Multilingual BERT for Library Classification in Romance Languages Using Basisklassifikation

José Calvo Tello
Enrique Manjavacas
Susanne Al-Eryani

SUB
NIEDERSÄCHSISCHE STAATS- UND UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN
UNIVERSITÄT GÖTTINGEN

1. Machine Learning for Subject Indexing

1. Machine Learning for Subject Indexing

- Application of algorithms for subject indexing in libraries
- Annif (Suominen 2019; Suominen, Lehtinen, and Inkinen 2022)
- Application to other institutions
  - German National Library (Uhlmann 2022)
  - Leibniz Information Centre for Economics (Kasprzik and Fürneisen 2022)

1. Machine Learning for Subject Indexing

- National library perspective
- Mainly one languages (in some cases two or three)
- Not representative for university libraries
- K10plus catalog contains records in more than 400 languages
- Risk of reinforcing already privileged languages
  - English and German
- Priority: Multilingual models

Romance Studies

2. Romance Studies

- Studies about languages derived from Latin, their cultures and literatures
  a. **Resources-rich Romance languages**: such as Spanish, French, or Italian
  b. **Romance languages with fewer resources**: Romanian, Galician
  c. **Middle cases**: Portuguese, Catalan
  d. **Resources-rich non-Romance languages:** specially German and English (used as communication languages)

2. Romance Studies

- ~~National perspective~~
- Multilingual
- Challenge for current ML and NLP approaches
- Superficial similarities, shared Latin script

3. Basisklassifikation (Basic Classification)

3. Basisklassifikation (Basic Classification)

- Middle sized mono-hierarchic library classification system (CS)
- One of the most frequently assigned CS in the German-speaking area
- Labels currently only expressed in German
- Published openly
- 48 main classes
- 2087 subclasses

3. Basisklassifikation (Basic Classification)

- Examples
  - 17.97 Texts by a single author
  - 18.38 Portuguese literature outside Portugal
  - 20.31 Visual artists
  - 33.23 Quantum physics
  - 54.75 Language processing

3. Basisklassifikation (Basic Classification)

- Identify publications from Romance Studies
- Publications classified with classes which match following regex: 18.[23]?
- Only consider classes in the 17 and 18 classes

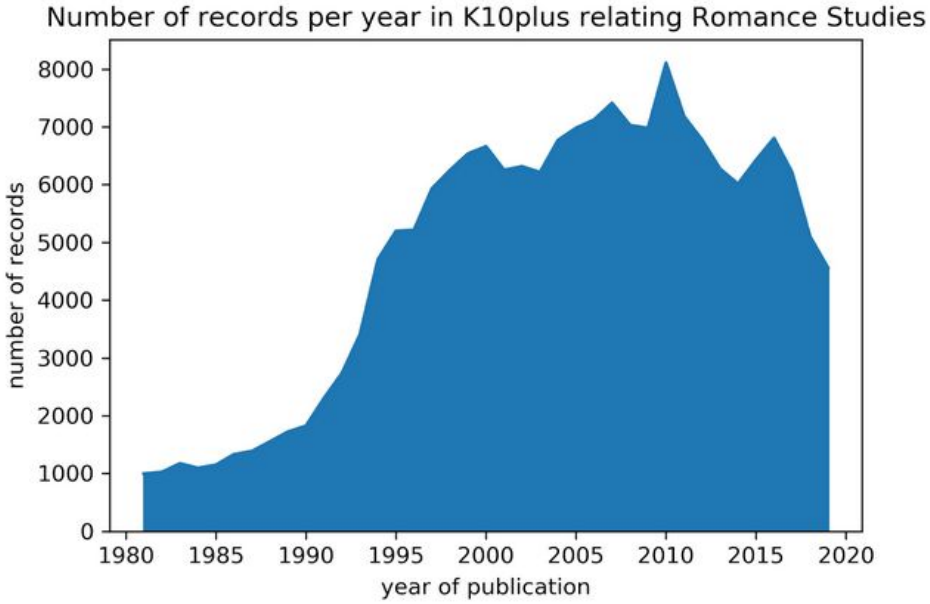
3. Basisklassifikation (Basic Classification)

- Total depth of 5 levels
- For this analysis, we consider three levels
  - **1st level**: only the main classes (17 and 18)
  - **2nd level**: direct child, with a total of 63
  - **Full**: the entire trees rooted at classes 17 and 18, with 157 possible classes

4. Dataset

4. Dataset

- K10plus: two German library networks (GBV and SWB)
- 1980-2019
- 189.134 records relating Romance Languages



4. Dataset

- Total number of labels: 157
- Extreme multi-label learning
- Multi-label
- Labels per record:
  - Mean: 1.8
  - Standard deviation: 0.7

4. Dataset

- Only data from the catalog
- ~~Full text~~
- Three different situations operationalize through different fields
  a. **Title**
  b. **Bibliographic data**
  c. **Extended data**

4. Dataset

- **Title**:
  - PICA+ field 021A $a
- **Bibliographic data**: more general case, for example, data from publishers or in a bibliography (Zotero)
  - Title (title, title supplement in PICA+ field 021A $d, and title in continuing resources from field
    036E $a)
  - Publisher (field 033A $n)
  - Place of publication (field 033A $a)
- **Extended data**: specific use case for libraries. Previous fields plus:
  - Summary (047I $a)
  - Work (022A $8 and $a)
  - Expression (039M $8 and $a)
  - Labels of the RVK (045R $j)
  - Keywords (044K $a)
  - LOC Keywords (044A $a)
- Hypothesis: Better results when more data is applied

5. Methods and Experiments

5. Methods and Experiments

- Two classification algorithms:
  a. **Multilingual BERT (mBERT):** (Devlin et al, 2019) which we fine-tune using a sigmoid loss function over the multiple classes
  b. **Support Vector Machines (linear)**: multilabel classifier, using a linear kernel

5. Methods and Experiments

- Input: vectorized representation of the raw metadata
- Tokenizer mBERT

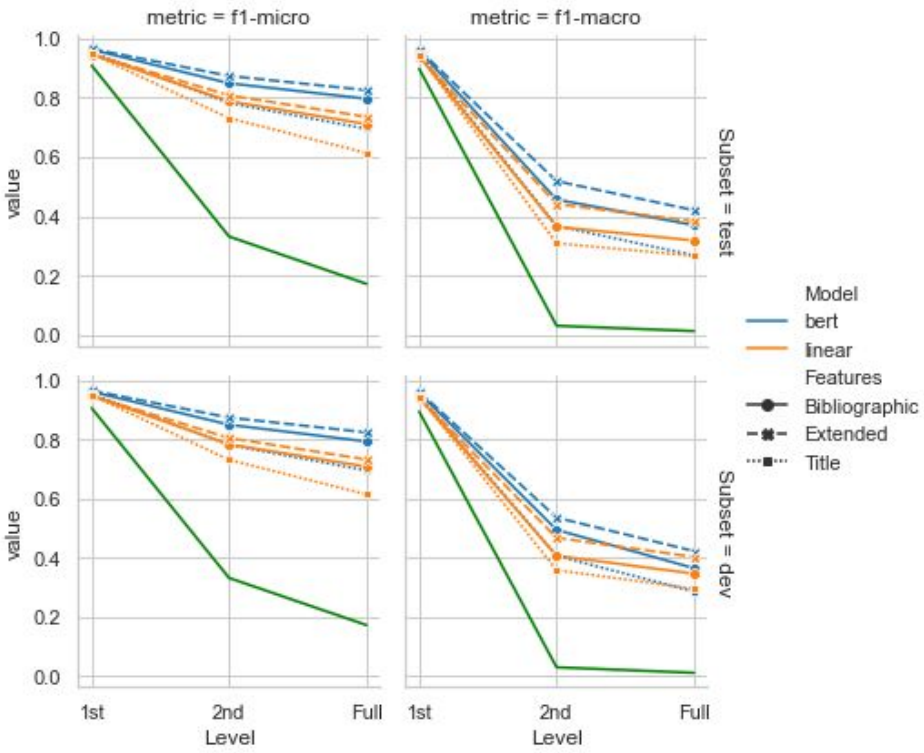
La @nuit tombée	'La', '@', 'nuit', 'tomb', '##ée'
Letteratura italiana contemporanea	'Letter', '##atura', 'italiana', 'contemporanea'
Politiques de l'amitié	'Pol', '##iti', '##ques', 'de', 'l', "'", 'amitié'
El @barco embrujado	'El', '@', 'barco', 'em', '##bru', '##jado'
Del texto a la iconografía	'Del', 'texto', 'a', 'la', 'i', '##cono', '##grafía'

5. Methods and Experiments

- Training, development and test splits with original distribution of labels
- F1-micro and F1-macro (considerable skewness)
- Random baseline following the distribution of labels
- Code is online available: https://github.com/morethanbooks/library_classification_rom

6. Results and Discussion

6. Results and Discussion

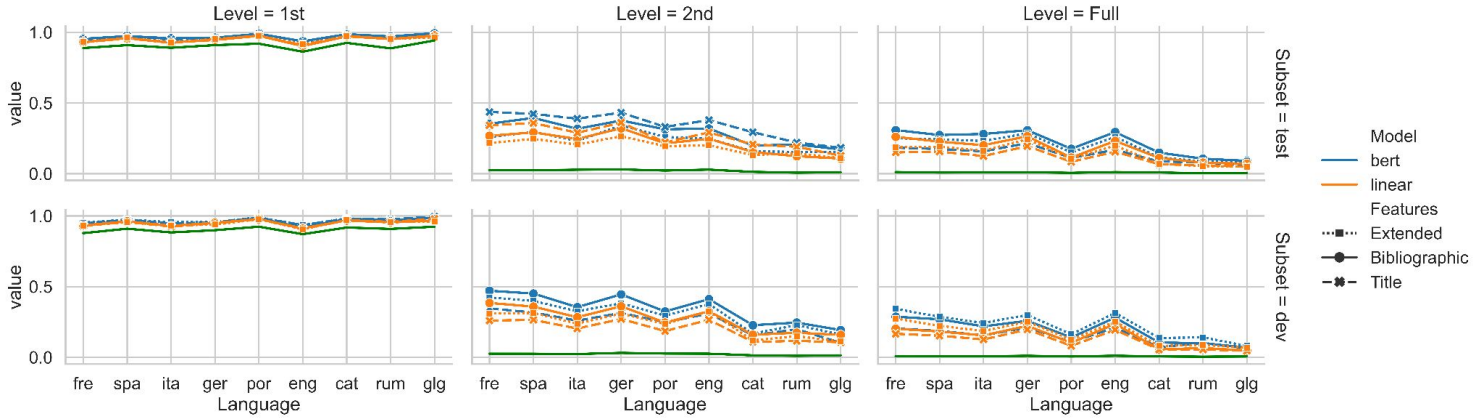


- **Levels of labels**: 1st too simple (17.00 and 18.00 classes)
- **Split**: Results in test and development are very similar
- **Algorithms**: mBERT obtains better results
- **Feature sets**: extended strongest performance, with basic bibliographic closer than expected

6. Results and Discussion: Languages

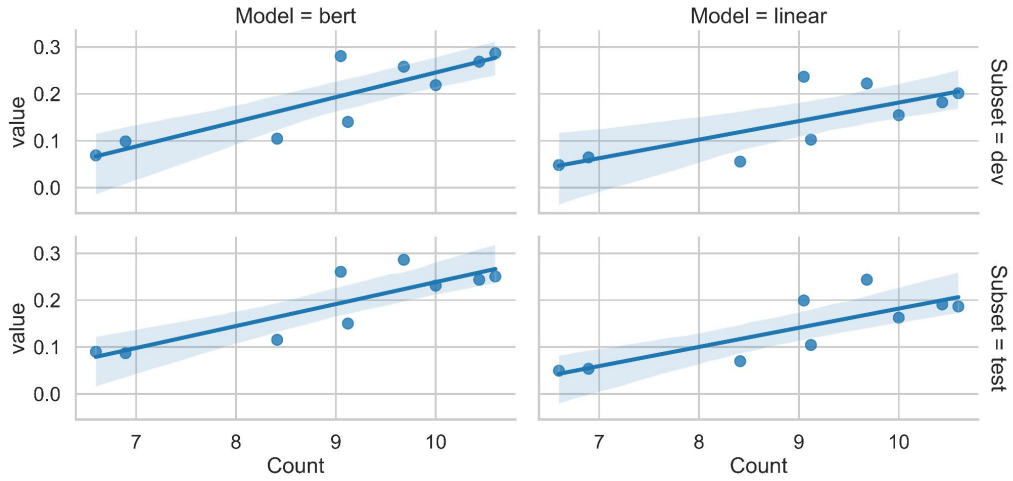
- Do the algorithms achieve higher scores for some languages?
  a. More publications → better classification results?
  b. Or more publications → more heterogeneity of data → worse classification results?

6. Results and Discussion: Languages



- Macro F1 scores
- Similar tendencies for all languages relating models and features
- Higher results for French, Spanish, German and English
- Correlation between number of publications and results?

6. Results and Discussion: Languages

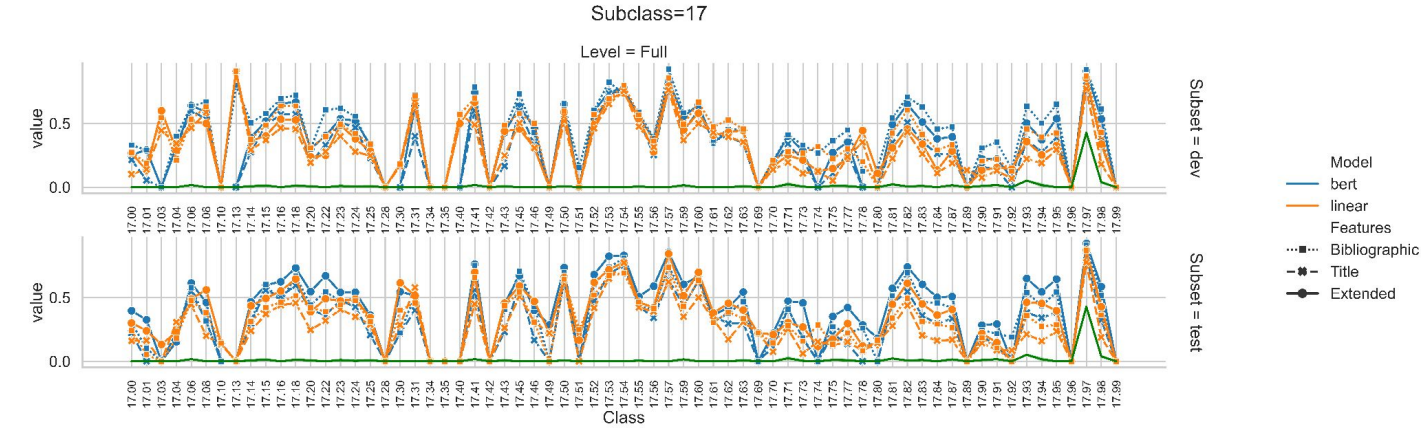


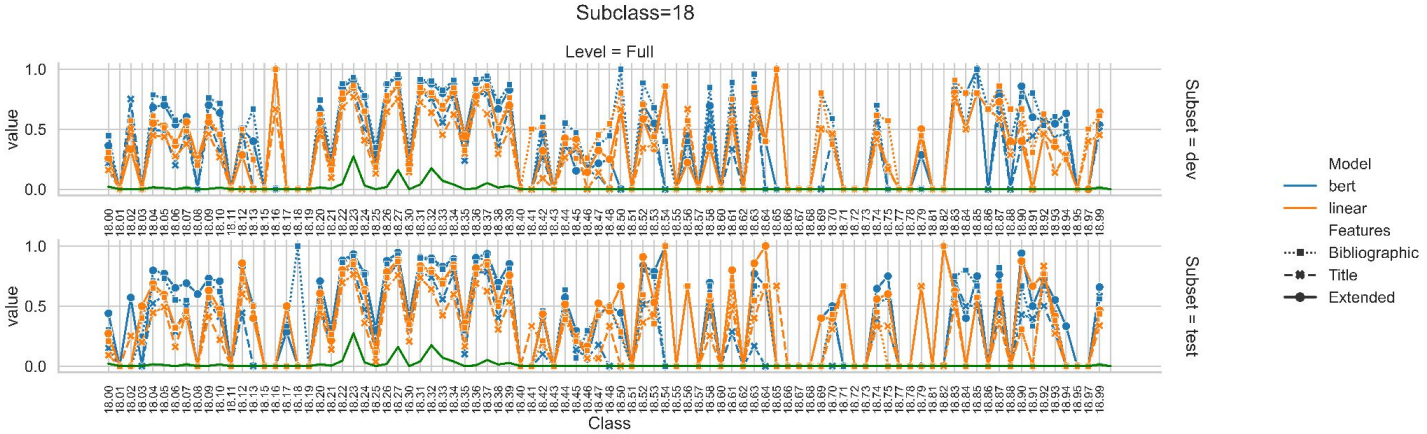
- Vertical axis = F1 score
- Horizontal axis = logarithmic number of instances
- Strong and very strong correlations (p-values < 0.05)
- Slightly stronger for mBERT than for linear model
- More data → better classification results
- **Reinforced already privileged languages**

6. Results and Discussion: Classes

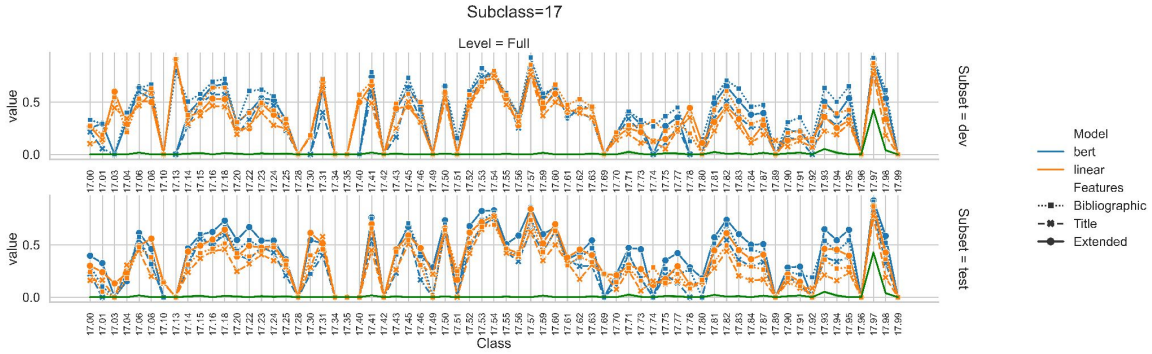
- Do the algorithms achieve higher scores for some classes?
- Supported by experience of subject librarians

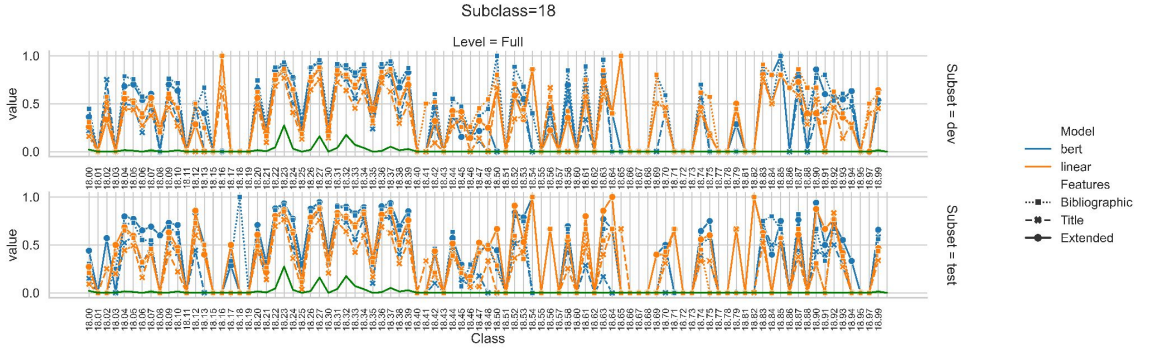
6. Results and Discussion: Classes





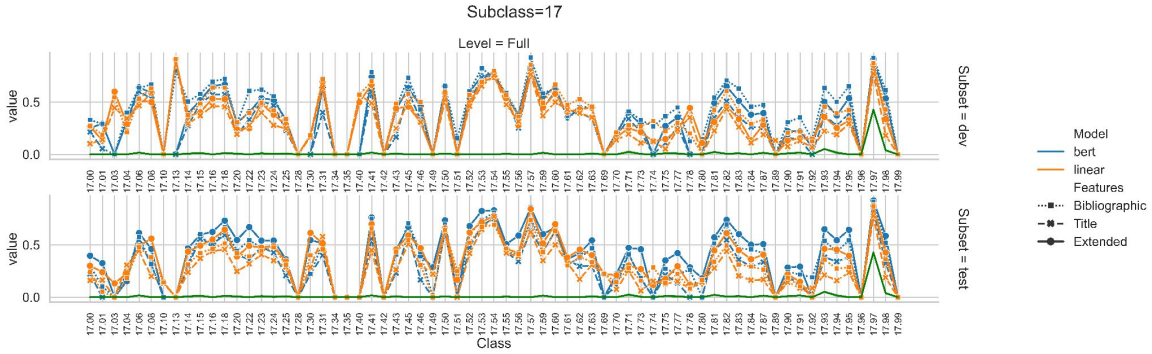
6. Results and Discussion: Classes

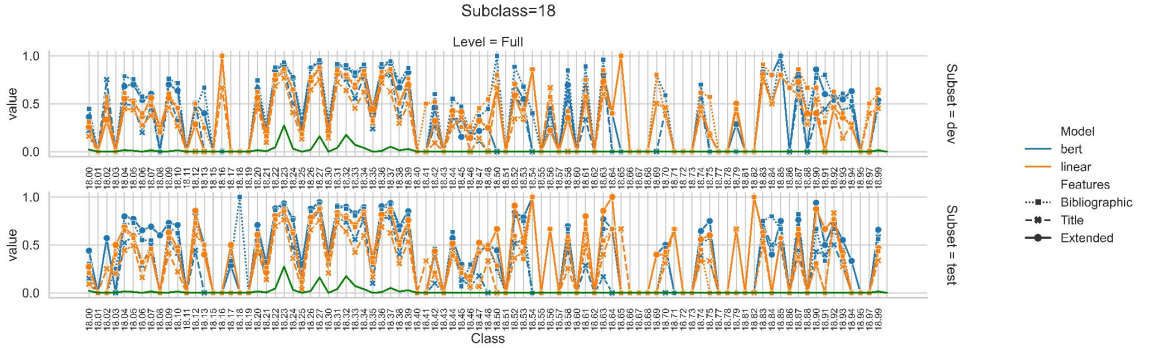




- Very high dispersion
- Opaque vs. clear classes for algorithms

6. Results and Discussion: Classes



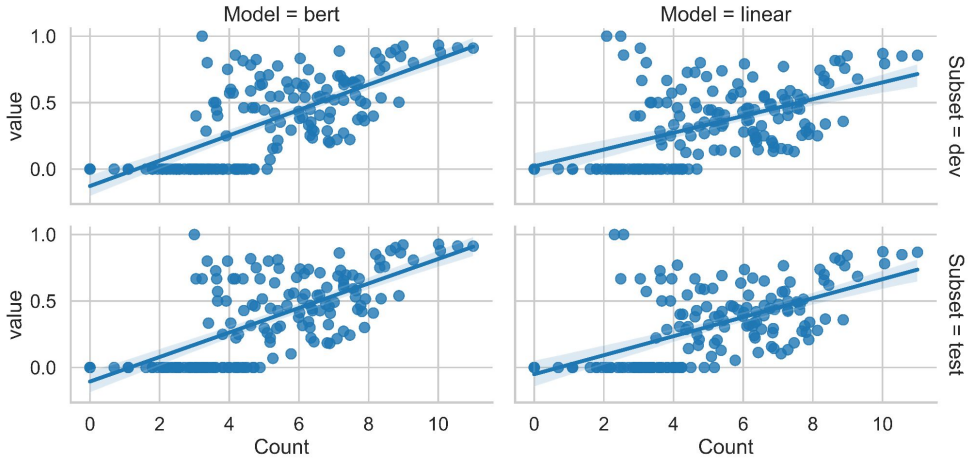


- In general, similar results for all classes
- Exceptions relating to the algorithm and features
- For a group of classes relating to Literary Studies, mBERT's gain is specially large
- Linear model works better with bibliographic features than with extended

6. Results and Discussion: Classes

- Classes with large number of publications → better scores?

6. Results and Discussion: Classes



- Few records: large dispersion
- Many records: high scores
- Statistical correlation
- Number of publications influences the results of the algorithms

8. Conclusions and Outreach

8. Conclusions and Outreach

- Only Romance Languages
- Macro F1 score around 0.4 (random baseline close to 0)
- Better results for mBERT
- The more publications, the better
- The more fields, the better
- Reasonable results with basic bibliographic data
- Basisklassifikation possibly better suited for ML tasks than other classification systems (DDC, RVK)

8. Conclusions and Outreach

- The majority of the catalogs contain publications in many languages
- Multilingual models
- Some languages obtain notably worse results
- Some classes are systematically mislabeled

8. Conclusions and Outreach

- Full-automatic subject indexing could be creating unacceptable metadata
  - for many classes
  - for many languages

8. Conclusions and Outreach

Stronger collaboration and more appreciation between classic librarian tasks and computational approaches!

9. References

Balakrishnan, Uma. 2016. “DFG-Projekt: Coli-conc. Das Mapping Tool ‘Cocoda.’” *o-bib. Das offene Bibliotheksjournal* 3 (1): 11–16. https://doi.org/10.5282/o-bib/2016H1S11-16.

Balakrishnan, Uma, and Jakob Voß. 2022. “Automatische Anreicherung der Sacherschließung des Verbundkatalogs K10plus mittels coli-rich.” In *#FreiräumeSchaffen*. Leipzig: Bibliothek und Information Deutschland. https://bid2022.abstractserver.com/program/#/details/presentations/27.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” *ArXiv:1810.04805 [Cs]*, May. http://arxiv.org/abs/1810.04805.

Kasprzik, Anna, and M. Fürneisen. 2022. “Aufbau eines produktiven Dienstes für die automatisierte Inhaltserschließung an der ZBW – ein Status- und Erfahrungsbericht.” In *#FreiräumeSchaffen*. Leipzig: Bibliothek und Information Deutschland. https://bid2022.abstractserver.com/program/#/details/presentations/22.

Kingma, Jelle. n.d. “Entstehungsgeschichte, Zweck Und Perspektiven Der Basisklassifikation in Den Niederlanden.” In *Aufbau Und Erschließung Begrifflicher Datenbanken: Beiträge Zur Bibliothekarischen Klassifikation ; Eine Auswahl von Vorträgen Der Jahrestagungen 1993 (Kaiserslautern) Und 1994 (Oldenburg) Der Gesellschaft Für Klassifikation*, 153–63.

Suominen, Osma. 2019a. “Annif: DIY Automated Subject Indexing Using Multiple Algorithms.” *LIBER Quarterly* 29 (1): 1–25. https://doi.org/10.18352/lq.10285.

———. 2019b. “Annif: DIY Automated Subject Indexing Using Multiple Algorithms.” *LIBER Quarterly: The Journal of the Association of European Research Libraries* 29 (1): 1–25. https://doi.org/10.18352/lq.10285.

Traiser, Walther. n.d. “Die Deutsche Bibliothek Und Die Basisklassifikation.” *Dialog Mit Bibliotheken* 7, 1995,2: 37–44.

Uhlmann, Sandro. 2022. “Automatische Inhaltserschließung an der Deutschen Nationalbibliothek.” In *#FreiräumeSchaffen*. Leipzig: Bibliothek und Information Deutschland. https://bid2022.abstractserver.com/program/#/details/presentations/100.

Zimmermann, Harald H. n.d. “Zur Struktur Und Nutzung von Klassifikationen Im Bibliothekswesen (Am Beispiel Der Klassifikation Der Deutschen Bibliothek Und Der Sog. Niederländischen Basisklassifikation).” In *Mehrwert von Information - Professionalisierung Der Informationsarbeit*, 187–200.

Thanks for the attention!

Vă mulțumim pentru atenție!

Merci pour l'attention!

Gràcies per la vostra atenció!

Obrigado pela atenção!

Grazie per l'attenzione!

Grazas pola súa atención!

¡Gracias por la atención!

Vielen Dank für die Aufmerksamkeit!