Maximilian Kähler

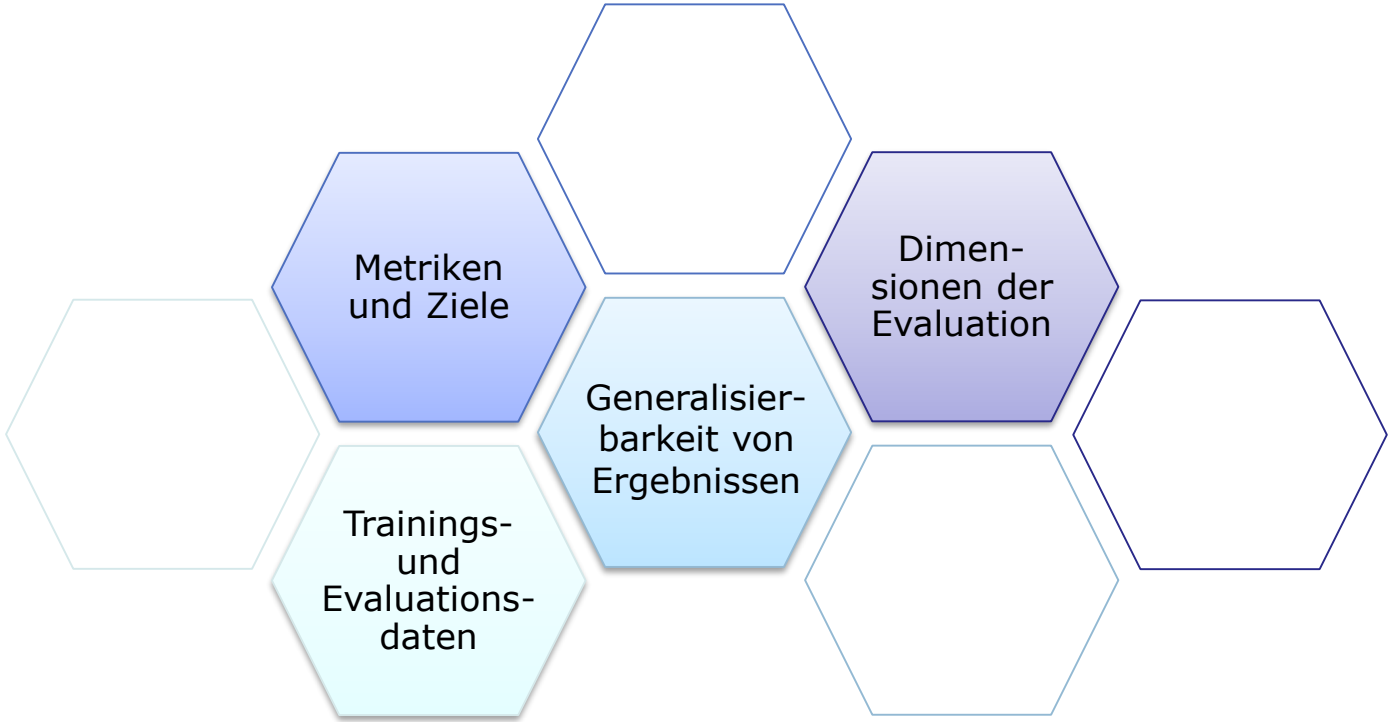
Evaluation im DNB-KI-Projekt

Drei ergänzende Wege zu einer besseren Beschlagwortung

- Entwicklung und Suche nach besseren Verfahren
- Vereinfachung des Problems
- **Besseres Erkennen von guter/ schlechter Qualität**

DEUTSCHE NATIONAL BIBLIOTHEK

Aspekte der Evaluation in ML-Projekten



Generalisierbarkeit von Evaluationsergebnissen

DEUTSCHE NATIONAL BIBLIOTHEK

Generalisierbarkeit von Ergebnissen

Die Messtheorie unterscheidet systematische und zufällige Messfehler:

Beispiele <u>systematischer</u> Fehler:

- Verteilungsunterschiede zwischen Trainings- und Produktionsdaten[1]
- “Information leakage” zwischen Trainings- und Evaluationsdaten

Beispiele <u>zufälliger</u> Fehler:

- Zufälliges Splitten von Trainings- und Evaluationsdaten
- ”Unbekannte” in einem komplexen Datenentstehungs-prozess

[1]vgl. Toepfer M, Seifert C 2020; Fusion architectures for automatic subject indexing under concept drift; https://doi.org/10.1007/s00799-018-0240-3

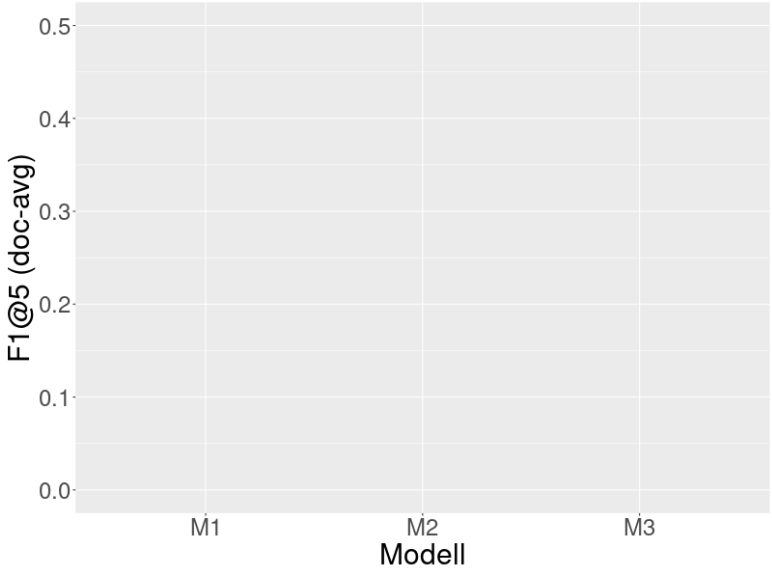
DEUTSCHE NATIONAL BIBLIOTHEK

Generalisierbarkeit von Ergebnissen

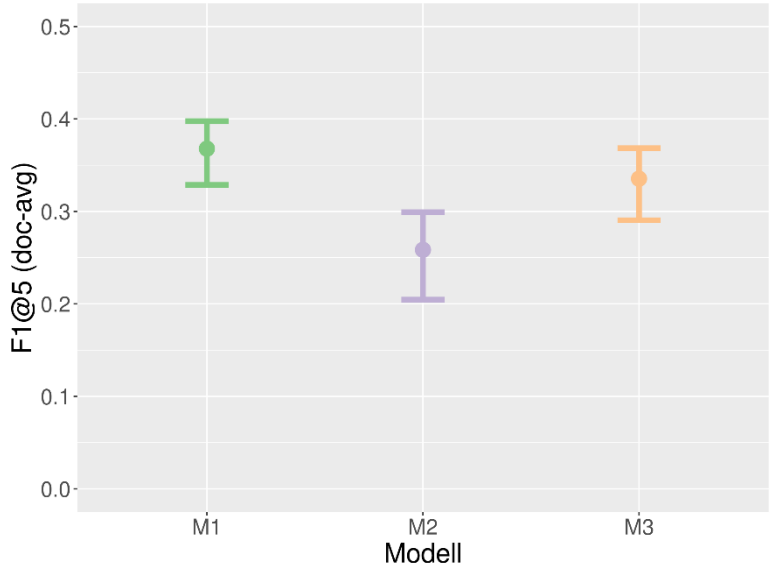
- Systematische Fehler können nur von Fall-zu-Fall behandelt werden
- Zufällige Messfehler werden durch **Konfidenzintervalle** quantifiziert
- Zufällige Messfehler hängen von der Größe des Test-Sets, und von der „Variabilität der Daten“ ab

Beispiel: Quantifizieren von Unsicherheit durch Bootstrap-Konfidenzintervalle

Wiederholtes Berechnen der Zielmetrik durch zufälliges Resampling des Test-Sets führt zu veräderten Ergebnissen mit einer empirischen Verteilung X



Perzentile aus der Verteilung X können als Konfidenzintervalle verwendet werden, um die Unsicherheit durch zufällige Fehler zu quantifizieren



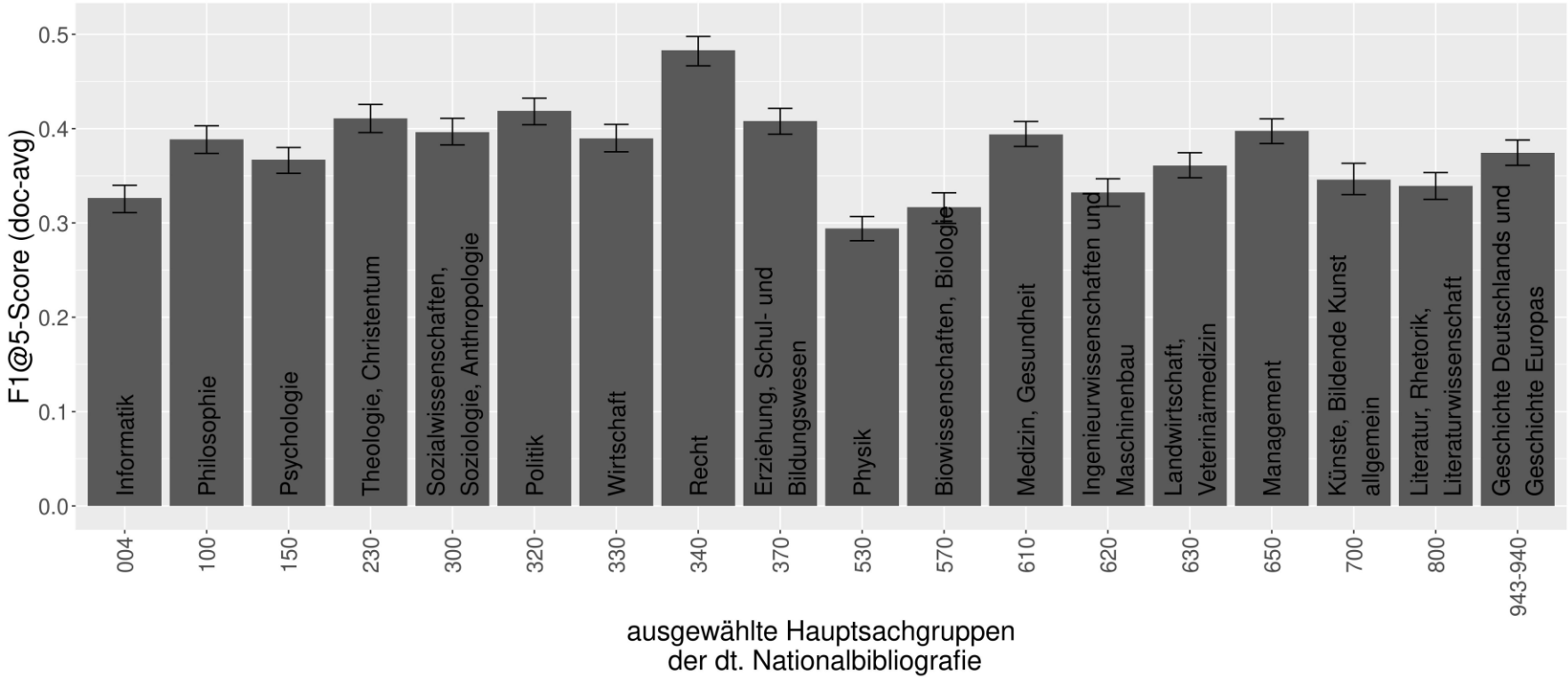
Dimensionen der Evaluation

DEUTSCHE NATIONAL BIBLIOTHEK

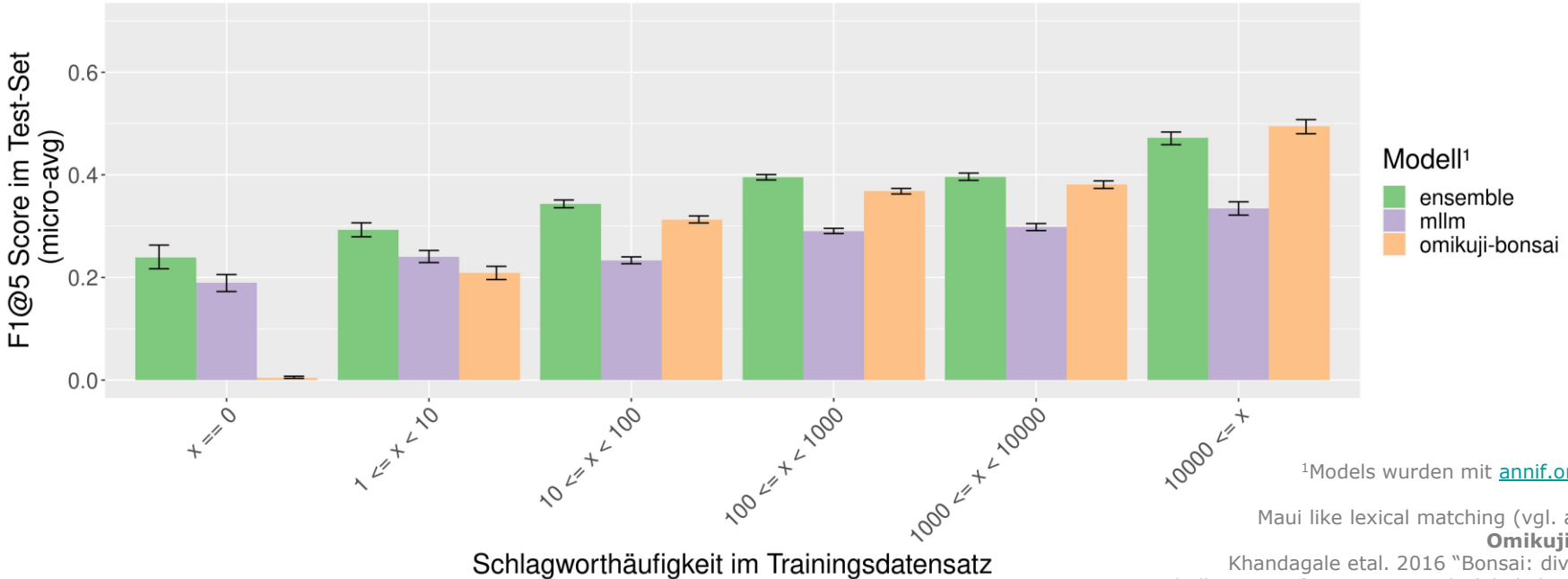
Dimensionen der Evaluation

- Ziel-Metriken zur Gesamt-Qualität ermöglichen keine Rückschlüsse über Ursachen von guter bzw. schlechter Qualität
- Eine sinnvolle Evaluation muss „Drill-Down"-Analyse der Ergebnisse ermöglichen, um Hypothesen zur weiteren Optimierung der Verfahren zu generieren
- Wichtige Dimensionen der Auswertung müssen mit Domänen-Wissen und Algorithmen-Wissen diskutiert werden

Beispiel: Stratifizierung von Ergebnissen aus Sicht der Dokumente



Beispiel: Stratifizierung von Ergebnissen aus Sicht des Vokabulars



[1]Models wurden mit annif.org erstellt

mllm:
Maui like lexical matching (vgl. annif.org)

Omikuji-bonsai:
Khandagale etal. 2016 "Bonsai: diverse and shallow trees for extreme multi-label classification"

DEUTSCHE NATIONAL BIBLIOTHEK

Berücksichtigung von Auswertungsdimensionen bei der Erstellung von Test-Sets

- Auswertungsdimensionen müssen bedacht werden bevor Daten in Trainings- und Evaluations-Daten aufgeteilt werden
- Die Gesamtgröße des Test-Sets hängt von der gewünschten Genauigkeit und dem Konfidenzniveau der Zielmetrik ab. Ausschlaggebend ist das kleinste Stratum einer Auswertungsdimension
- Mit gerichteten Stichproben kann sichergestellt werden, dass alle Strata in allen Auswertungsdimensionen ausreichend im Test-Set vertreten sind

DEUTSCHE NATIONAL BIBLIOTHEK

Summary: Where do we meet?

- Plan your evaluation scheme, before you start training models
- Discuss the dimensions of your data that need to be looked at
- Choose metrics that reflect your goals
- Discuss uncertainty and generalizability of your results

DEUTSCHE NATIONAL BIBLIOTHEK

Vielen Dank!

Fragen und Anregungen jederzeit an:

Maximilian Kähler

E-Mail: m.kaehler@dnb.de

Unser DNB-KI-Projekt: https://www.dnb.de/ki-projekt

DEUTSCHE NATIONAL BIBLIOTHEK

Related Work:

- **Golub etal 2016**; *A framework for evaluating automatic indexing or classification in the context of retrieval*. https://doi.org/10.1002/asi.23600
- **Raschka S 2020**; *Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning*. *https://arxiv.org/abs/1811.12808v3*
- **Toepfer M, Seifert C 2018**; *Content-Based Quality Estimation for Automatic Subject Indexing of Short Texts Under Precision and Recall Constraints*. https://doi.org/10.1007/978-3-030-00066-0_1