

# Automatisierte Typklassifikation von Normdaten mit BERT

## Hebeis, Maximilian

maximilian.hebeis@uni-bamberg.de  
Otto-Friedrich-Universität Bamberg, Lehrstuhl für  
Medieninformatik, An der Weberei 5, 96047 Bamberg,  
Deutschland  
ORCID: 0009-0008-2531-3616

## Fruth, Leon

leon.fruth@uni-bamberg.de  
Otto-Friedrich-Universität Bamberg, Lehrstuhl für  
Medieninformatik, An der Weberei 5, 96047 Bamberg,  
Deutschland  
ORCID: 0009-0001-2128-3025

## Gradl, Tobias

tobias.gradl@uni-bamberg.de  
Otto-Friedrich-Universität Bamberg, Lehrstuhl für  
Medieninformatik, An der Weberei 5, 96047 Bamberg,  
Deutschland  
ORCID: 0000-0002-1392-2464

## Henrich, Andreas

andreas.henrich@uni-bamberg.de  
Otto-Friedrich-Universität Bamberg, Lehrstuhl für  
Medieninformatik, An der Weberei 5, 96047 Bamberg,  
Deutschland  
ORCID: 0000-0002-5074-3254

## Motivation

Normdaten spielen eine zentrale Rolle in den Digital Humanities. Durch eindeutige Identifikatoren für Entitäten wie Personen, Orte, oder Organisationen tragen sie wesentlich zur Wiederverwendbarkeit und Interoperabilität geisteswissenschaftlicher Datensätze bei (Busch und Müller 2023). Das an der Otto-Friedrich-Universität Bamberg entwickelte *Authority Data Integration Search System* (ADISS) verknüpft verschiedene Normdateien und Wissensdatenbanken, darunter die GND, Wikidata, OpenStreetMap und Geonames. Die einzelnen Datensätze werden dabei in ein integriertes Schema überführt, um eine einheitliche Suche über verschiedene Quellen hinweg zu ermöglichen. Projekte wie die Text+ Registry<sup>1</sup> und Oral-History.Digital<sup>2</sup> nutzen ADISS, um Forschungsdaten semi-automatisch oder manuell mit Normdaten zu annotieren (Fruth u. a. 2025). Hinsichtlich der Typisierung der durch die Normdaten be-

schriebenen Entitäten erfolgt diese Integration bisher jedoch nur oberflächlich: Die Typenangabe wird aus der jeweiligen Quelldatenbank als Freitext übernommen. So erscheinen beispielsweise Normdatensätze zu „Bamberg“ in ADISS mit den Typen *TerritorialCorporateBodyOrAdministrativeUnit* (GND), *college town*, *major regional center*, ... (Wikidata) oder *city*, *village* (Geonames).

Eine Facettierung der Suche in ADISS könnte die Relevanz und Nutzbarkeit der Suchergebnisse für Endnutzer:innen deutlich erhöhen: Je nach Anwendungskontext wäre eine gezielte Eingrenzung auf bestimmte Entitätsklassen (etwa Geografika, Personen oder Organisationen) hilfreich, um die Treffermenge zu präzisieren. Ein manuelles Mapping der bestehenden Typinformationen wird jedoch durch die heterogenen Schemata der eingebundenen Datenquellen erschwert: Während klassische Normdateien wie die GND auf rigide Ontologien setzen, verfolgen offene Wissensdatenbanken wie Wikidata oder OpenStreetMap einen kollaborativen Ansatz, bei dem Nutzer:innen neue Typen selbst definieren können. Für die Umsetzung einer effektiven Facettensuche wäre daher eine Zuordnung der in ADISS enthaltenen Normdatensätze zu einem reduzierten, übergreifenden Zielschema erforderlich.

## Ziel

Im Rahmen dieses Beitrags soll ein System zur automatisierten Typisierung von Normdatensätzen aus Datenbanken mit offenen Ontologien vorgestellt werden. Hierbei werden Methoden des überwachten maschinellen Lernens aus dem Bereich der Textklassifikation genutzt. Algorithmen zur Worteinbettung und besonders Sprachmodelle auf Basis der Transformer-Architektur wie BERT (Devlin u. a. 2019) wurden in der jüngeren Vergangenheit angewendet, um Texte verschiedener Domänen zu klassifizieren (Cevik u. a. 2023; Chi u. a. 2023). Speziell zeigen Arbeiten wie die von (Peng u. a. 2023), dass BERT-Modelle auch für das automatisierte Mapping zwischen heterogenen Datenschemata und Ontologien geeignet sind. Aufgrund der hierarchischen Struktur von Typenontologien sind Methoden der hierarchischen Textklassifikation von besonderem Interesse, da Entitäten so grob- und feingranular typisiert werden könnten: Auch hier existieren eine Vielzahl an Vorarbeiten in verschiedenen Datendomänen (Zangari u. a. 2024; Zhang u. a. 2022; Biswas u. a. 2022; Chen u. a. 2020). In einem ersten Schritt wurde prototypisch ein Zielschema und Workflow für die automatisierte Typklassifikation von Normdaten erstellt und auf Wikidata-Datensätzen evaluiert.

## Datenbasis und Workflow

Aufgrund der thematisch breit gefächerten Typologie wurde eine Untermenge von Schema.org<sup>3</sup> als geeignetes integriertes Zielschema ausgewählt. Um eine Abbildung des größten Teils der relevanten Normdaten zu gewährleisten, wurden zusätzlich noch zwei Typen für Wege, Straßen, und

Grenzen (*PathOrBoundary*) und Personengruppen (*GroupOfPersons*) definiert. Ein Ausschnitt aus dem so entstandenen Zielschema mit insgesamt 33 Typen ist in Abbildung 1 dargestellt.

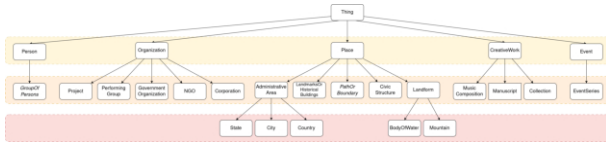


Abbildung 1: Ein Ausschnitt aus dem Zielschema für die Typklassifikation, basierend auf schema.org.

Für das Trainieren eines auf einem BERT-Modell basierenden Klassifikators werden annotierte Trainingsdaten benötigt. Hierfür werden bestehende Verlinkungen zwischen GND- und Geonames-Normdaten auf der einen Seite und Wikidata-Datensätzen auf der anderen Seite genutzt: Auf Grundlage eines selbst erstellten Mappings zwischen der GND- und Geonames-Ontologie und den Entitätsklassen unseres Zielschemas kann so jeder verlinkte Wikidata-Datensatz mit einem Typenlabel versehen werden. Diese mit einem Zieltypen versehenen Datensätze können dann als Trainings- und Testmaterial für den Klassifikator dienen. Eine Abfrage der Typenstatistik der in ADISS verfügbaren Wikidata-Datensätze, die mit GND- oder Geonames-Normdaten verlinkt sind, offenbart das starke Klassenungleichgewicht als eine erste Hürde bei der automatisierten Typisierung: Während Normdaten zu Personen dominieren, ist das Datenmaterial zu einigen Typen wie *Manuscript*, *Project* und *Collection* stark unzureichend (siehe Tabelle 1).

Tabelle 1: Anzahl an mit GND oder Geonames verlinkten Wikidata-Entitäten in ADISS.

Typ	Anzahl Verlinkungen Wikidata-GND/Geonames
Person	1 610 178
Mountain	1 335 944
BodyOfWater	1 308 570
City	693 214
CivicStructure	395 805
AdministrativeArea	292 831
Organization	46 481
Landform	34 395
Thing	28 663
CreativeWork	17 043
LandmarksOrHistoricalBuildings	12 701
Event	5 457
MusicComposition	4 961
PerformingGroup	4 433
Corporation	3 616
GovernmentOrganization	3 475
EventSeries	2 433
GroupOfPersons	2 421
NGO	1 582
Place	1 532
PathOrBoundary	1 197
Language	905
State	830
Brand	651
Country	536
SoftwareApplication	521
BioChemEntity	507
Manuscript	228
Vehicle	177
Project	59
Collection	21

Als Eingabefeatures für das Modell dienen Name, Beschreibung und Quelltypen (Wikidata-Typen) der jeweiligen Entität. Aufgrund des eher flachen Zielschemas wurden *local per-level classifiers* als Modellarchitektur gewählt (Zangari u. a. 2024): Ein multilinguales DistilBERT-Modell<sup>4</sup> dient als Encoder für die Eingabefeatures; drei *classification heads* (Multi-Layer-Perzeptrene) ordnen die so semantisch eingebettete Instanz anschließend jeweils einer der drei Ebenen der Typenhierarchie zu (siehe Abbildung 2).

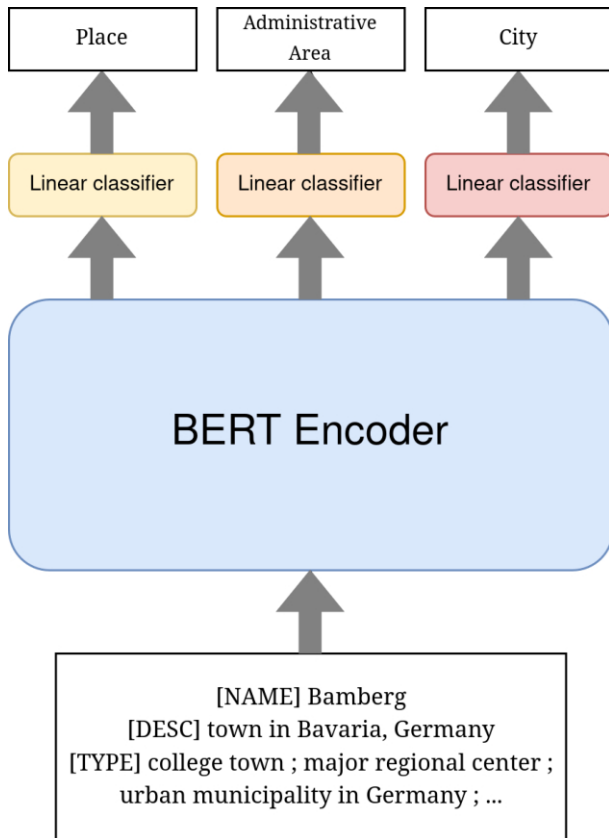


Abbildung 2: Die Modellarchitektur des Typklassifikators.

## Ausblick

Erste Experimente wurden auf einem Datensatz mit 139 011 Normdateninstanzen durchgeführt (mit bis zu 4 000 Instanzen pro Typ), bei einem Trainings-/Testsplit von 4:1. Bei unterrepräsentierten Typen wurden fehlende Wikidata-Instanzen durch zusätzliche unverlinkte GND- und Geonames-Datensätze ergänzt, um eine robuste Trainingsdatenbasis zu schaffen. Diese erste Evaluation liefert vielversprechende Makro-F1-Werte über alle Hierarchieebenen hinweg, mit leicht abfallender Performanz bei zunehmender Klassifikationstiefe (siehe Tabelle 2).

Tabelle 2: Evaluationsergebnisse für die automatisierte Typklassifikation von Wikidata-Datensätzen.

Hierarchieebene	Anzahl der Klassen (N)	Macro F1 ( % )
1	8	93,41
2	19	89,31
3	5	87,97

In einem nächsten Schritt soll die Generalisierbarkeit auf unverlinkte Wikidata-Datensätze geprüft werden. Eine Anpassung oder Vereinfachung des Zielschemas könnte zudem für eine Verbesserung der Modellleistung führen, da die geringe Trennschärfe zwischen semantisch eng verwandten Klassen zu Fehlklassifikationen führen kann. Weiterhin soll untersucht werden, ob *local per-parent classi-*

*fiers* die Performanz zusätzlich erhöhen können (Zangari u. a. 2024).

## Fußnoten

1. <https://registry.text-plus.org/>
2. <https://www.oral-history.digital/>
3. <https://www.schema.org/docs/full.html>
4. <https://huggingface.co/distilbert/distilbert-base-multi-lingual-cased>

## Bibliographie

- Biswas, Russa, Jan Portisch, Heiko Paulheim, Harald Sack, und Mehwish Alam.** 2022. „Entity Type Prediction Leveraging Graph Walks and Entity Descriptions“. In *The Semantic Web – ISWC 2022*, herausgegeben von Ulrike Sattler, Aidan Hogan, Maria Keet, u. a. Springer International Publishing. [https://doi.org/10.1007/978-3-031-19433-7\\_23](https://doi.org/10.1007/978-3-031-19433-7_23).
- Busch, Nathanael, und Diana Müller.** 2023. „Normdaten in den Geisteswissenschaften“. *Zeitschrift für Literaturwissenschaft und Linguistik* 53 (3): 781–96. <https://doi.org/10.1007/s41244-023-00295-1>.
- Cevik, Mucahit, Savas Yildirim, Devang Parikh, und Ayse Basar.** 2023. „Adaptive Fine-tuning for Multiclass Classification over Software Requirement Data“. Preprint, Research Square. <https://doi.org/10.21203/rs.3.rs-2434133/v1>.
- Chen, Tongfei, Yunmo Chen, und Benjamin Van Durme.** 2020. „Hierarchical Entity Typing via Multi-Level Learning to Rank“. arXiv:2004.02286. Preprint, arXiv. <https://doi.org/10.48550/arXiv.2004.02286>.
- Chi, Te-Yu, Yu-Meng Tang, Chia-Wen Lu, Qiu-Xia Zhang, und Jyh-Shing Roger Jang.** 2023. „WC-SBERT: Zero-Shot Text Classification via SBERT with Self-Training for Wikipedia Categories“. arXiv:2307.15293. Preprint, arXiv. <https://doi.org/10.48550/arXiv.2307.15293>.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, und Kristina Toutanova.** 2019. „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding“. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, herausgegeben von Jill Burstein, Christy Doran, und Tamar Solorio. Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>.
- Fruth, Leon, Tobias Gradl, Maximilian Hebeis, und Andreas Henrich. 2025. „A Flexible Search System for Integrated Authority Data—ADISS“. *Datenbank-Spektrum, Online-Vorab-Publikation*. <https://doi.org/10.1007/s13222-025-00515-7>.
- Peng, Yiwen, Mehwish Alam, und Thomas Bonald.** 2023. „Ontology Matching Using Textual Class

Descriptions“. Conference paper presented auf The 18th International Workshop on Ontology Matching co-located with the 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece. <https://telecom-paris.hal.science/hal-04459105>.

**Zangari, Alessandro, Matteo Marcuzzo, Matteo Rizzo, Lorenzo Giudice, Andrea Albarelli, und Andrea Gasparetto.** 2024. „Hierarchical Text Classification and Its Foundations: A Review of Current Research“. *Electronics* 13 (7): 7. <https://doi.org/10.3390/electronics13071199>.

**Zhang, Shu, Ran Xu, Caiming Xiong, und Chetan Ramaiah.** 2022. „Use All The Labels: A Hierarchical Multi-Label Contrastive Learning Framework“. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 16639–48. <https://doi.org/10.1109/CVPR52688.2022.01616>.