

Automated classification of German job titles according to KldB: Challenges and novel methods

Ralf Dorau¹, Kristine Hein ¹, Jens Dörpinghaus ^{1,2,3}, and Michael Tiemann ^{1,2}

Abstract: The automated classification of job titles constitutes a critical component of labor market research, survey analysis, and administrative data processing. The present study explores the classification of German job titles according to the German Classification of Occupations (KldB), with a particular emphasis on the linguistic and structural challenges that are inherent to this task. This study builds upon previous research by incorporating a variety of heterogeneous data sources, including manually annotated survey responses, a comprehensive synonym dataset, online job advertisements (OJAs), and vocational education and training titles from DAZUBI. Conventional machine learning models, including logistic regression, naive Bayes, and random forest, are employed to assess the classification performance at varying taxonomic levels of the KldB. The findings of the present study demonstrate that while substantial results can be achieved for broad occupational categories, fine-grained classification, particularly at the level of performance (5th digit), remains challenging. The findings underscore the limitations of relying solely on job titles and underscore the importance of richer contextual information and more expressive models. This work provides both an expanded dataset and a systematic analysis of classification performance, thereby establishing the foundation for future research on context-aware occupational coding in the German labor market.

Keywords: Computational Social Sciences, Job titles, Classification, Labor Market Research

1 Introduction

The automated classification of job titles is a significant topic in both academic research and for practitioners in labor market analysis. For instance, survey respondents may provide open-text answers regarding their occupation, which then need to be mapped to standardized classifications such as the German Classification of Occupations (KldB) or 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 [HD24]. Subsequent research endeavors seek to ascertain the presence of occupations within textual data, such as that found in parliamentary debates [Bi24].

¹ Federal Institute for Vocational Education and Training (BIBB), Bonn, Germany, dorau@bibb.de; kristine.hein@bibb.de,  <https://orcid.org/0009-0005-6414-4294>; doerpinghaus@uni-koblenz.de,  <https://orcid.org/0000-0003-0245-7752>; mtiemann@uni-koblenz.de,  <https://orcid.org/0000-0001-7136-2744>

² University of Koblenz, Germany, doerpinghaus@uni-koblenz.de,  <https://orcid.org/0000-0003-0245-7752>; mtiemann@uni-koblenz.de,  <https://orcid.org/0000-0001-7136-2744>

³ Linnaeus University, Department of Computer Science and Media Technology, Växjö, Sweden, doerpinghaus@uni-koblenz.de,  <https://orcid.org/0000-0003-0245-7752>

The German language presents distinctive 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’). Additionally, 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 example is the surname ‘Bäcker,’ which is derived from the German word for ‘baker.’

The integration of diverse labor market data sources is widely acknowledged as a multifaceted undertaking [FD24b]. 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. In the present study, the approach previously outlined in [DH25] is expanded upon through the incorporation of job titles derived from online job advertisements and vocational education and training (VET) titles. The aforementioned sources present a series of challenges that will be examined in detail in the following sections.

The primary contribution of this paper is a substantial collection of novel training data and a systematic analysis of the challenges associated with training classifiers for German job titles according to KldB. This analysis focuses particularly on the different levels of occupations and their performance levels.

The present document is organized into five sections. The initial section of the text offers an introduction to the subject matter. The subsequent section provides a comprehensive overview of the current state of the field and relevant works that have been conducted in related areas. The third section delineates the data and methodological background. The fourth section is dedicated to the presentation of experimental results and their subsequent evaluation. The conclusions of this study are presented in the final section.

2 Related Work

We follow our study of automated classifications presented in [DH25]. 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 2008⁴. 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

⁴ See <https://www.ilo.org/public/english/bureau/stat/isco/isco08/>.

of Occupations (KldB) serves as the reference classification for the Federal Employment Agency (BA) and its research institute (IAB)⁵. 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 [Co80].

A portion of the research has focused on the classification of OJAs according to the O*NET framework [Ja16]. This has included the application of normalization approaches [ZJO17] and similarity-based methods [Ra23]. The classification of job titles is also employed in the context of online job recruitment [Ja15].

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 [BM23] with challenges on level 4, but with promising results on level 1. Malte Schierholz's 2018 publication [Sc18] 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 [Mü14]. 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 [Ka22]. In a similar vein, a scholarly article was published that compared the classification of survey data using BERT and GPT-3, see [Sa23]. 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 [DH25] 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.

3 Data and Methods

3.1 Survey data

The dataset aggregates 60,022 manually annotated occupational titles from the 2012, 2018, and forthcoming 2024 waves of the Erwerbstätigenbefragung (ETB; German Employment Survey)⁶, each coded to KldB 2010. These representative surveys – conducted by BIBB and BAuA – target Germany's core workforce (employed individuals aged 15+ working ≥10

⁵ See <https://statistik.arbeitsagentur.de/DE/Navigation/Grundlagen/Klassifikationen/Klassifikation-der-Berufe/KldB2010-Fassung2020/KldB2010-Fassung2020-Nav.html>.

⁶ Main datasets: 10.7803/501.12.1.1.60 (2012), 10.7803/501.18.1.1.10 (2018). Supplementary variables: 2012: 10.7803/501.12.1.4.10 (full-text), 10.7803/501.12.1.3.20 (regional identifiers), 10.7803/501.12.1.5.30 (special variables); 2018: 10.7803/501.18.1.4.10 (full-text), 10.7803/501.18.1.3.10 (regional identifiers), 10.7803/501.18.1.5.10 (special variables).

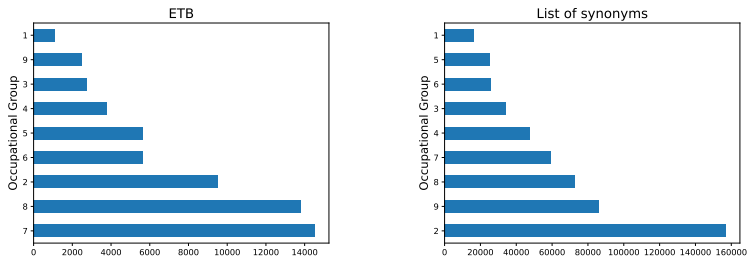


Fig. 1: Distribution of training data according to occupational groups

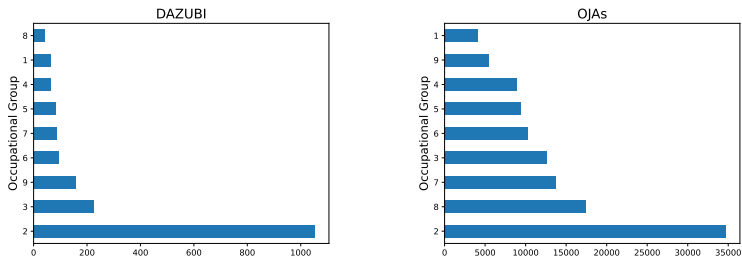


Fig. 2: Distribution of training data according to occupational groups

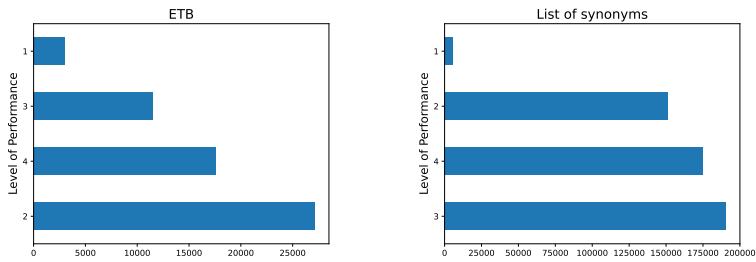


Fig. 3: Distribution of training data according to level of performance (5th-digit)

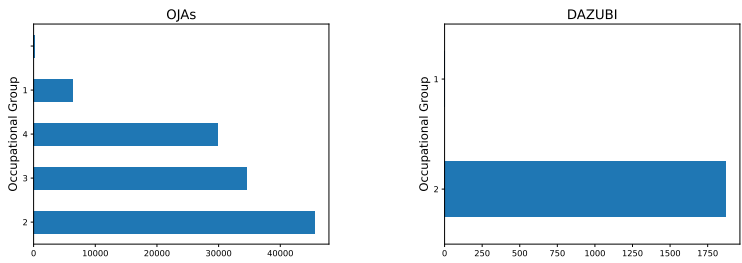


Fig. 4: Distribution of training data according to level of performance (5th-digit)

hours/week). While historical predecessor surveys (1979-1999 BIBB/IAB studies⁷) exist, their integration was precluded by fundamental incompatibilities between KldB versions, despite partial crosswalk feasibility.

3.2 Job titles and synonyms

A secondary 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)⁸.

3.3 OJAs

For the present dataset, a total of 116,674 titles of job advertisements from the Jobsuche portal were collected⁹. These titles range from generic titles like ‘technischer Mitarbeiter (m/w/d)’ with additional requirements, e.g., ‘Logistikmitarbeiter mit Staplerschein (m/w/d),’ to specific titles with different gender forms and additional phrases, for example:

- Zytologieassistent/in oder MTLA (m/w/d)
- MTLA - Zytologieassistent/in
- Zytologie- Assistent/in Bereich gynäkologischer Forschung
- MTLA / Zytologieassistent/in (m/w/d) gesucht!

Consequently, this dataset presents its own challenges, yet it also offers a more extensive array of job titles and equivalent titles. However, these labels may lack specificity, contain extraneous text, or even include duplicate job titles.

3.4 DAZUBI

This dataset is supplementary information derived from the ‘Datensystem Auszubildende’ (DAZUBI) database, which is provided by the Federal Institute for Vocational Education and Training (BIBB). The scope of the data encompasses vocational training statistics from Federal and State Statistical Offices, along with annual survey data concerning vocational training for trainees, contracts, and examinations. However, the dataset utilized exclusively contains current and deprecated labels, which are employed for training, retraining, and other vocational education purposes. This modest dataset encompasses 1,875 labels, which frequently contain supplementary information, such as the testing organization.

⁷ GESIS SUF: 10.4232/1.1243, 10.4232/1.12563, 10.4232/1.2565, 10.4232/1.12247

⁸ Available at <https://www.arbeitsagentur.de/institutionen/dkz-downloadportal>.

⁹ See <https://www.arbeitsagentur.de/jobsuche/>.

Consequently, the terms ‘Änderungsschneider/-in (Hw)’ and ‘Änderungsschneider/-in (IH)’ are encountered. Additionally, other information, for example in ‘Anlagenmechaniker/-in (Monoberuf) (HwEx)’ and ‘Werkzeugmaschinenwerker/-in (§ 66 BBiG) (IH)’ are also presented. The fifth digit is typically 2, as the dataset encompasses vocational education. However, four data points containing the digit 1 are identified:

- Blinde/-r/Sehbehinderte/-r Telekommunikationsoperator/-in (§ 66 BBiG) (IH)
- Bürohelfer/-in (§ 42r HwO) (Hw)
- Bürohelfer/-in (§ 66 BBiG) (IH)
- Telefonist/-in (§ 66 BBiG) (IH)

3.5 Data biases and cleaning

As illustrated in Figures 1 and 2, it is evident that all datasets are subject to inherent biases. A preliminary investigation of the ETB dataset reveals that the majority of occupations are located in groups 7 and 8. However, when the list of synonyms is considered, it is evident that these occupations are primarily situated in groups 2 and 9. The list of OJAs demonstrates a discrepancy in the representation of groups, as evidenced by the underrepresentation of groups 9 and 5. DAZUBI data is subject to an inherent bias due to the unequal distribution of available VET programs across the specified groups.

A similar bias is identified with respect to the level of performance, which is encoded in the fifth digit of KldB, see Figures 3 and 4. Once more, it is important to note that DAZUBI data is distinctive in that it exclusively contains data for digit 2, while also comprising four records for digit 1.

Data cleaning is an existentiell step. We remove occupational are 0 from the dataset as we omit military occupations. Second, we clean all data with errors, for example too large or small KldB numbers. In Dataset 1, 59,535 and in dataset 2, 522,197 entries remain.

3.6 Method

In this paper, we adopt the classification approach to labor market data as presented in [FD24a; FDF23]. The baseline model was trained using three standard classifiers with default values from the scikit-learn library [Pe11] with logistic regression ($C = 10$), naive Bayes, and random forest. The data was divided into two distinct sets: 20% was designated for the test set, while the remaining 80% was allocated for the training set. In our previous work, we identified a methodology that was found to be inferior to traditional approaches, see [DH25]. In this study, a prediction model was developed using 10-fold cross-validation for three distinct datasets, employing the aforementioned three machine learning (ML)

		time	F_1 (w)	F_1 (m)	Precision	Recall
all	Logistic Regression	248,06	0,74	0,73	0,75	0,74
	Naive Bayes	1,28	0,71	0,69	0,74	0,72
	Random Forrest	10108,56	0,67	0,65	0,70	0,66
\OJAs	Logistic Regression	222,63	0,74	0,73	0,76	0,74
	Naive Bayes	1,05	0,71	0,68	0,74	0,72
	Random Forrest	6435,05	0,67	0,64	0,70	0,66
\OJAs&DAZUBI	Logistic Regression	67,25	0,74	0,73	0,76	0,74
	Naive Bayes	1,02	0,71	0,68	0,73	0,72
	Random Forrest	6354,38	0,67	0,65	0,71	0,66

Tab. 1: Weighted results for 1-digit, weighted (w) and macro (m)

techniques. In our methodology, we employ all available datasets, while omitting the OJAs as a highly specific dataset. Alternatively, we utilize a combination of the OJAs and DAZUBI.

However, another open question is the choice of text representation. As shown by this can significantly change the performance of classifiers. We can use the combination of term frequency and inverse document frequency as TF.IDF (term frequency–inverse document frequency). Another approach is the Word to Vector (W2V) which applies unsupervised learning to represent textual data and produces rather low-dimensional data. As we have seen in paper [DH25], the later produces rather poor results.

4 Experimental Results

4.1 Occupational areas (1-digit)

Table 1 presents the outcomes for occupational categories with a one-digit classification. These outcomes are contrasted with the utilization of all datasets with OJAs and DAZUBI collectively. The findings of this study are largely consistent with those documented in the extant literature and in our previous work, see [DH25]. The results reported in this study are consistent with the results reported in the aforementioned studies. A comparative analysis reveals that, while all other models demonstrate comparable performance, Logistic Regression consistently exhibits superior performance in various applications.

In contrast to the results of [DH25], the findings of this study demonstrate that for occupational extension, the dataset has no substantial impact on the quality of the results for detecting occupational areas. For instance, the F_1 score demonstrates minimal change, even when utilizing OJAs or DAZUBI.

		time	F_1 (w)	F_1 (m)	Precision	Recall
all	Logistic Regression	2163,74	0,72	0,72	0,75	0,72
	Naive Bayes	5,58	0,60	0,53	0,72	0,60
	Random Forrest	15198,02	0,69	0,70	0,72	0,69
\OJAs	Logistic Regression	837,82	0,75	0,75	0,78	0,75
	Naive Bayes	4,74	0,63	0,54	0,73	0,62
	Random Forrest	9510,90	0,73	0,73	0,76	0,72
\OJAs&DAZUBI	Logistic Regression	594,44	0,76	0,76	0,78	0,75
	Naive Bayes	3,94	0,63	0,55	0,74	0,62
	Random Forrest	9453,18	0,73	0,74	0,77	0,72

Tab. 2: Weighted results for 3-digit, weighted (w) and macro (m)

4.2 Occupational group (3-digit)

As demonstrated in Table 2, the findings concerning the occupational group (3-digit) are displayed. A direct comparison with occupational areas reveals that the performance is comparable. However, the employment of disparate datasets has been demonstrated to result in a substantial enhancement of the outcomes. In particular, the employment of OJAs, as well as DAZUBI, has been demonstrated to yield suboptimal outcomes. Therefore, the allocation of job titles to occupational groups is an ineffective process for OJAs. As illustrated in Figure 5, we present the comprehensive results for the optimal outcomes on the dataset excluding OJAs and DAZUBI. The findings presented in [DH25] are consistent with the established range. However, a general tendency towards overestimation is observed.

However, a more thorough examination reveals that this approach is particularly ineffective for groups with limited data. A preliminary analysis of dataset 1 indicates the presence of significant missing data points, specifically those related to areas 114 (fishing occupations), 213 (glass-making and -processing occupations), and 214 (ceramic-making and -processing occupations). As previously mentioned in Section 3.5, a strong bias for DAZUBI data has been identified. For occupational groups, the OJA data is also subject to significant bias. For instance, there are 20 OJAs for 110, 26 for 712 (i.e., legislators and senior officials of special interest organizations), but 4,259 for 263 (i.e., occupations in electrical engineering) or 3,244 for 251 (occupations in machine-building and -operating). For future research endeavors, there is a necessity for not only improved training data that is more balanced, but also for the evaluation of cross-validation approaches.

4.3 Level of Performance (5th digit)

The results of the classification of the level of performance (5th digit) demonstrate a number of noteworthy outcomes, as illustrated in Table 3. It is noteworthy that the extended dataset

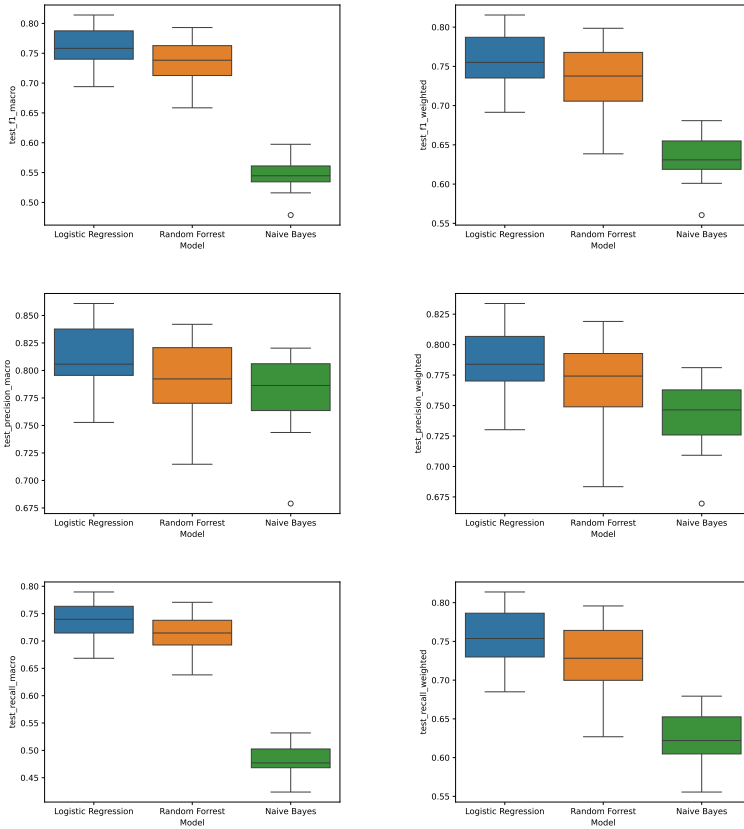


Fig. 5: Results for digit 3 and all data without OJAs and DAZUBI

does not yield substantial improvements in the results; rather, it yields contrary results. OJAs and DAZUBI seem to have a negative, but small influence on the results, as evidenced by the mean F_1 score. Secondly, we observe that different levels yield different F_1 scores, as also mentioned in [DH25]. Logistic regression demonstrates superior performance for both level 2 and level 4, while levels 1 and 3 exhibit substandard performance. A similar outcome is observed for all other methods, although level 1 and 3 demonstrate an even poorer performance. Therefore, it can be posited that the fifth digit plays a significant role in the classification of job titles. As was previously observed, the addition of more data results in improved performance, although a careful selection seems to be necessary. While OJAs appear to disrupt the performance metrics for occupational groups classified under three-digit categories, they do not seem to encode data corresponding to the level of performance represented by the fifth digit.

		time	F_1 (w)	F_1 (m)	Precision	Recall
all	Logistic Regression	167,73	0,65	0,62	0,66	0,66
	Naive Bayes	1,21	0,60	0,42	0,61	0,61
	Random Forrest	12832,42	0,63	0,60	0,64	0,63
\OJAs	Logistic Regression	134,16	0,62	0,58	0,62	0,62
	Naive Bayes	0,99	0,56	0,39	0,57	0,57
	Random Forrest	8425,47	0,60	0,63	0,62	0,60
\OJAs&DAZUBI	Logistic Regression	34,89	0,65	0,63	0,66	0,66
	Naive Bayes	1,19	0,60	0,43	0,61	0,61
	Random Forrest	12646,54	0,63	0,60	0,65	0,63

Tab. 3: Weighted results for Level of Performance (5th digit), weighted (w) and macro (m)

Therefore, the results support the results in our previous work [DH25] that the available data from a job title alone is insufficient for effective classification. However, it should be noted that not all data can be used for all data encoded in KldB. The efficacy of job titles in OJAs in enhancing performance detection is a notable finding. However, the collective performance is not particularly noteworthy. The various methods that utilize the different data sets are presented in Figures 6 and 7. It is evident that a reduced volume of data yields slightly superior outcomes. However, it should be noted that the outcomes exhibit significant variability. Therefore, while the selection of data appears to be a challenging task, it is generally the case that the collection of more data results in more stable results.

5 Discussion and outlook

This study builds upon our prior research [DH25] by expanding the classification task to include additional data sources, such as online job advertisements (OJAs) and vocational education and training titles from DAZUBI. These modifications have the effect of significantly increasing the linguistic diversity and contextual variability of job titles, thereby providing a broader foundation for examining automated classification according to the KldB taxonomy.

The findings indicate that conventional machine learning approaches, particularly logistic regression combined with TF-IDF text representations, demonstrate robustness for higher-level occupational categories, such as the 1-digit and 3-digit classifications. However, their performance exhibits a marked degradation at finer levels, particularly at the fifth-digit level of performance. This decline underscores the fundamental limitations of relying exclusively on job title strings, particularly when they are devoid of context or influenced by domain-specific formulations, gendered variations, and certification notations.

The integration of the synonym dataset, in conjunction with the incorporation of OJA and DAZUBI entries, yielded performance enhancements across various categories. However,

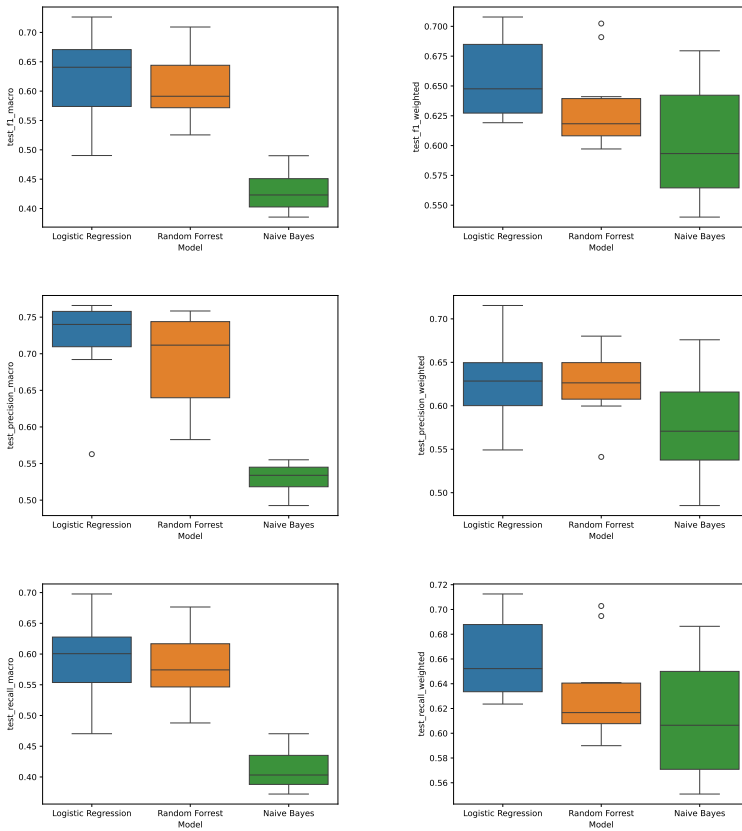


Fig. 6: Results for digit 5 and all data

these enhancements were accompanied by the introduction of biases due to imbalanced class distributions. These biases were particularly pronounced in underrepresented occupational groups and levels of performance. Consequently, while the overall classification accuracy remained stable, the reliability of predictions for less frequent classes remained limited. This outcome underscores the necessity of more balanced datasets and more granular training data.

A significant finding of this study is that job titles, when evaluated in isolation, are insufficient for fine-grained classification tasks. In order to enhance performance at deeper classification levels, it is essential to incorporate supplementary textual or contextual information, such as job descriptions, formal qualification requirements, or metadata about the occupational environment. The previously mentioned information can facilitate the disambiguation of similar or ambiguous job titles, e.g., from OJAs.

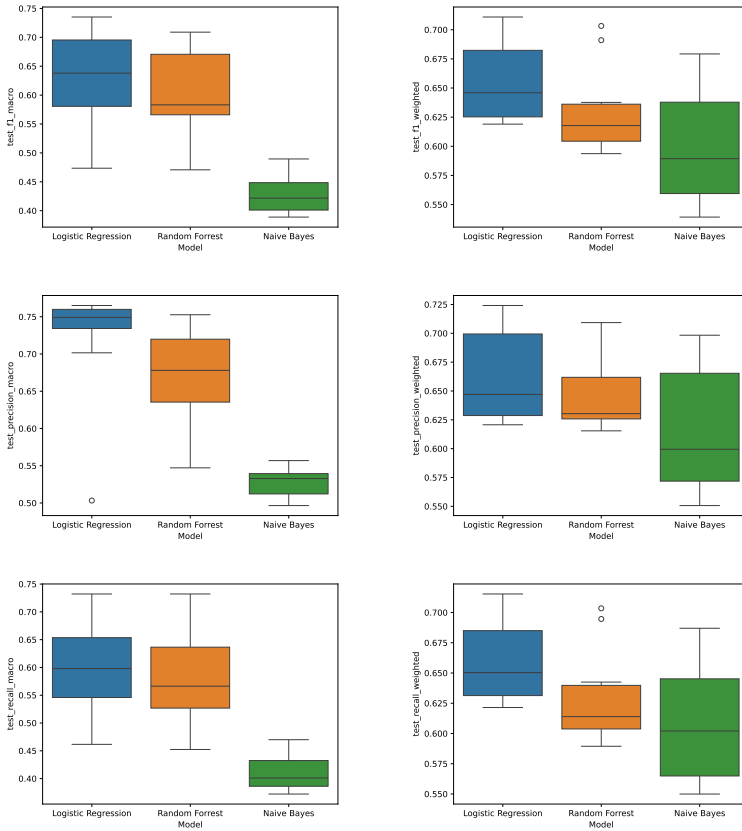


Fig. 7: Results for digit 5 and all data without OJAs and DAZUBI

References

- [Bi24] Binnewitt, J.: 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). Pp. 221–230, 2024.
- [BM23] Baskaran, R.; Müller, J.: Classification of German job titles in online job postings using the KldB-2010 taxonomy. 2023.
- [Co80] Council, N. R. et al.: Work, jobs, and occupations: A critical review of the Dictionary of Occupational Titles. 1980.
- [DH25] Dorau, R.; Hein, K.: Towards the automated classification of German job titles according to KldB. In: 205th Conference on Computer Science and Information Systems (FedCSIS). 2025.

-
- [FD24a] Fechner, R.; Dörpinghaus, J.: No Train, No Pain? Assessing the Ability of LLMs for Text Classification with no Finetuning. In: Proceedings of the Position Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), Belgrade, Serbia. Pp. 8–11, 2024.
- [FD24b] Fischer, A.; Dörpinghaus, J.: Web Mining of Online Resources for German Labor Market Research and Education: Finding the Ground Truth? *Knowledge* 4 (1), pp. 51–67, 2024.
- [FDF23] Fechner, R.; Dörpinghaus, J.; Firll, A.: Classifying industrial sectors from German textual data with a domain adapted transformer. In: 2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS). IEEE, pp. 463–470, 2023.
- [HD24] Hein, K.; Dörpinghaus, J.: What is said about VET on social media in Germany? Trends, demands, and opinions. In: NORDYRK BOOK OF ABSTRACTS. P. 109, 2024.
- [Ja15] Javed, F. et al.: 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, pp. 286–293, 2015.
- [Ja16] Javed, F. et al.: Towards a job title classification system. arXiv preprint arXiv:1606.00917, 2016.
- [Ka22] Karanam, V. P. V.: Occupation coding using a pretrained language model by integrating domain knowledge. 2022.
- [Mü14] Müller, A.: The implementation of the German Classification of Occupations 2010 in the IAB Job Vacancy Survey: documentation of the implementation process, tech. rep., IAB-Forschungsbericht, 2014.
- [Pe11] Pedregosa, F. et al.: Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12, pp. 2825–2830, 2011.
- [Ra23] Rahhal, I. et al.: Two stage job title identification system for online job advertisements. *IEEE Access* 11, pp. 19073–19092, 2023.
- [Sa23] Safikhani, P. et al.: Automated occupation coding with hierarchical features: A data-centric approach to classification with pre-trained language models. *Discover Artificial Intelligence* 3 (1), p. 6, 2023.
- [Sc18] Schierholz, M.: An auxiliary classification with work activity descriptions for occupation coding. *AStA Wirtschafts-und Sozialstatistisches Archiv* 12, pp. 285–298, 2018.
- [ZJO17] Zhu, Y.; Javed, F.; Ozturk, O.: Document Embedding Strategies for Job Title Classification. In: FLAIRS. Pp. 221–226, 2017.