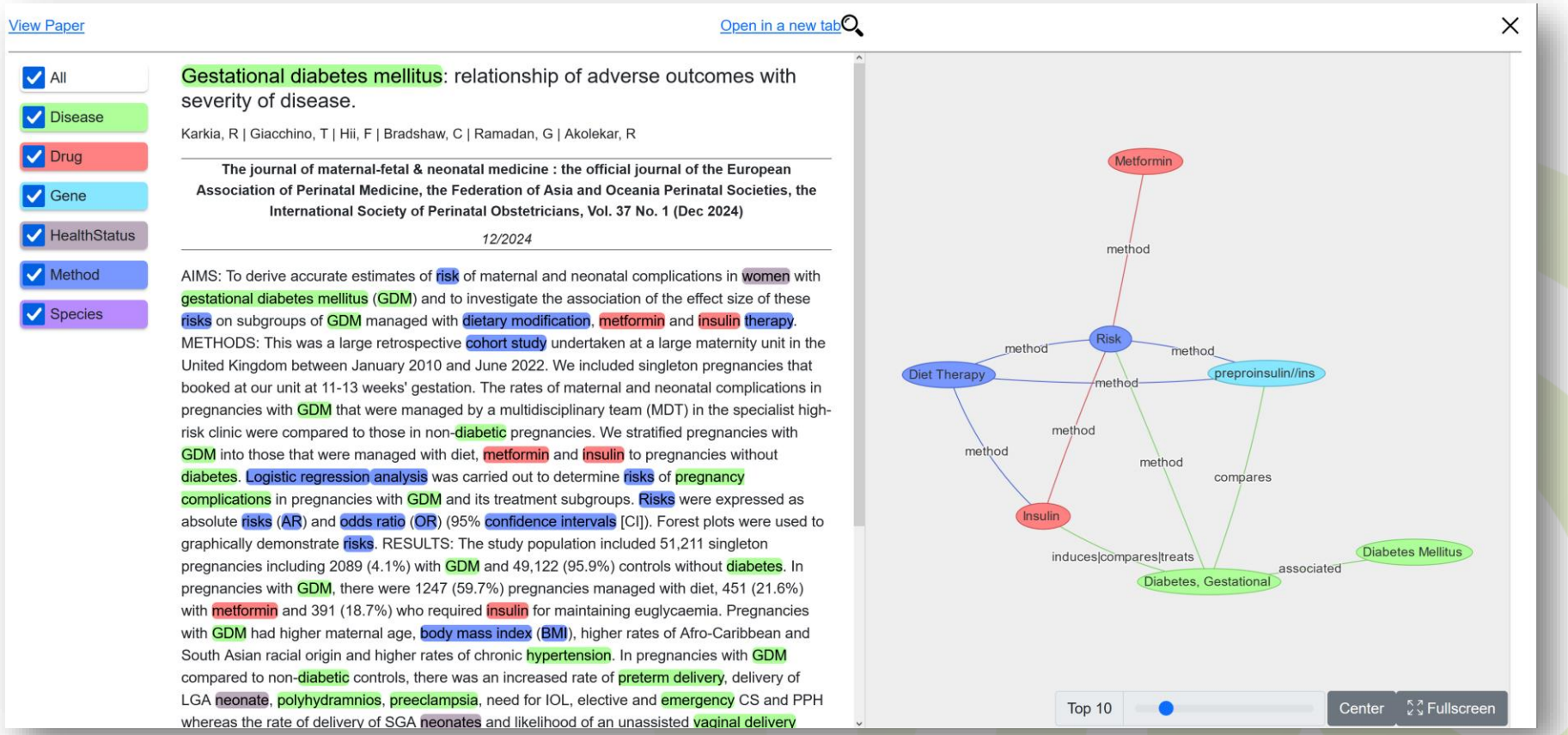






www.narrative.pubpharm.de

Document Graphs



www.narrative.pubpharm.de

Supervised Text Processing

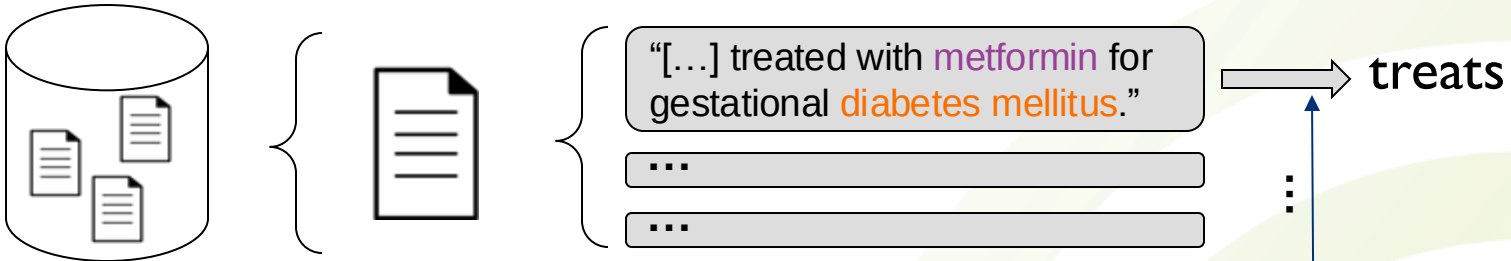
https://github.com/HermannKroll/SupervisedTextProcessing

Research Questions

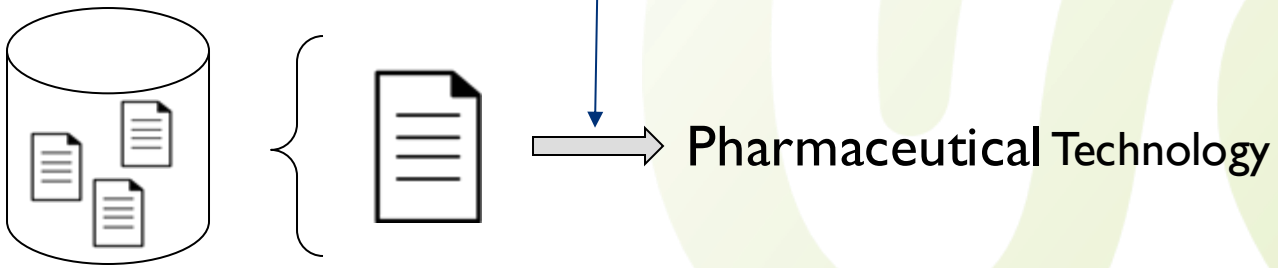
- **RQ1**: Which model should we use in a digital library project?
  - Tradeoff between accuracy and training/runtime costs
- **RQ2**: How to design a full digital library pipeline?
  - One model for all purposes? Multiple models?
- **RQ3**: How can we label training data?
  - By experts? By distant supervision? By large language models?

Tasks

- Relation extraction:



- Text classification:



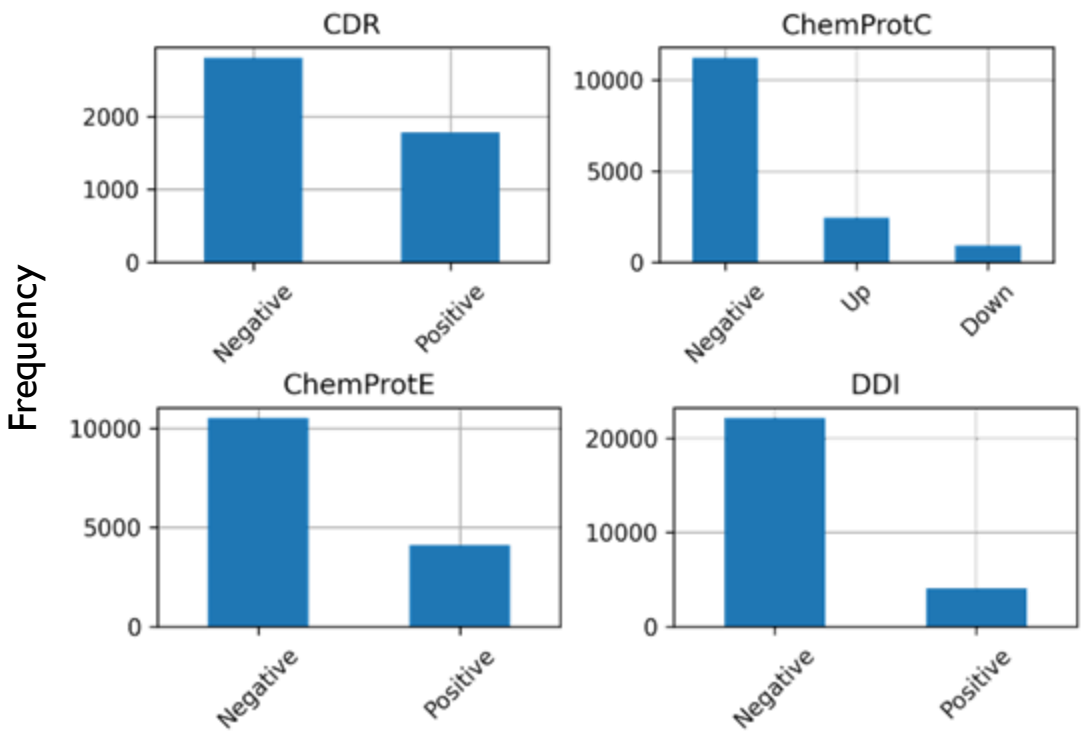
Classification Models

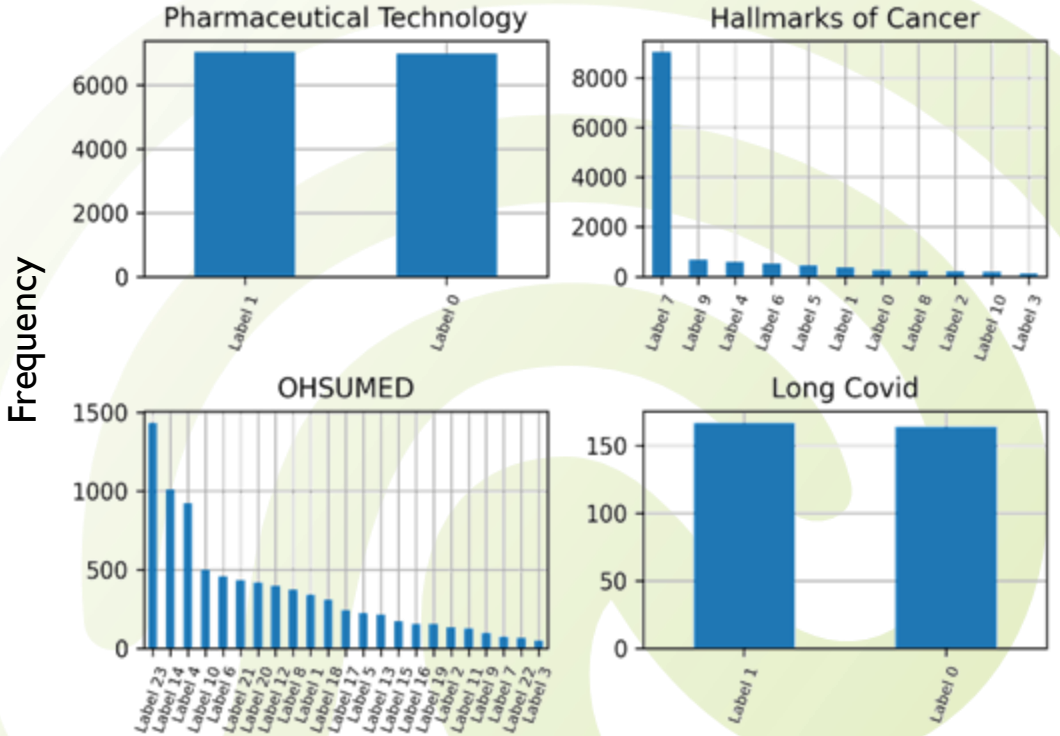
- Traditional classification models:
  - Support Vector Classifier
  - Extreme Gradient Boosting (XGBoost)
  - Random Forest
  - tfidf and sBERT for embedding texts

- Language Models:
  - Generic: BERT, RoBERTa, XLNet
  - Domain-specific: BioBERT, BioLinkBERT, PubMedBERT

Datasets

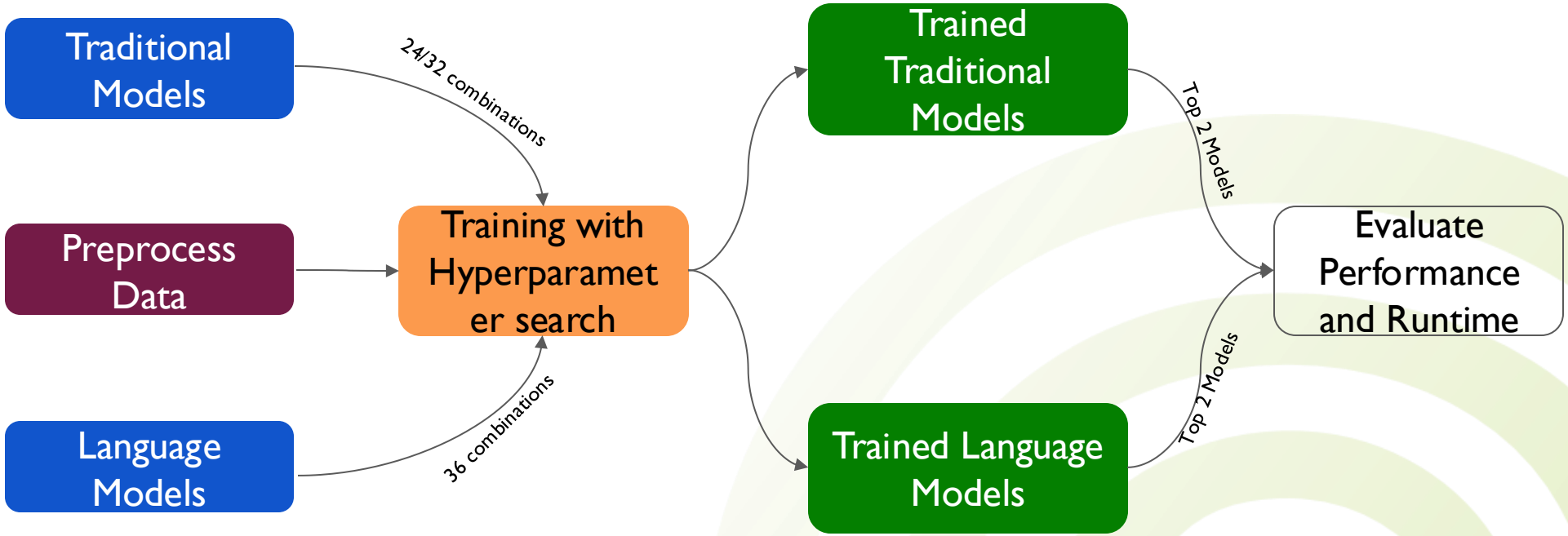
- 8 biomedical data sets (4 for each task)





RQ1: Which model?

RQ1: Evaluation Strategy



Model	Hyperparameter Grid
SVC	**C**: {0.1, 1, 10, 100}, **kernel**: {poly, rbf, sigmoid}, **degree**: {1, 2, 3, 4, 5, 6}
XGBoost	**n_estimators**: {50, 100}, **max_depth**: {3, 5}, **learning_rate**: {0.01, 0.1}, **subsample**: {0.8, 1.0}, **colsample_bytree**: {0.8, 1.0}
Random Forest	**n_estimators**: {50, 100}, **max_depth**: {None, 10, 20}, **min_samples_split'**: {2, 5}, **min_samples_leaf**: {1, 2}
Language Models	**learning_rate**: {1e-3, 1e-4, 1e-5}, **epochs**: {1, 3, 5}, **weight_decay**: {0.0, 0.1, 0.2, 0.3}

RQ1: Results

Relation Extraction

Model	CDR			DDI		
	P	R	F1	P	R	F1
Traditional Classification Models						
SVC + tfidf	**0.49**	0.58	**0.53**	**0.22**	0.81	**0.35**
SVC + sBERT	**0.49**	0.58	**0.53**	**0.22**	0.81	**0.35**
XGBoost + tfidf	0.45	0.63	**0.53**	0.21	0.78	0.34
XGBoost + sBERT	0.45	0.63	**0.53**	0.21	0.78	0.34
Random Forrest + tfidf	0.39	0.61	0.47	0.18	0.92	0.3
Random Forrest + sBERT	0.43	**0.69**	**0.53**	0.18	**0.93**	0.3
Language Models						
BERT	0.57	0.7	0.63	0.56	0.93	0.7
RoBERTa	0.57	0.75	0.65	0.54	0.93	0.68
XLNet	0.54	0.55	0.55	0.59	0.88	0.71
BioLinkBERT	**0.59**	**0.79**	**0.68**	**0.67**	0.92	**0.78**
BioBERT	0.58	0.8	**0.68**	0.59	0.92	0.72
PubMedBERT	0.6	0.78	**0.68**	0.59	**0.94**	0.73

Text Classification

Model	Hallmark			Pharm. Tech.		
	P	R	F1	P	R	F1
Traditional Classification Models						
SVC + tfidf	**0.36**	**0.67**	**0.44**	**0.87**	**0.89**	**0.88**
SVC + sBERT	**0.36**	**0.67**	**0.44**	**0.87**	**0.89**	**0.88**
XGBoost + tfidf	0.27	0.57	0.31	0.86	0.84	0.85
XGBoost + sBERT	0.27	0.57	0.31	0.86	0.84	0.85
Random Forrest + tfidf	0.24	0.5	0.26	0.77	0.86	0.81
Random Forrest + sBERT	0.23	0.5	0.25	0.77	0.88	0.82
Language Models						
BERT	0.36	0.76	0.44	0.89	0.90	0.89
RoBERTa	0.37	0.75	0.45	0.88	**0.93**	0.91
XLNet	0.27	0.64	0.31	0.90	0.90	0.90
BioBERT	0.44	0.8	0.54	**0.91**	**0.93**	**0.92**
BioLinkBERT	0.41	0.8	0.51	0.90	0.91	0.91
PubMedBERT	**0.46**	**0.82**	**0.56**	0.90	0.92	0.91

RQ1: Performance (Relation Extraction)

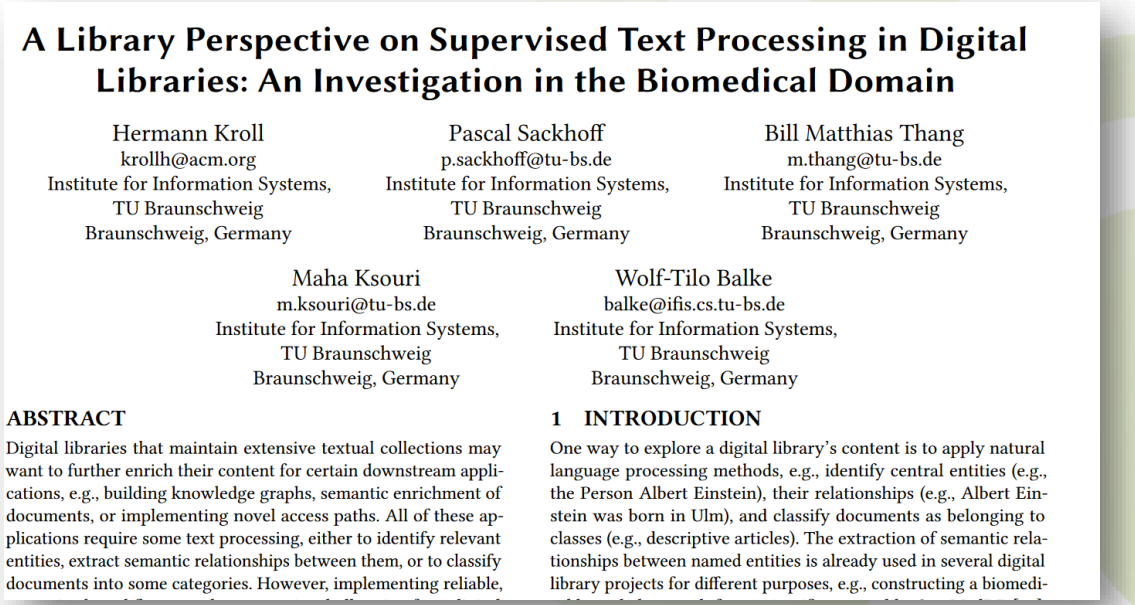
Model	Training	HS	Application	ET PubMed
Traditional Classification Models				
SVC + tfidf	57 min 28 s	42 h 41 min 7 s	9.79e-04 s	9 h 56 min 40.26 s
SVC + sBERT	1 h 0 min 5 s	53 h 17 min 10 s	9.70e-04 s	9 h 50 min 47.59 s
XGBoost + tfidf	**9 s**	6 min 51 s	**5.63e-05 s**	**34 min 17.15 s**
XGBoost + sBERT	**9 s**	**6 min 37 s**	7.00e-05 s	42 min 39.25 s
Random Forest + tfidf	2 min 51 s	7 min 38 s	7.96e-05 s	48 min 28.16 s
Random Forest + sBERT	3 min 27 s	8 min 24 s	7.68e-05 s	46 min 48.54 s
Language Models on GPU				
BERT	27 min 24 s	12 h 14 min 4 s	4.02e-03 s	1 d 16 h 52 min 12.99 s
RoBERTa	**9 min 48 s**	8 h 40 min 25 s	**3.92e-03 s**	**1 d 15 h 49 min 11.95 s**
XLNet	12 min 12 s	22 h 6 min 47 s	1.15e-02 s	4 d 20 h 34 min 6.06 s
BioBERT	10 min 6 s	10 h 44 min 4 s	3.97e-03 s	1 d 16 h 18 min 5.54 s
BioLinkBERT	9 min 53 s	**5 h 57 min 51** s	4.00e-03 s	1 d 16 h 34 min 23.02 s
PubMedBERT	28 min 29 s	9 h 30 min 4 s	3.99e-03 s	1 d 16 h 33 min 50.04 s

RQ1: Take-a-ways

	Relation Extraction	Text Classification
Measure	Effectiveness	
Quality	LMs **outperform** TMs	TMs are **comparable** to LMs
Specificity	Domain specific models **mostly outperform** generic models	
	Efficiency	
Model	TMs are **much faster** than LMs (even for CPU vs GPU)	
Hardware	LMs **require** a GPU for a **large scale application**	
Specificity	Domain specific models are **faster** than the generic models	Both types have **comparable** application times

RQ2: How to design a full digital library pipeline?

Please have a look at our paper!



A Library Perspective on Supervised Text Processing in Digital Libraries: An Investigation in the Biomedical Domain

Hermann Kroll
krollh@acm.org
Institute for Information Systems,
TU Braunschweig
Braunschweig, Germany

Pascal Sackhoff
p.sackhoff@tu-bs.de
Institute for Information Systems,
TU Braunschweig
Braunschweig, Germany

Bill Matthias Thang
m.thang@tu-bs.de
Institute for Information Systems,
TU Braunschweig
Braunschweig, Germany

Maha Ksouri
m.ksouri@tu-bs.de
Institute for Information Systems,
TU Braunschweig
Braunschweig, Germany

Wolf-Tilo Balke
balke@ifis.cs.tu-bs.de
Institute for Information Systems,
TU Braunschweig
Braunschweig, Germany

ABSTRACT

Digital libraries that maintain extensive textual collections may want to further enrich their content for certain downstream applications, e.g., building knowledge graphs, semantic enrichment of documents, or implementing novel access paths. All of these applications require some text processing, either to identify relevant entities, extract semantic relationships between them, or to classify documents into some categories. However, implementing reliable,

1 INTRODUCTION

One way to explore a digital library's content is to apply natural language processing methods, e.g., identify central entities (e.g., the Person Albert Einstein), their relationships (e.g., Albert Einstein was born in Ulm), and classify documents as belonging to classes (e.g., descriptive articles). The extraction of semantic relationships between named entities is already used in several digital library projects for different purposes, e.g., constructing a biomedi-

RQ3: How can we label training data?

RQ3: Relabeling

- Relabel existing **training data** by:
  - Experts: provided by the benchmark
  - Distantly supervised: relabel using knowledge bases
  - Prompting large language models
    - 1-Prompt: take the result of the first prompt
    - 3-Prompt; 1-Yes: Positive label if one of the three prompts is positive
    - 3-Prompt; 2-Yes: Positive label if two of the three prompts are positive
    - 3-Prompt; 3-Yes: Positive label if all of the prompts are positive
- This just works for **relation extraction**

RQ3: Relabeling Quality (RE)

Training Data Generation	$t_{sen}\downarrow$	CDR P	CDR R	CDR F1	CDR NA	DDI P	DDI R	DDI F1	DDI NA
Distantly-supervised	< 0.1s	0.44	**0.99**	0.61	-	0.14	0.27	0.19	-
OlMo 7B (1-Prompt)	0.17s	0.52	0.95	0.67	1	0.18	0.98	0.3	2
OlMo 7B (3-Prompt, 1-Yes)	0.29s	0.52	0.96	0.67	-	0.17	**0.99**	0.29	-
OlMo 7B (3-Prompt, 2-Yes)	0.42s	0.6	0.89	0.71	-	0.18	0.98	0.31	-
OlMo 7B (3-Prompt, 3-Yes)	0.49s	0.64	0.83	**0.72**	-	0.22	0.94	0.35	-
Llama 3 8B (1-Prompt)	0.21s	0.74	0.42	0.54	-	0.26	0.81	0.39	-
Llama 3 8B (3-Prompt, 1-Yes)	0.37s	0.74	0.65	0.69	-	0.24	0.91	0.38	-
Llama 3 8B (3-Prompt, 2-Yes)	0.51s	0.76	0.46	0.57	-	0.26	0.84	0.4	-
Llama 3 8B (3-Prompt, 3-Yes)	0.6s	0.73	0.26	0.39	-	0.28	0.72	0.41	-
GPT-4o (1-Prompt)	-	0.77	0.59	0.67	-	0.55	0.88	0.67	-
GPT-4o (3-Prompt, 1-Yes)	-	0.75	0.67	0.71	-	0.53	0.93	**0.68**	-
GPT-4o (3-Prompt, 2-Yes)	-	0.77	0.59	0.67	-	0.55	0.89	**0.68**	-
GPT-4o (3-Prompt, 3-Yes)	-	**0.78**	0.5	0.61	-	**0.56**	0.84	0.67	-

RQ3: Training on Relabeled Data (RE)

1. Train model on relabeled training
2. Test model on original test set

Model	Labeling	CDR			DDI		
		P	R	F1	P	R	F1
SVC + tfidf	Experts	0.49	0.58	0.53	0.22	0.81	0.35
XGBoost + tfidf	Experts	0.45	0.63	0.53	0.21	0.78	0.34
BioLinkBERT	Experts	0.59	**0.79**	**0.68**	**0.67**	0.92	**0.78**
PubMedBERT	Experts	**0.6**	0.78	**0.68**	0.59	**0.94**	0.73
SVC + tfidf	Distant.	0.39	**0.79**	**0.53**	**0.21**	0.22	0.22
XGBoost + tfidf	Distant.	0.37	0.59	0.46	0.16	0.2	0.18
BioLinkBERT	Distant.	**0.41**	0.78	**0.53**	0.17	**0.39**	**0.23**
PubMedBERT	Distant.	**0.41**	0.73	**0.53**	0.17	**0.39**	**0.23**
SVC + tfidf	LLama 3 (3Y)	0.47	0.27	0.34	0.27	0.46	0.34
XGBoost + tfidf	LLama 3 (3Y)	0.44	0.31	0.36	0.23	0.5	0.31
BioLinkBERT	LLama 3 (3Y)	0.51	0.49	0.5	0.34	0.77	0.47
PubMedBERT	LLama 3 (3Y)	**0.53**	**0.5**	**0.52**	**0.37**	**0.83**	**0.51**
SVC + tfidf	GPT-4o (2Y)	0.5	0.39	0.43	0.24	0.73	0.36
XGBoost + tfidf	GPT-4o (2Y)	0.46	0.36	0.41	0.27	0.72	0.39
BioLinkBERT	GPT-4o (2Y)	0.55	**0.67**	**0.61**	**0.48**	**0.95**	**0.64**
PubMedBERT	GPT-4o (2Y)	**0.62**	0.56	0.59	**0.48**	**0.95**	**0.64**

RQ3: Findings for Relation Extraction

	Distantly supervised	Llama 3	GPT-4o
	Binary labels		
Quality	Depends on the knowledgebase	**Comparable results** with the experts labeling → GPT-4o still better	
	Multiple labels		
Quality	**Not** applicable	**Bad** results; invalid answers occur	
	Other		
Require-ments	Good knowledgebases	GPU to run the model efficiently	OpenAI Account + API access ~130$ to relabel 4 benchmarks
Pricing	Depends on the knowledgebase source	**Free** to use after verification	**Expensive** for real scale applications

RQ3: Text Classification

fractions
100%
75%
50%
25%

Flip data randomly

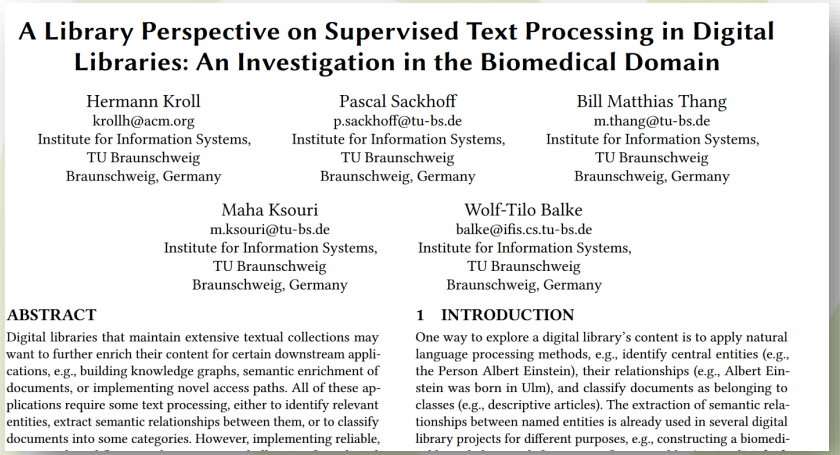
Introduce **noise**

Reduce data randomly

Train on **two best performing** models

Compare results

Look at our paper ☺



A Library Perspective on Supervised Text Processing in Digital Libraries: An Investigation in the Biomedical Domain

Hermann Kroll
krollh@acm.org
Institute for Information Systems, TU Braunschweig
Braunschweig, Germany

Pascal Sackhoff
p.sackhoff@tu-bs.de
Institute for Information Systems, TU Braunschweig
Braunschweig, Germany

Bill Matthias Thang
m.thang@tu-bs.de
Institute for Information Systems, TU Braunschweig
Braunschweig, Germany

Maha Ksouri
m.ksouri@tu-bs.de
Institute for Information Systems, TU Braunschweig
Braunschweig, Germany

Wolf-Tilo Balke
balke@ifis.cs.tu-bs.de
Institute for Information Systems, TU Braunschweig
Braunschweig, Germany

ABSTRACT

Digital libraries that maintain extensive textual collections may want to further enrich their content for certain downstream applications, e.g., building knowledge graphs, semantic enrichment of documents, or implementing novel access paths. All of these applications require some text processing, either to identify relevant entities, extract semantic relationships between them, or to classify documents into some categories. However, implementing reliable,

1 INTRODUCTION

One way to explore a digital library's content is to apply natural language processing methods, e.g., identify central entities (e.g., the Person Albert Einstein), their relationships (e.g., Albert Einstein was born in Ulm), and classify documents as belonging to classes (e.g., descriptive articles). The extraction of semantic relationships between named entities is already used in several digital library projects for different purposes, e.g., constructing a biomedi-

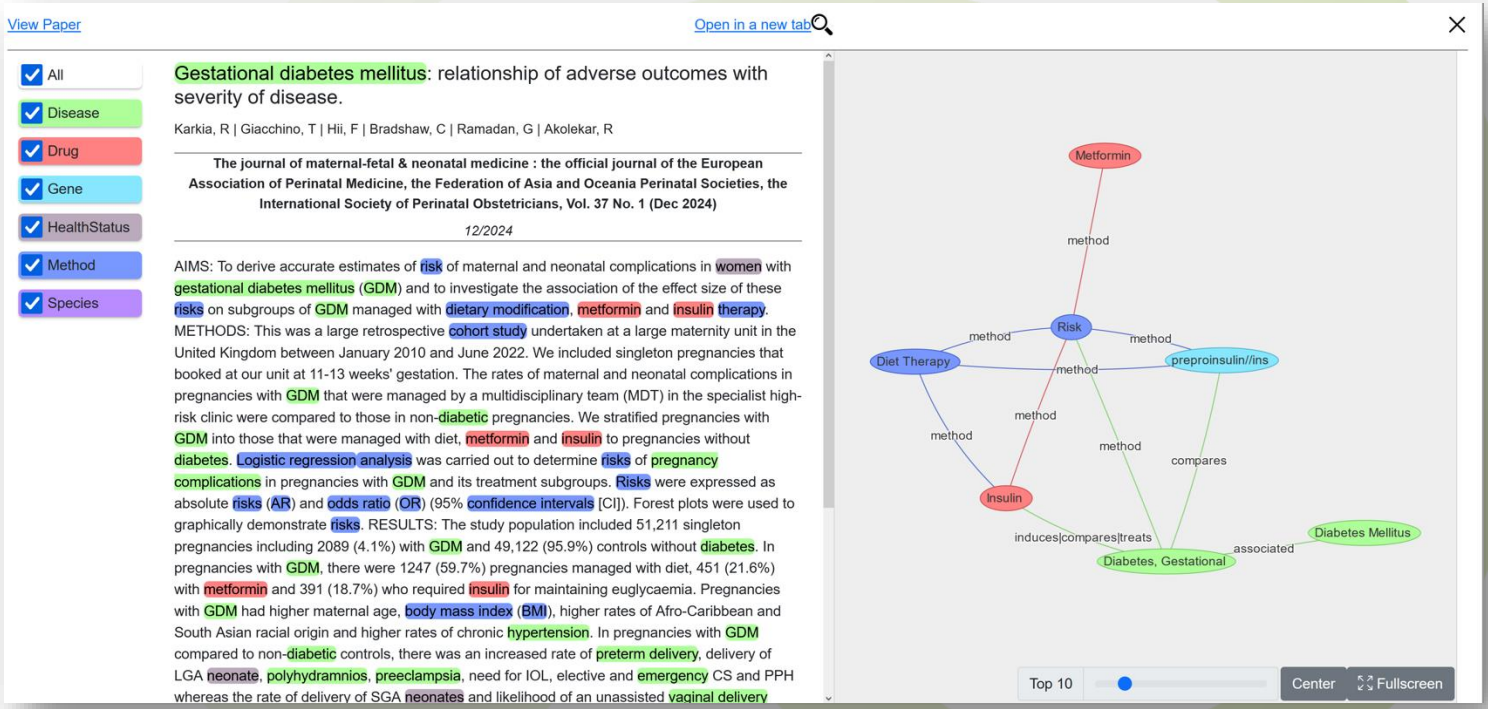
Conclusion

Conclusion

Contributions	
Model comparison	**LMs** more **robust** and **accurate** compared to **shallow models**
	GPUs are a **must-have** when working with **LMs**
	Shallow models like **SVC/XGBoost** may still **worth using**
Data Labeling	**LLMs** can **label training** data with a **moderate quality and costs** → overall **classification quality** is then **decreased**
Future Work	
MultiTask	Examination of **more reliable MultiTask-Setups**
Data Labeling	**Prompt Engineering** and **Instruction Tuning**
	Possibility of generating **useful** datasets using **LLMs**

Why did we do the research?

- Our service handles about 38M document abstracts
  - Methods need to be robust and scalable
  - We do not have training data available for every required relation



But why should you read?

- *"While some of our findings were expected, e.g., that LMs are more robust and accurate on classification tasks, our paper contributes a library perspective when applying them."*

  ~ claimed by our paper

- Briefly, it's a **library perspective** on NLP tools
  - We shed a light on the tradeoff between quality and runtimes
  - We share our code ☺

Thank You!

FACHINFORMATIONSDIENST PHARMAZIE
TU Braunschweig

Technische Universität Braunschweig

If you have any questions, contact me via:

@hkroll@fosstodon.org

krollh@acm.org

www.hkroll.de

https://github.com/HermannKroll/SupervisedTextProcessing