

Extraction of contract types from German OJA

Bertelsmann Stiftung

Version

1

Autor*innen

Rahkakavee Baskaran

Erstellt am

29. Juni 2026

&effect

Table of Contents

1	Executive Summary	3
2	Contract Type Classification in Online Job Postings	3
3	Data & Annotation	3
3.1	Data Source	3
3.2	Sampling	4
3.3	Annotations	4
3.4	Relabeling	4
3.5	Concept Distribution.....	4
4	Model	5
4.1	Intended Use	5
4.2	Architecture	5
4.3	Post-processing	6
4.4	Training.....	6
5	Evaluation	7
6	Limitations & Recommendations	8
6.1	Data	8
6.2	Model.....	9

1 Executive Summary

This report documents the dataset, model, and evaluation for the contract-type classifier, a single-label text classification model that assigns German online job postings to one of eight employment categories: *Arbeit*, *Berufsausbildung*, *Duales Studium*, *Praktikum*, *Werks-Studierende*, *HiWi*, *Selbstständigkeit*, and *FSJ/BUFDI*. The classifier enables contract-type-specific labor market analyses and serves as a filter for downstream models, where non-standard contract types such as internships or working-student roles are excluded prior to further processing. The training dataset consists of 2,360 manually annotated German online job postings. The model is built on spaCy's TextCatEnsemble.v2 architecture and achieves an overall accuracy of 0.89 and a macro F1 of 0.87 on 521 held-out postings, with strong performance across most categories.

2 Contract Type Classification in Online Job Postings

Analyzing online job postings at scale requires enriching raw posting texts with structured attributes. Contract type is one such attribute: it distinguishes standard job openings from internships, vocational training, working-student roles, and other non-standard employment forms that require separate treatment in downstream analysis.

The contract-type classifier assigns German online job postings to one of eight mutually exclusive categories: *Arbeit* (standard employment), *Berufsausbildung* (vocational training), *Duales Studium* (dual study programme), *Praktikum* (internship), *Werks-Studierende* (working students), *HiWi* (student research assistant), *Selbstständigkeit* (freelance / self-employment), and *FSJ/BUFDI* (voluntary service).

The predicted labels serve two purposes: they enable contract-type-specific labor market analyses, and they function as a filter for downstream models, where non-standard contract types are excluded prior to further processing.

3 Data & Annotation

3.1 Data Source

The dataset consists of 2,360 manually annotated German online job postings. Job postings were sourced from a large corpus of German online vacancies collected between 2018 and August 2024, with posting texts retrieved via [Textkernel's Jobs Data Feed](#).

3.2 Sampling

For each annotation round, a candidate pool was drawn via random sampling from the full posting corpus. To increase coverage of underrepresented classes, pre-labels from three sources were used to guide post-hoc selection from this pool: LLM annotations (Gemini 2.5 Flash), Textkernel’s existing contract-type field (which covers a subset of the target categories), and regex-based matching rules. These pre-labels served as selection signals only; all final labels were determined by human annotators. In total, four annotation datasets were assembled this way.

3.3 Annotations

Postings were annotated using the software Elinor by a mixed group of labor market domain experts and ML practitioners from Bertelsmann Stiftung and &effect data solutions GmbH. Each posting was assigned exactly one contract-type label from the defined taxonomy. A subset of postings was annotated by multiple annotators per round to assess consistency: 108 postings in the test round, 60 to 100 in subsequent rounds. Disagreements were resolved through structured group discussion and iteratively refined annotation guidelines.

3.4 Relabeling

Two automated, rule-based relabeling rules were applied to the gold-standard labels before producing the final train/test splits.

The first rule addresses systematic confusion between *Arbeit* and *Selbstständigkeit*. A manual analysis showed that postings from non-freelancer platforms are ambiguous in this regard even to human annotators, and that source platform is a more reliable signal for this distinction than posting content alone. A keyword-based alternative was evaluated (screening for terms such as “freelancer” or “freiberuflich”) but performed worse in practice, as these terms appear frequently in job portal boilerplate, regardless of actual employment type. The applied rule relabels all *Selbstständigkeit* postings not originating from a known freelancer portal (source_type_id = 7) to *Arbeit*; postings from freelancer portals retain the *Selbstständigkeit* label.

The second rule addresses *Ehrenamt*. With only 20 training and 5 test examples, the category is too rare for reliable classification. *Ehrenamt* is retained in the training set so the model learns the boundary with *Arbeit*, but relabeled to *Arbeit* in the test set to reflect production behavior, where *Ehrenamt* is not part of the operational taxonomy. All reported evaluation metrics reflect this mapping.

3.5 Concept Distribution

The following table shows the class distribution of the final dataset after relabeling, as used by the deployed model. *Arbeit* is the dominant class at ~45%, reflecting its prevalence in the

underlying corpus. The remaining classes range from ~11% (*Werks-Studierende*) down to ~1% (*Ehrenamt*). The distribution is consistent across train and test splits.

Concept	Train (abs)	Test (abs)	Train (%)	Test (%)
Arbeit	953	237	45.82	45.49
Werks-Studierende	237	60	11.39	11.52
Berufsausbildung	226	56	10.87	10.75
Praktikum	187	47	8.99	9.02
Duales Studium	170	43	8.17	8.25
Selbstständigkeit	120	31	5.77	5.95
HiWi	86	22	4.13	4.22
FSJ/BUFDI	81	20	3.89	3.84
Ehrenamt	20	5	0.96	0.96
Total	2,080	521		

4 Model

4.1 Intended Use

The model is intended for the automated classification of German online job postings by contract type as part of labor market analysis pipelines. It accepts the full text of a job posting as input and returns one of eight contract-type labels. The model is optimized for postings sourced via Textkernel and has not been systematically evaluated on postings from other aggregators or collection methodologies. It is not suitable for CVs, applicant profiles, or other text types.

4.2 Architecture

The classifier is built on spaCy's TextCatEnsemble.v2 architecture, which combines two sub-models. The first is a bag-of-words component (TextCatBOW.v3) that represents each document as a unigram count vector. The second is a contextual tok2vec encoder that embeds each token using normalized form, prefix, suffix, and word shape, concatenated with 300-dimensional static word vectors from the base model `de_core_news_lg`. Token embeddings are

passed through a 4-layer MaxoutWindowEncoder; the outputs of both sub-models are combined before the final classification layer.

4.3 Post-processing

Two deterministic rules are applied to every prediction before the result is returned, encoding domain knowledge that the model cannot reliably learn from data alone. Where available, the job posting title should be passed to the model alongside the full posting text, as it is consumed by the FSJ/BUFDI post-processing rule and improves classification recall for that category.

Rule	Condition	Action
Ehrenamt fallback	pred == "Ehrenamt"	→ "Arbeit"
FSJ/BUFDI guard	pred == "FSJ/BUFDI" and neither text nor title contains ["fsj", "bufdi", "freiwilliges soziales", "bundesfreiwillig"]	→ "Arbeit"

4.4 Training

The model was trained using spaCy's built-in training loop with early stopping on macro F1. The best checkpoint across 20,000 maximum steps was retained as the final model.

Parameter	Value
Optimizer	Adam
Learning rate	0.001
Dropout	0.15
Max steps	20,000
Early stopping patience	1,600 steps
Batch size	compounding 300 to 1,000
Model selection	macro F1 on dev set

5 Evaluation

5.1.1 Results

The model was evaluated on 521 held-out postings (80/20 split). All metrics are computed after post-processing rules have been applied, reflecting the performance of the full inference pipeline. The pipeline achieves an overall accuracy of 0.89 and a macro F1 of 0.87. The lowest individual F1 scores are observed for *Praktikum* (0.79) and *Duales Studium* (0.83), both of which show recall deficits due to confusion with adjacent categories.

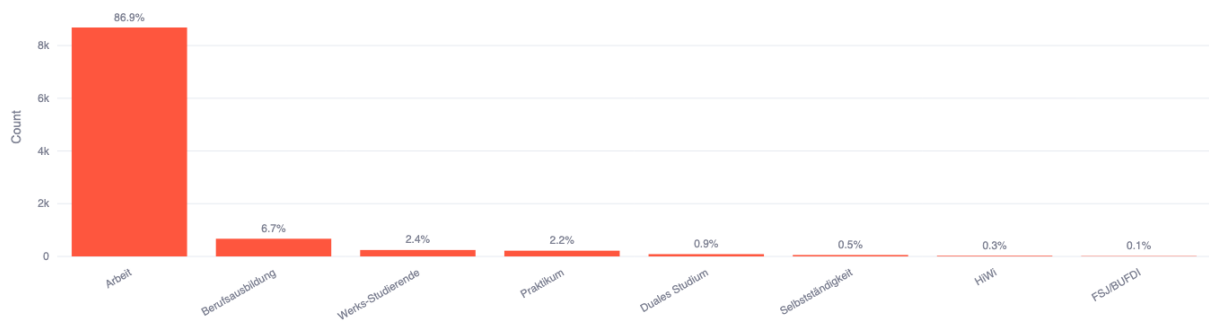
Category	Precision	Recall	F1	Support
Arbeit	0.90	0.96	0.93	242
Berufsausbildung	0.83	0.88	0.85	56
Duales Studium	0.97	0.72	0.83	43
Praktikum	0.87	0.72	0.79	47
Werks-Studierende	0.83	0.88	0.85	60
HiWi	0.82	0.82	0.82	22
Selbstständigkeit	1.00	0.87	0.93	31
FSJ/BUFDI	1.00	0.95	0.97	20
Macro avg	0.90	0.85	0.87	521

The dominant error pattern across all classes is misclassification into *Arbeit*, consistent with its role as the majority class and catch-all for ambiguous cases.

5.1.2 Plausibility Checks

Beyond formal evaluation, three checks were conducted on a random sample of approximately 10,000 production postings.

First, the predicted label distribution was compared against prior expectations about the composition of the German online job market. The distribution is consistent with expectations and shows no signs of systematic over- or under-prediction.



Second, for categories where the contract type is typically stated explicitly in the job title, keyword match rates between 64% and 71% provide a high-level confirmation that the majority of predictions are correct. The remaining predictions, where no keyword match was found in the title, were manually spot-checked and show plausible classifications based on the full posting text.

Label	Predictions	Title matches	Match rate
Berufsausbildung	671	466	69%
Praktikum	217	148	68%
FSJ/BUFDI	14	10	71%
Duales Studium	87	56	64%

Third, approximately 4.8% of postings in the production sample are not in German, of which 91% are classified as *Arbeit*. A manual spot-check of these cases found the predictions to be plausible: the postings are predominantly standard job openings where shared vocabulary, named entities, and document structure provide sufficient signal for a sensible classification. The model is optimized for German text and was not systematically evaluated on other languages; results on non-German postings should be treated as indicative only.

6 Limitations & Recommendations

6.1 Data

The *Selbstständigkeit* class is defined by source platform rather than posting content: only postings from known freelancer portals are labeled *Selbstständigkeit*; all others are defined as *Arbeit*. This is a taxonomic decision that yields a clean decision boundary in practice but introduces known label noise for postings that explicitly mention freelance work while originating from non-freelancer platforms. Several posting types in the grey area between employment and self-employment (promoter jobs, tutoring roles, commission-based sales positions, network marketing, and informal household help postings) are assigned to *Arbeit* by

taxonomic decision, as textual evidence alone is insufficient to distinguish them from standard job postings.

Some contract-type boundaries are inherently ambiguous, particularly between *Arbeit* and *Selbstständigkeit*, and between *Praktikum* and *Berufsausbildung*. Disagreements were resolved through structured group discussion among domain experts.

6.2 Model

The categories *FSJ/BUFDI* (n=20), *HiWi* (n=22), and *Selbstständigkeit* (n=31) have small test sets; reported metrics for these classes should be interpreted with caution. The model was trained and evaluated on German-language postings; performance on non-German postings has not been systematically evaluated and may vary.