Detecting Occupations in German Texts: Challenges and Data

Thomas Reiser[1,*,†], Jens Dörpinghaus[1,2,3,*,†] and Petra Steiner[2]

[1]University of Koblenz, Department of Computer Science, Germany
[2]Federal Institute for Vocational Education and Training (BIBB), Bonn, Germany
[3]Linnaeus University, Department of Computer Science and Media Technology, Växjö, Sweden

Abstract

This paper is concerned with the detection and classification of occupational titles in German texts, with a focus on the linguistic and structural challenges that are unique to the German language. The study utilizes an extensive dataset on job title variants to assess rule-based and language model-based methodologies across diverse corpora, encompassing historical documents and parliamentary proceedings. The findings indicate that rule-based methods demonstrate robust performance, particularly in structured texts, while large language models exhibit complementary strengths in recognizing complex terms. The work makes a significant contribution to the field by providing valuable annotated data and methodological insights.

Keywords

Text analysis, NER, name detection, computational social sciences

1. Introduction

The automated classification of job titles is a significant topic in both academic research and for practitioners in labor market analysis. It is also applicable to many other use cases, including occupations in literature and other texts, such as parliamentary debates [1]. In certain applications, such as surveys, it is necessary to map data to standardized classifications. Two prominent examples of such classifications are the German Classification of Occupations (KldB) and the International Standard Classification of Occupations (ISCO). Additional use cases include the classification of online job advertisements (OJAs) and the alignment of occupational titles from other sources, such as online platforms like Kununu [2].

The German language poses unique challenges in the context of occupation classification. Occupational titles may manifest not only as single nouns but also as complex noun phrases, potentially including information such as certifying institutions (e.g., IHK). Furthermore, there are gender-specific and gender-neutral variants of titles. A further complication arises in general texts from surnames that derive from historical professions. One notable example is the surname “Bäcker”, which is derived from the German word for “baker”.

In previous studies, we expanded upon the incorporation of job titles derived from online job advertisements and vocational education and training (VET) titles to detect particular digits and exactly match labels and single occupations (see [3]). Nevertheless, this is not invariably feasible and is only requisite in specific instances. Consequently, the present study explores alternative research inquiries. The following list contains the items in question:

- The following question is posed for consideration: What methodology could be employed to efficiently map occupational titles in German texts to the German classification of occupations?
- The following inquiry seeks to ascertain whether divergent textual categories present distinctive challenges.

AISEER ’25: 2nd International Workshop on AI in Society, Education and Educational Research, Bologna, Italy
*Corresponding author.
†These authors contributed equally.
treiser@uni-koblenz.de (T. Reiser); doerpinghaus@uni-koblenz.de (J. Dörpinghaus); steiner@bibb.de (P. Steiner)
0009-0007-5452-7800 (T. Reiser); 0000-0003-0245-7752 (J. Dörpinghaus)

In addressing the second research question, the focus is directed towards two distinct datasets. The first dataset comprises German parliamentary debates. The second dataset encompasses GDR job descriptions (Berufsbilder). The principal contribution of this paper is a substantial collection of novel training data and a systematic analysis of the challenges associated with the detection of German job titles according to KldB.

The present document is organized into five sections.

2. Related Work

The integration of diverse labor market data sources is widely acknowledged as a multifaceted undertaking [4]. However, the present study focuses on the mapping and automated classification of job titles in the German language. While dictionary-based methods are commonly employed, machine learning (ML)-based approaches have also been explored. Existing training datasets are often compiled from survey responses or classification systems, including KldB, ESCO, and other synonym collections.

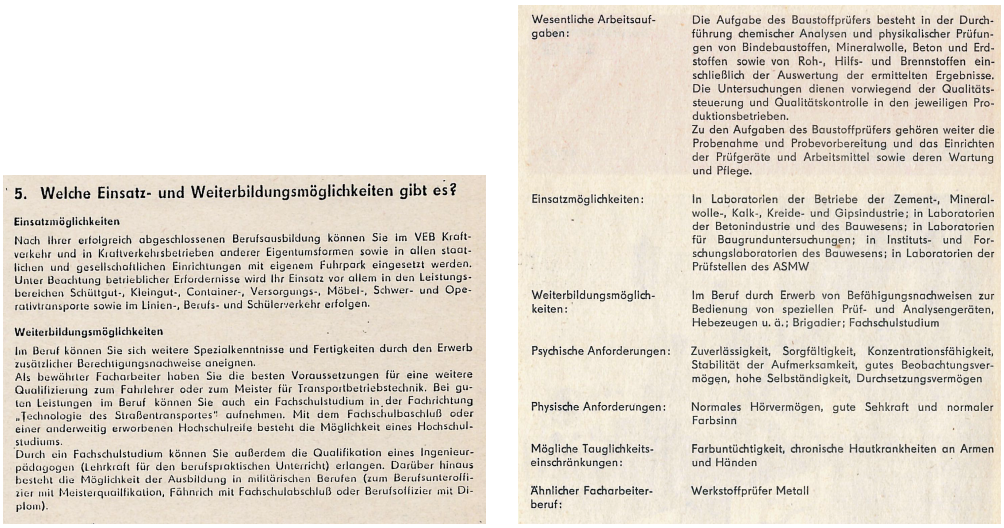
We follow our study of automated classifications presented in [3]. A multitude of classification categories are recognized for occupations. The International Standard Classification of Occupations (ISCO) was developed by the International Labor Organization (ILO) and published in 1958, 1968, 1988, and most recently in 2008)[1]. The ISCO 2008 has also been utilized within the European Union (EU), with certain German-speaking countries (Germany, Austria, and Switzerland) developing a customized version of the classification. The International Standard Classification of Occupations (ISCO) is structured at a skill level and linked to the "European Skills, Competences, Qualifications and Occupations" (ESCO) ontology, which adds another hierarchy level to the data. In Germany, the Classification of Occupations (KldB) serves as the reference classification for the Federal Employment Agency (BA) and its research institute (IAB)[2]. In this organization, occupations are structured at a task level. The most recent version is the 2020 revision of the KldB 2010, which has undergone a comprehensive redesign, thereby rendering the previous versions from 1988 and 1992 obsolete. The development of this system was undertaken with the objective of ensuring compatibility with the ISCO-08 standard. The study of job titles and taxonomies has a long history, extending even before the advent of computer technology [5].

A portion of the research has focused on the classification of OJAs according to the O*NET framework [6]. This has included the application of normalization approaches [7] and similarity-based methods [8]. The classification of job titles is also employed in the context of online job recruitment [9].

A limited number of publications have been published on the subject of German job titles, with a particular emphasis on the German KldB. For instance, a technical report based on OJAs [10] with challenges on level 4, but with promising results on level 1. Malte Schierholz's 2018 publication [11] introduced the concept of auxiliary classifications in the field of occupational coding. For further research on the subject of occupational coding in surveys, we refer to [12]. A master's thesis endeavors to predict KldB 5-digit job titles from survey data, thereby highlighting the persistent challenges associated with this endeavor, see [13]. In a similar vein, a scholarly article was published that compared the classification of survey data using BERT and GPT-3, see [14]. However, the absence of a standardized reporting methodology precludes the direct comparison of their results. Nevertheless, they evince analogous challenges to those observed in other studies. Our previous study [3] lends support to this assertion, particularly in terms of the conclusion that large language models (LLMs) are not capable of enhancing the quality of automated classifiers to a significant extent. Consequently, both the classification of occupational areas and, in particular, the level of performance (5th digit) persist as arduous tasks.

[1] See https://www.ilo.org/public/english/bureau/stat/isco/isco08/.

[2] See https://statistik.arbeitsagentur.de/DE/Navigation/Grundlagen/Klassifikationen/Klassifikation-der-Berufe/KldB2010-Fassung2020/KldB2010-Fassung2020-Nav.html.



5. Welche Einsatz- und Weiterbildungsmöglichkeiten gibt es?

Einsatzmöglichkeiten

Nach Ihrer erfolgreich abgeschlossenen Berufsausbildung können Sie im VEB Kraftverkehr und in Kraftverkehrsbetrieben anderer Eigentumsformen sowie in allen staatlichen und gesellschaftlichen Einrichtungen mit eigenem Fuhrpark eingesetzt werden. Unter Beachtung betrieblicher Erfordernisse wird Ihr Einsatz vor allem in den Leistungsbereichen Schüttgut-, Kleingut-, Container-, Versorgungs-, Möbel-, Schwer- und Operativtransporte sowie im Linien-, Berufs- und Schülerverkehr erfolgen.

Weiterbildungsmöglichkeiten

Im Beruf können Sie sich weitere Spezialkenntnisse und Fertigkeiten durch den Erwerb zusätzlicher Berechtigungsnachweise aneignen.
Als bewährter Facharbeiter haben Sie die besten Voraussetzungen für eine weitere Qualifizierung zum Fahrlehrer oder zum Meister für Transportbetriebstechnik. Bei guten Leistungen im Beruf können Sie auch ein Fachschulstudium in der Fachrichtung „Technologie des Straßentransportes" aufnehmen. Mit dem Fachschulabschluß oder einer anderweitig erworbenen Hochschulreife besteht die Möglichkeit eines Hochschulstudiums.
Durch ein Fachschulstudium können Sie außerdem die Qualifikation eines Ingenieurpädagogen (Lehrkraft für den berufspraktischen Unterricht) erlangen. Darüber hinaus besteht die Möglichkeit der Ausbildung in militärischen Berufen (zum Berufsunteroffizier mit Meisterqualifikation, Fähnrich mit Fachschulabschluß oder Berufsoffizier mit Diplom).

Wesentliche Arbeitsaufgaben:	Die Aufgabe des Baustoffprüfers besteht in der Durchführung chemischer Analysen und physikalischer Prüfungen von Bindebaustoffen, Mineralwolle, Beton und Erdstoffen sowie von Roh-, Hilfs- und Brennstoffen einschließlich der Auswertung der ermittelten Ergebnisse. Die Untersuchungen dienen vorwiegend der Qualitätssteuerung und Qualitätskontrolle in den jeweiligen Produktionsbetrieben. Zu den Aufgaben des Baustoffprüfers gehören weiter die Probenahme und Probevorbereitung und das Einrichten der Prüfgeräte und Arbeitsmittel sowie deren Wartung und Pflege.
Einsatzmöglichkeiten:	In Laboratorien der Betriebe der Zement-, Mineralwolle-, Kalk-, Kreide- und Gipsindustrie; in Laboratorien der Betonindustrie und des Bauwesens; in Laboratorien für Baugrunduntersuchungen; in Instituts- und Forschungslaboratorien des Bauwesens; in Laboratorien der Prüfstellen des ASMW
Weiterbildungsmöglichkeiten:	Im Beruf durch Erwerb von Befähigungsnachweisen zur Bedienung von speziellen Prüf- und Analysengeräten, Hebezeugen u. ä.; Brigadier; Fachschulstudium
Psychische Anforderungen:	Zuverlässigkeit, Sorgfältigkeit, Konzentrationsfähigkeit, Stabilität der Aufmerksamkeit, gutes Beobachtungsvermögen, hohe Selbständigkeit, Durchsetzungsvermögen
Physische Anforderungen:	Normales Hörvermögen, gute Sehkraft und normaler Farbsinn
Mögliche Tauglichkeitseinschränkungen:	Farbuntüchtigkeit, chronische Hautkrankheiten an Armen und Händen
Ähnlicher Facharbeiterberuf:	Werkstoffprüfer Metall

Figure 1: GDR job descriptions for Berufskraftfahrer (1987, B 58 2 01, left) and Baustoffprüfer (1985, B 54 2 02, right)

3. Data and Methods

3.1. Data

To identify job titles, an extensive dataset was utilized, encompassing 526,535 synonyms and variants of male, female, and neutral job titles. This dataset was provided by the German Federal Employment Agency (BA)[3]. For illustrative purposes, the following examples are given: “Meister – Maßschneiderei” and “Herrenschneidermeisterin”, both of which link to KldB 28293-901. However, it should be noted that the first five digits are utilized exclusively, and a link to 28293 is established. The presence of numerous duplicates results in a non-unique linkage. The most prominent example is the term “Meister” (Master), which is linked to nearly all crafts. Furthermore, the dataset under consideration contains terms associated with occupations. For instance, the terms “Kohle”, “Naturwerkstein”, and “Anlagenführung” are linked to 21212-129. Consequently, this dataset can also be utilized to identify any implicit relations in a text that link to occupations. Nevertheless, for the aforementioned approach, a blacklist of these terms was created. While these terms are classified under the gender-neutral section, the removal of these terms would also result in the removal of all gender-neutral occupational titles.

The initial test dataset comprises job descriptions (Berufsbilder) utilized in the former German Democratic Republic (GDR) for the purpose of vocational guidance, see [15, 16]. These documents are available for a variety of occupations and contain eight pages of information regarding requirements, the content of (vocational) education, skills, and workplaces. Furthermore, these documents often include information regarding additional training opportunities (Weiterbildungsmöglichkeiten). However, as shown in Figure 1, the layout and content varies through the years. A preliminary investigation of job postings reveals a heterogeneity in the occupational requirements. While some positions do not specify additional occupations or training prerequisites, other positions do. For instance, the job description for “Berufskraftfahrer” (professional driver) includes qualifications such as “Fahrlehrer” (driving instructor), “Meister für Transportbetriebstechnik”, and even university degrees. Therefore, in order to methodically ascertain the interrelationships between disparate occupations, it is necessary to assess which other occupations are referenced.

The Corpus of Plenary Proceedings of the German Bundestag (CPP-BT) is another dataset that will be utilized in this study, see [17]. This data set consists of 4,566 plenary proceedings of the German Bundestag. The document under consideration is a compilation of all plenary minutes from the first legislative period until the twenty-fourth. May 2025. The initial XML data was retrieved from the Open

[3]Available at https://www.arbeitsagentur.de/institutionen/dkz-downloadportal.

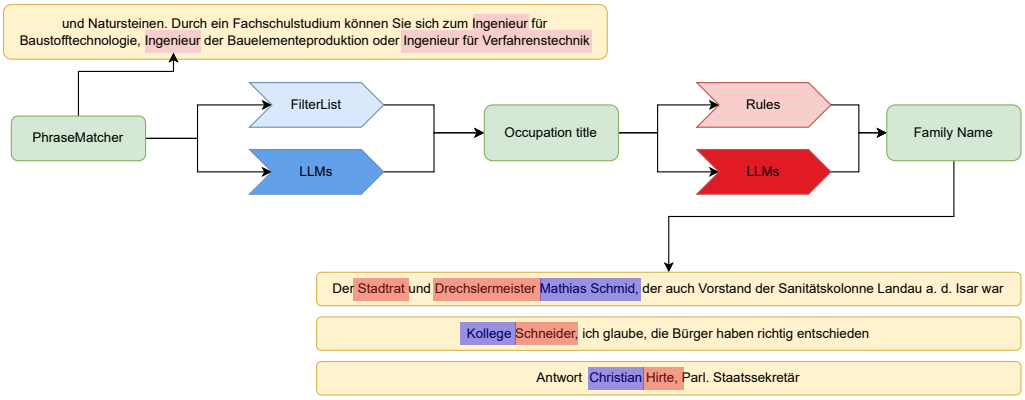


Figure 2: Proposed workflow

Data Portal of the German Bundestag and the Documentation and Information System for Parliamentary Materials (DIP) up to the respective cut-off date.

While the initial dataset does not include any individual names, the subsequent dataset highlights the complexity of identifying individuals by their family names when these names are associated with specific job titles, as previously discussed. Examples include:

- Carsten Schneider spricht jetzt für die SPD-Fraktion. (Carsten Schneider now speaks on behalf of the SPD parliamentary group.)
- Kollege Schneider, ich glaube, die Bürger haben richtig entschieden
- Da wird Herr Weber nicht begeistert sein! (Mr. Schneider, I believe that the citizens have made the right decision!)
- Herr Koch war sowieso dagegen. (Mr. Koch was against it anyway.)
- Stefan Müller [Erlangen] [CDU/CSU]
- Zimmermann, Sabine DIE LINKE

3.2. Methods

The workflow delineates two sequential steps: initial identification of job titles, followed by subsequent determination of whether said titles refer to an individual or pertain to a previously mentioned occupation. To identify all relevant candidates, a PhraseMatcher is employed, leveraging the data set comprising job titles and synonyms. The present study utilizes Python 3.11.2 and the Spacy library. For a visual representation of the complete workflow, refer to Figure 2.

The initial step in this process is to either substantiate or refute the candidates' claims, as certain terms may not be applicable to a specific position. The approach employed involves the utilization of a filter list, which comprises a manually curated list, or the execution of queries against diverse LLMs to ascertain the nature of the term in question, specifically determining whether it pertains to an occupation or not. In the evaluation, the initial approach is designated as "Rule based."

- If the previous word is a well-known given name (list retrieved from Wikidata), return true.
- If the previous word is a particular title (Kollege, Kollegin, Herr, Frau, or Dr.) or a party name (SPD, CDU/CSU, FDP, ...) return true.
- If the previous or next word is a title or begins with a bracket, return true.

The subsequent task is to ascertain whether the term in question is a family name. The initial approach is a rudimentary rule-based strategy that utilizes a concise sequence of other words:

The second approach once again delegates this question to different LLMS. In the evaluation, the initial approach is again designated as "Rule based."

Table 1
Results on GDR job description data

	Approach	Metrics (Macro/Weighted) Precision	Recall	F_1-score
Occupation Title	Rule based	**0.85/0.86**	**0.83/0.86**	**0.84/0.86**
	llama3.2	0.78/0.81	0.79/0.80	0.78/0.80
	qwen3	0.70/0.74	0.72/0.73	0.71/0.73
	phi4	0.70/0.73	0.71/0.73	0.70/0.73
	gemma3	0.70/0.74	0.72/0.73	0.71/0.73
Family Name	Rule based	**1.00/1.00**	**1.00/1.00**	**1.00/1.00**
	llama3.2	0.50/1.00	0.47/0.95	0.49/0.97
	qwen3	**1.00/1.00**	**1.00/1.00**	**1.00/1.00**
	phi4	0.50/1.00	0.47/0.95	0.49/0.97
	gemma3	0.50/1.00	0.28/0.56	0.36/0.71

4. Experimental Results

4.1. GDR job descriptions

As illustrated in Table 1, the following metrics are presented on GDR job descriptions. The rule-based approach is the most effective method for identifying job title candidates. However, it is subject to certain limitations when it comes to terms such as “Aufbereitung”.

It has been observed that certain models appear to encounter difficulties with extended terms, such as “Ingenieur für Verfahrenstechnik”. This particular term was identified by both the rule-based approach and llama3.2, yet it did not elicit any results from other models. Additional challenges were identified, including the term “Ausbildung” (training), which was identified by the smaller models. Nevertheless, the rule-based approach was the only method that correctly identified more exotic terms such as “Fähnrich” and even “Berufskraftfahrer”.

It is not surprising that the rule-based approach is the most effective method for name detection. As this text category does not include names, the application of all rules is not guaranteed. Nevertheless, it is noteworthy that qwen3 also achieved this optimal performance.

4.2. German Bundestag

As illustrated in Table 2, the result metrics for an annotated subset of 20 plenary discussions of the German Bundestag are presented. Once more, the rule-based approach to identifying job titles evidently surpasses the capabilities of LLM approaches. In particular, the qwen3 model has been observed to detect a considerable number of generic terms, such as “Beratung” (consulting), as occupational titles. However, both the phi4 and qwen3 models have demonstrated challenges in identifying terms such as “Wohnungsbau” and “Recht”. However, the performance of phi4 appears to be satisfactory, as evidenced by its satisfactory performance in both macro and weighted F_1-score metrics. In contrast, the performance of all other models is clearly biased towards macro weights.

With regard to the identification of family names, the precision and recall metrics demonstrate efficacy across all models for weighted scores. However, the efficacy of both phi4 and the rule-based approach has been demonstrated. The annotated data set exhibits significant imbalance, with nearly 94% of all candidates comprising non-job titles and only 0.2% consisting of family names. Consequently, the macro values are not readily discernible. However, preliminary findings suggest the potential efficacy of LLM approaches over the proposed rule-based approach.

Table 2
Results on German Bundestag data

	Approach	Metrics (Macro/Weighted) Precision	Recall	F_1-score
Title	Rule based	**0.97**/**1.00**	**1.00**/**1.00**	**0.98**/**1.00**
	llama3.2	0.55/0.89	0.58/0.86	0.56/0.88
	qwen3	0.60/**0.95**	**0.85**/0.72	0.57/0.79
	phi4	**0.75**/0.94	0.78/**0.94**	**0.77**/**0.94**
	gemma3	0.47/0.87	0.60/0.93	0.48/0.90
Name	Rule based	0.50/**1.00**	0.50/**1.00**	0.50/**1.00**
	llama3.2	0.50/**1.00**	0.49/0.98	0.49/0.99
	qwen3	**0.53**/**1.00**	**0.98**/0.95	**0.54**/0.97
	phi4	0.50/**1.00**	0.50/**1.00**	0.50/**1.00**
	gemma3	0.52/**1.00**	0.97/0.95	0.53/0.97

5. Conclusions and Outlook

This study investigated the complex task of detecting and classifying occupational titles in German texts, highlighting the linguistic and contextual challenges that render this problem particularly challenging. The research utilizes a comprehensive dataset comprising 526,535 job title variants and synonyms, and it analyzes two distinct textual domains: GDR vocational descriptions, and plenary proceedings of the German Bundestag. This approach offers a comprehensive perspective on the nuances involved in occupational detection.

The experimental results underscore the continued effectiveness of rule-based approaches, particularly in structured texts such as GDR job descriptions. The efficacy of these methods was demonstrated by their high precision and recall in identifying occupational titles and family names, with several cases exhibiting superior performance in comparison to large language models (LLMs). However, LLMs such as llama3.2 demonstrated proficiency in detecting more complex, multi-word occupational expressions that were occasionally overlooked by rule-based systems. Despite these advancements, the process of mapping occupations to fine-grained classification codes, such as the 5-digit level of the German Klassifikation der Berufe (KldB), persists as a challenging endeavor. The task is further complicated by the presence of ambiguities caused by overlapping uses of names and job titles, as well as the presence of gender-specific, gender-neutral, and historically derived terms.

In light of these findings, this work suggests several promising areas for future research. First, the usage of other AI-approaches like BERT, see [1], could improve the quality. Second, the integration of rule-based methodologies with LLMs within hybrid systems has the potential to achieve a harmonious equilibrium between precision and adaptability, see for example [18, 19]. The augmentation of training data to encompass a more diverse array of contemporary sources, including user-generated content from online platforms and digitized historical documents, has the potential to enhance the robustness and applicability of models.

Furthermore, the implementation of cross-lingual mapping and alignment with international classification systems such as ISCO and ESCO would expand the practical relevance of these methods beyond national boundaries. As researchers increasingly experiment with LLMs for occupational coding, the development of standardized benchmarks and evaluation frameworks will be essential to ensure comparability and reproducibility of results.

In sum, the present study furnishes a substantial dataset and a series of methodological insights that establish the foundation for more precise and extensible systems in occupational text analysis. The findings are of particular pertinence for applications in labor market research, policy analysis, and computational social science, where the automated identification of occupational information persists as a salient concern.

Declaration on Generative AI

During the preparation of this work, the authors used DeepL in order to: Grammar and spelling check. After using these tool(s)/service(s), the authors reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] J. Binnewitt, Recognising occupational titles in german parliamentary debates, in: Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024), 2024, pp. 221–230.

[2] K. Hein, J. Dörpinghaus, What is said about vet on social media in germany? trends, demands, and opinions., in: NORDYRK BOOK OF ABSTRACTS, 2024, p. 109.

[3] R. Dorau, K. Hein, Towards the automated classification of german job titles according to kldb, in: 205th Conference on Computer Science and Information Systems (FedCSIS), 2025.

[4] A. Fischer, J. Dörpinghaus, Web mining of online resources for german labor market research and education: Finding the ground truth?, Knowledge 4 (2024) 51–67.

[5] N. R. Council, D. of Behavioral, S. Sciences, C. on Occupational Classification, Analysis, Work, jobs, and occupations: A critical review of the dictionary of occupational titles (1980).

[6] F. Javed, M. McNair, F. Jacob, M. Zhao, Towards a job title classification system, arXiv preprint arXiv:1606.00917 (2016).

[7] Y. Zhu, F. Javed, O. Ozturk, Document embedding strategies for job title classification., in: FLAIRS, 2017, pp. 221–226.

[8] I. Rahhal, K. M. Carley, I. Kassou, M. Ghogho, Two stage job title identification system for online job advertisements, IEEE Access 11 (2023) 19073–19092.

[9] F. Javed, Q. Luo, M. McNair, F. Jacob, M. Zhao, T. S. Kang, Carotene: A job title classification system for the online recruitment domain, in: 2015 IEEE First International Conference on Big Data Computing Service and Applications, IEEE, 2015, pp. 286–293.

[10] R. Baskaran, J. Müller, Classification of german job titles in online job postings using the kldb-2010 taxonomy (2023).

[11] M. Schierholz, An auxiliary classification with work activity descriptions for occupation coding, AStA Wirtschafts-und Sozialstatistisches Archiv 12 (2018) 285–298.

[12] A. Müller, The implementation of the German Classification of Occupations 2010 in the IAB Job Vacancy Survey: documentation of the implementation process, Technical Report, IAB-Forschungsbericht, 2014.

[13] V. P. V. Karanam, Occupation coding using a pretrained language model by integrating domain knowledge (2022).

[14] P. Safikhani, H. Avetisyan, D. Föste-Eggers, D. Broneske, Automated occupation coding with hierarchical features: A data-centric approach to classification with pre-trained language models, Discover Artificial Intelligence 3 (2023) 6.

[15] T. Reiser, J. Dörpinghaus, P. Steiner, M. Tiemann, Towards a datatset of digitalized historical german vet and cvet regulations, Data 9 (2024).

[16] T. Reiser, J. Dörpinghaus, P. Steiner, Analyzing historical legal textcorpora: German vet and cvet regulations, in: INFORMATIK 2024, Gesellschaft für Informatik eV, 2024, pp. 2007–2018.

[17] S. Fobbe, Corpus der plenarprotokolle des deutschen bundestages (cpp-bt), 2025. URL: https://doi.org/10.5281/zenodo.15462956. doi:10.5281/zenodo.15462956.

[18] S. Laqrichi, A hybrid framework for cosmic measurement: Combining large language models with a rule-based system, IWSM-Mensura (2024).

[19] M. Billi, A. Parenti, G. Pisano, M. Sanchi, A hybrid approach for accessible rule-based reasoning through large language models, in: 18th International Workshop on Juris-Informatics, 2024.

A. Online Resources

The sources for the ceur-art style are available via

- GitHub,
- Overleaf template.