

# Talking about tasks or just sharing job offers? A case study on job-related tweets

Kristine Hein <sup>1</sup>, Michael Tiemann <sup>1,2</sup>, and Jens Dörpinghaus <sup>1,2,3</sup>

**Abstract:** The skills and qualifications of IT professionals are constantly changing and under discussion. In particular, we see the impact of emerging technologies and tools, such as AI, on the labor market and their reflection in the broader scientific community, media and everyday life [Do19; He16]. Several approaches have discussed how to assess the impact of computer science education from the perspective of education and labor market research [HBE22; SHL22]. However, in order to uncover the complex dynamics surrounding vocational education, we need to take a closer look at the labor demand as well as the public perception and valuation of professionals. Therefore, we present a novel approach to work with social media data, in particular Twitter/X data. We have collected a large dataset of job-related X data (2007-2023) but see that a word list-based searching strategy is not enough. Most tweets are actually job offers and we will discuss several strategies to categorize tweets accordingly to make a meaning out of it. We focus on occupations in natural sciences, geography and informatics.

**Keywords:** Online Social Networks, Computational Social Sciences, Classifier, LLMs

## 1 Introduction

In today's digital age, social media platforms have become essential tools for understanding public sentiment and discourse, see [HBE22; SHL22]. We explore the broad world of social media focusing on Twitter/X to uncover the complex narrative surrounding Vocational Education and Training (VET) in Germany. Our study examines the public perception of various occupations, going beyond a surface-level examination of vocational education and schools. For this, we have collected 3,407,481 X tweets (2007-2023) linked to occupations according to the German Classification of Occupations (KldB). However, not all tweets give information about occupations. Modern state of the art Text Classification approaches depend heavily on well annotated and diverse data capturing the intricacies of the unknown data distribution. As discussed in [FD24], LLMs may provide an alternative to these classical approaches which we will also discuss in this paper.

---

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

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

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

The German labour market is dynamic: technical innovations like digitalization and changes in society lead to novel skills needed for employees. Vocational education and training, re-training and advanced vocational qualification are key to respond to these novel requirements [Do19; He16]. The central importance of continuing vocational training and lifelong learning for the success of digital transformation has been increasingly addressed by policymakers as part of the National Continuing Education Strategy [Sc22]). Here, continuing education is described as a central prerequisite for securing skilled labour, for ensuring the employability of all employees and thus also for national competitiveness and innovation. With regard to the relevance resulting from the digital transformation of work, it can be stated out that all sectors of the economy and the majority of companies are affected by forms of digitalization [WW17]. It is thus a cross-cutting issue affecting a wide range of topics in vocational education and training.

This paper addresses two primary research questions. The first is to ascertain the nature of data yielded by a word-list searching approach. The second is to determine whether it is possible to classify tweets according to their relationship to occupations. To address the second question, the results of classical machine learning (ML) approaches and large language model (LLM) approaches are compared.

This paper is divided into six sections. The first section provides an introduction and gives a brief overview of the background of the research question, the second section presents related work. Section three presents the data and an overview about existing resources. In section four, we will introduce the methods used to answer the research question. The fifth section is dedicated to experimental results and the evaluation of these methods. In the final section, we discuss our findings and give a detailed interpretation of the results. Our conclusions and outlook onto further research are drawn in the final section.

## 2 Related Work

There are many studies focusing on labor market research in social media and in particular Twitter/X. Most studies focus on the tweet text, only some work with biographies and other meta data [Gu24] or communication networks [Th13]. In general, research questions seem to be inspired by questions from social sciences, including demographical or geographical aspects or identity and social status. Other questions are related to the reaction to events like the COVID-19 pandemic or new tools and topics, see for example [ZXZ21]. Quite a lot of research focused on the prediction of user's occupations [Pa19a; YZD22].

In Table 1 we present an overview of work carried out studying occupations on social media, in particular Twitter/X. Since research relies on data, not all platforms can be used equivalently, we still added example studies on Facebook to show the interest in this kind of social media data for occupational research.

However, very little work has been done in this area for German-language social media data. Older studies mostly work with manually curated small datasets or do not apply digital

| Data Source | Research Questions       | Literature                                                                                                                                                                                           |
|-------------|--------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Twitter     | Demography, Geography    | García-Albacete; Theocharis [GT14], Kim; Wan; Paris [KWP16], Lansley; Longley [LL16], Sloan et al. [S115], and Theocharis [Th13]                                                                     |
|             | Language                 | Hu et al. [Hu16] and Zainab; Srivastava; Mago [ZSM21]                                                                                                                                                |
|             | Identity, Social classes | Abitbol; Karsai; Fleury [AKF18], Guo et al. [Gu24], Pan et al. [Pa19a], Preoțiuc-Pietro; Lampos; Aletras [PLA15], Sloan et al. [S115], Yan; Zhao; Deng [YZD22], and Zainab; Srivastava; Mago [ZSM21] |
|             | Reactions                | Miyazaki et al. [Mi24] and Zhao; Xi; Zhang [ZXZ21]                                                                                                                                                   |
| Facebook    | Employers and Facebook   | Gee; Jones; Burke [GJB17] and Smith; Kidder [SK10]                                                                                                                                                   |

Tab. 1: Related work studying occupations on social media

methods [FAW22; MLG21]. We find a study working on the same corpus of tweets trying to evaluate essential occupations [TUF23]. This corpus was investigated several times [DT23; HD24; VP23]. Social media data's representativeness for general publications is a frequently raised question [Tu14]. Generally, researchers claim that it is not representative [Ha20; MP17; Ro13]. However, it can also be used to infer real-world data from social media [Wa19]. Additionally, even for marginalized groups, data is available [Ha23].

However, more generic research focused on other data for the labor market in Germany, which depends on academic and vocational education, as well as training, retaining, and continuing education to meet future expectations. Going beyond a surface-level examination of academic and vocational education and schools, we can analyse digital labor market data from online job advertisements (OJAs) or social media data, for example from Twitter/X and YouTube. These can be used for the analysis of OJA can help identify and describe labor market needs, which can help identify skill and educational needs [DJ19; VLV21]. Furthermore, the analysis of social media data can substantiate our knowledge about the broader perception and reception of labor market demands as well as changing skills and qualifications [Pa19b]. In addition, over the past few decades, there has been a growing interest in extracting data from educational databases, advertisements, and information systems [DIH17; MT13; RV07]. It is critical to support decision-making and process management in education. The typical hurdles include automatically extracting knowledge from data, interpreting passages from texts, and mapping to present datasets. However, there remain challenges to be addressed regarding data and data integration [KKM20] and classification [FD24; FDF23].

While Twitter has been utilized as a source for further analysis, its use has primarily been confined to homophily [Pa19a], measuring demographic characteristics [AKF18; FW15; S115], happiness [Gu16], and socioeconomic status [AKF18]. However, existing studies are limited to English texts and no method attempts to produce interoperable data, such as

| Area | Occupations in ...                                                                             | # Tweets |
|------|------------------------------------------------------------------------------------------------|----------|
| 0    | Military occupations                                                                           | 12,425   |
| 1    | Agriculture, forestry, gardening                                                               | 123,531  |
| 2    | Production of raw materials and goods, and manufacturing                                       | 859,805  |
| 3    | Construction, architecture, surveying and technical building services                          | 214,294  |
| 4    | Natural sciences, geography and informatics                                                    | 273,029  |
| 5    | Traffic, logistics, safety and security                                                        | 220,832  |
| 6    | Commercial services, trading, sales, the hotel business and tourism                            | 281,153  |
| 7    | Business organisation, accounting, law and administration                                      | 410,809  |
| 8    | Health care, the social sector, teaching and education                                         | 580,802  |
| 9    | Philology, literature, humanities, social sciences, economics, media, art, culture, and design | 430,801  |

Tab. 2: Number of tweets to a particular occupational area.

connecting tweets to existing data schemes. Although various approaches are available, they are typically constrained to English texts, including resumes [BA11; Ko10]. Investigations into contextualizations with social media have explored occupational differences, but no significant association was identified [TV18]. Since we can only rely on very limited previous work, we will continue with a detailed discussion of the methods to encourage further research in this field.

As shown in Dörpinghaus; Tiemann [DT23] the only ontology available for the German labor market is GLMO. The International Labor Organization (ILO) developed the International Standard Classification of Occupations (ISCO)<sup>4</sup> in 1958 and published subsequent revisions in 1968, 1988, and 2008, with the latest version being in 2008. It has been utilized in the European Union (EU), and German-speaking countries such as Germany, Austria, and Switzerland have created a distinct adaptation of ISCO 2008. ISCO occupations are organized based on skill level and connected to the “European Skills, Competences, Qualifications and Occupations” (ESCO) ontology, which introduces an additional level of hierarchy to the information.

In Germany, the Classification of Occupations, “Klassifikation der Berufe” (KldB), is the reference classification for the Federal Employment Agency (BA) and its research institute (IAB). This classification system ensures occupations are structured at a task level, allowing for clear and concise categorization. The 2020 revision of KldB 2010 is the latest version, rendering the obsolete 1988 and 1992 versions. Its redesign ensured compatibility with ISCO-08. GLMO incorporates both ISCO and KldB, making data interoperable with other systems.

### 3 Data

The data was retrieved using a generic workflow to analyze twitter data, see [DT23]. It comprises three parts: First, reading the data retrieved from Twitter API using the

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

search-term list, then transferring the data using a custom UIMA Typesystem to the second part which analyzes for example the sentiment of words within a Tweet and annotates entities from ESCO and tools, and ultimately passes the result to the writer for storing the output. Besides of UIMA, we use Python with several libraries like SpaCy and Tweepy. The search-term list comprises 526,535 (old and contemporary) synonyms, female, male or neutral forms of occupations. For example, for "B 21112" we find beside others "Bergmechanikerin", "Erzhauer", "Vortriebbergmann" but also "Facharbeiter für Bergbautechnologie - Spezialisierungsrichtung Vortrieb und Abbau". However, with this search strategy we will also find Tweets where occupations are mentioned in other contexts. For anonymized data and data documentation and to reproduce our findings all data is available at <https://zenodo.org/records/14604143>.

So in summary, the dataset contains 3,407,481 X tweets (2007-2023) with occupational IDs, corresponding occupational groups, and occupational areas based on German Classification of Occupations (KldB) codes (e.g., 32103, 321, 3), along with job titles as search terms. Additionally, the dataset includes tweet ID, tweet text, creation timestamp, and, if available, language tags, geographic information, and sentiment data. Thus, data is linked to occupations according to the KldB. For this, we applied a list of occupations and their German synonyms to query X. In Table 2 we present the number of tweets in occupational areas. Our data does not follow population-specific distribution, as social media data never does. However, as our research questions addresses social media, we do not rely on representative data distribution.

However, this dataset contains a broad variety of textual data. We inductively generated five categories. The first category contains data which does not any information related to occupations, for example<sup>5</sup>: „A mathematician in a lab coat. Applied mathematics is more exciting than I expected!“, „Then one without a findable paper. I'll take your word for it that you believe that humans could change anything about the climate. But please don't make yourselves so vulnerable.“, „Career aspiration: To drive one of those cars that push the planes around at the airport Profession: Actuary“.

A second category describes information about a particular occupation, for example: „Field robots are already taking over labor-intensive tasks in the fields and helping agriculture to produce more resource-efficiently. In the future, digital helpers will play an increasingly important role in efficient and sustainable agriculture.“, „Working as an #environmentalmanager is simply extremely exciting. - Portrait of [Person X]“

A third category described people associated with a particular occupation, for example a mathematician or computer scientist working at a new institute or getting a particular award.

The forth category is related to educational aspects, for example VET or CVET: „As an apprentice at lofty heights. Chimney sweep on roofs and constantly on the move. #Chimney sweep“, „Bio + math = biomathematics? How can you become a biomathematician and what professional fields are open to you?“.

<sup>5</sup> Here and in the following, we have translated and paraphrased German tweets.

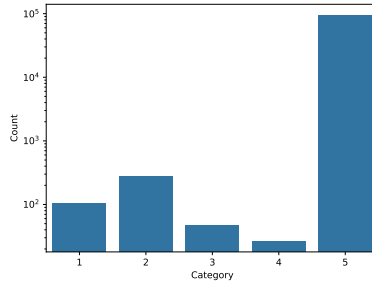


Fig. 1: The distribution of samples training and evaluation data across the different categories. Data are very unbalanced. Especially categories 3 and 4 are strongly underrepresented.

A fifth category holds job offers, either referring to external portals or directly naming in the tweet. These categories help to filter out relevant information on the labor market and particular occupations.

We have manually annotated 93,799 random tweets from occupational area 4 with these five categories. However, nearly all data contains job offers (99.5%), and 0.2% of the annotated data falls in category 2. In Figure 1 we show the distribution of the samples. In the last section, we will discuss that this unbalanced data does not reflect the real data.

## 4 Method

In this paper, we adopt the classification approach to labor market data presented in Fechner; Dörpinghaus [FD24] and Fechner; Dörpinghaus; Firll [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. In order to address the issue of the unbalanced dataset, the fifth class was subjected to downsampling (up to 300 samples with random state=42). We split the data data so that 20% was used as test data and 80% as training data.

As previously mentioned, due to data privacy regulations, we are unable to test API models as GPT-4o or Claude Sonnet. It should be noted that the majority of LLMs were likely trained on social media data. This allows for the possibility of utilizing prompts, as it may be assumed that the model has some degree of comprehension regarding the classes it is intended to map onto, such as individuals or occupations. The experiments were conducted using a group of four openly available (open weights) contemporary finetuned LLMs with the Python-ollama library<sup>6</sup>. The experiment was conducted on a local cluster with NVIDIA L40 GPU.

<sup>6</sup> See <https://github.com/ollama/ollama>.

It should be noted that, due to data privacy laws, API-models such as GPT-4o or Claude Sonnet could not be tested. A variety of prompt strategies were tested. In designing the prompts, we adhered to the general principle for success when engineering a prompt with clear and concise, step-by-step reasoning, as outlined in [FD24]. The outputs of the model are verified heuristically at each intermediate step, and at the end, an output parser makes a final response validation, ensuring that the predicted class is valid. In the event of an invalid assistant response, the query is marked as failed for the specific datapoint. The most common cause of failure was an invalid output format or an invalid classification (i.e. hallucination) of the model.

## 5 Experimental Results

### 5.1 Evaluation of Classifiers

In Table 3 we see that state of the arts models produce reasonable results and still outperform text generation based methods. All baselines apart from the Naive Bayes classifier are within the same performance class, with the Logistic Regression Classifier producing the best  $F_1$  score. As discussed in [FD24] we can assume it also excels compared to Transformerbased language model. Even with down sampled data, all classifiers perform best for category 5, detecting job offers.

| Model               | Metrics (Macro/Weighted) |                  |                  |
|---------------------|--------------------------|------------------|------------------|
|                     | Precision                | Recall           | $F_1$ -Score     |
| Logistic Regression | <b>0.68/0.82</b>         | <b>0.58/0.83</b> | <b>0.61/0.81</b> |
| Naive Bayes         | 0.50/0.70                | 0.43/0.76        | 0.41/0.69        |
| Random Forest       | 0.65/0.80                | 0.55/0.80        | 0.57/0.78        |

Tab. 3: (macro-)  $F_1$ -score, precision and recall for classifiers trained with tf-idf.

In Table 4 we present the results using LLMs. We see that large models generally outperform small ones, e.g., qwen2.5:72b and llama3:70b. In addition, there is not one-fits-all solution for different models and prompting strategies. While for some models, direct strategies lead to best results, SE or ME significantly increased the scores for other models. In addition, we cannot see similar results to [FD24], we do not reach competitive scores on the benchmark without any finetuning. While we do not see high failure rates, the models do not perform as good as the baseline models, the best  $F_1$ -Score has value of 0.55, which does not make them a good choice for solving the task on tweets. A special observation can be made for different categories: While the baseline models performed significantly better on category 5, this does not hold for the LLMs. Here, scores do not indicate that these approaches work best on one particular category.

Thus, for further exploration of the data, we chose Logistic Regression to classify all tweets belonging to category 5.

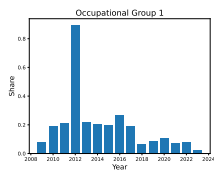
| Model         | Strategy | Metrics (Macro/Weighted) |                   |                   | Failed Samples |
|---------------|----------|--------------------------|-------------------|-------------------|----------------|
|               |          | Precision                | Recall            | $F_1$ -Score      |                |
| llama3.3:70b  | direct   | 0.44/0.53                | 0.47/0.29         | 0.28/0.27         | 0              |
|               | SE       | 0.50/0.60                | 0.54/0.38         | 0.39/0.39         | 0              |
|               | ME       | 0.47/0.58                | 0.57/0.38         | 0.41/0.37         | 0              |
| llama3.2:3b   | direct   | 0.20/0.39                | 0.19/0.20         | 0.14/0.23         | 4              |
|               | SE       | 0.27/0.48                | 0.26/0.42         | 0.24/0.44         | 0              |
|               | ME       | 0.26/0.51                | 0.29/0.32         | 0.23/0.35         | 0              |
| llama3:8b     | direct   | 0.44/0.52                | 0.28/0.51         | 0.27/0.45         | 0              |
|               | SE       | 0.45/0.50                | 0.28/0.46         | 0.28/0.45         | 0              |
|               | ME       | 0.25/0.43                | 0.26/0.38         | 0.22/0.36         | 0              |
| llama3:70b    | direct   | 0.38/0.48                | 0.47/0.30         | 0.29/0.29         | 0              |
|               | SE       | <b>0.54</b> /0.57        | 0.52/0.35         | 0.36/0.35         | 0              |
|               | ME       | 0.53/0.63                | <b>0.60</b> /0.42 | <b>0.43</b> /0.40 | 0              |
| qwen2.5:72b   | direct   | 0.51/ <b>0.64</b>        | 0.51/ <b>0.55</b> | 0.41/ <b>0.55</b> | 0              |
|               | SE       | 0.31/0.51                | 0.33/0.51         | 0.31/0.50         | 0              |
|               | ME       | 0.36/0.52                | 0.33/0.32         | 0.26/0.28         | 0              |
| mistral:7b    | direct   | 0.49/0.57                | 0.44/0.51         | 0.37/0.51         | 0              |
|               | SE       | 0.39/0.55                | 0.52/0.38         | 0.38/0.40         | 0              |
|               | ME       | 0.39/0.54                | 0.45/0.49         | 0.40/0.51         | 1              |
| command-r:35b | direct   | 0.38/0.52                | 0.40/0.31         | 0.27/0.32         | 0              |
|               | SE       | 0.38/0.60                | 0.39/0.28         | 0.27/0.27         | 0              |
|               | ME       | 0.41/0.63                | 0.45/0.34         | 0.32/0.32         | 0              |

Tab. 4: (macro-)  $F_1$ -score, precision and recall for LLM classifiers trained with tf-idf.

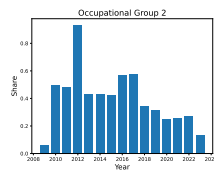
5.2 Evaluation of job offers

As illustrated in Table 5, the distribution of job advertisements across various occupational categories is depicted. It is notable that Group 0, which primarily contains news and advertisements for military occupations, exhibits a paucity of job offers for military occupations. This dataset encompasses both news and advertisements for military occupations in general. However, the proportion of job offers for military occupations is notably high in all other groups, particularly in groups 2 (production of raw materials and goods, and manufacturing) and 3 (construction, architecture, surveying, and technical building services).

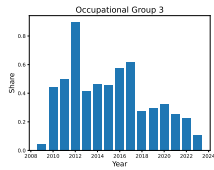
However, a longitudinal perspective reveals a noteworthy shift in the share of job advertisements (see Figures 2a–3). A notable peak emerges in 2012, followed by a significant decline in 2018. A comparison of job offers in 2012 with other years reveals the presence of several authors who demonstrated particularly high levels of activity in 2012. For occupational group 9, in 2012, we identified 13 unique authors who collectively accounted for at least 2.5% of all tweets. In contrast, for all other years, we only identified three unique authors who shared more than 2.5% of the total tweet share. A similar trend was observed in all other occupational groups; however, it was especially pronounced in Occupational Group 4, where one author published nearly 6% of all job advertisements in 2012. Moreover, more than six other authors published over 3.3% of all job advertisements in 2012. This mirrors



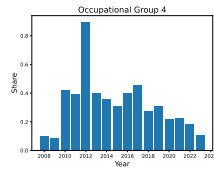
(a). Share of job advertisements in tweets



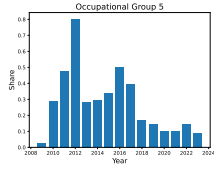
(b). Share of job advertisements in tweets



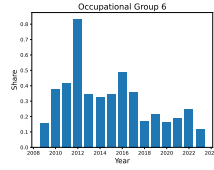
(c). Share of job advertisements in tweets



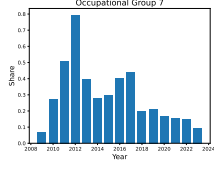
(d). Share of job advertisements in tweets



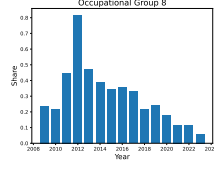
(e). Share of job advertisements in tweets



(f). Share of job advertisements in tweets



(g). Share of job advertisements in tweets



(h). Share of job advertisements in tweets

the growth in employment in Germany for 2012. Despite the European recession gradually also affecting German markets employment would still grow in this year, in particular in services, production and health- and social occupations.

| Occupational group | 0    | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     |
|--------------------|------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Share of ads       | 0.3% | 23.7% | 50.5% | 47.9% | 34.2% | 24.2% | 32.5% | 27.7% | 23.9% | 14.8% |

Tab. 5: After classification, this table presents the share of job advertisements/offers in category 5 according to the total number of tweets.

## 6 Discussion and Outlook

This study shows that the data used can be analysed to yield sensible results. It is shown that they are in parallel to overall labour market trends. While these data and results could then

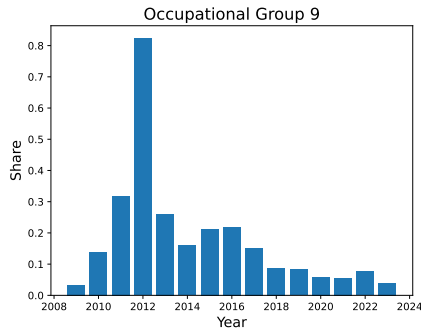


Fig. 3: Share of job advertisements in tweets

be used to enrich and enhance other data (statistical and survey data), there are still clear limitations. One is that we can only see correlations as in parallel trends, but we cannot be sure, for example, whether discussions in social media data are a result of other phenomena, a possible cause or maybe a conjugation. We also have to be cautious in analysing and interpreting these data, as we often only know they will be biased (e.g. the user-base of different social media platforms are not only distinct but also non-representative for the population) and sometimes cannot be sure in what way (it is, e.g. hard to figure out to what extent online job advertisements are representative of all job offerings).

This study has presented an exhaustive examination of job-related tweets, emphasizing the categorization of Twitter data according to the German Classification of Occupations (KldB). The findings offer critical insights into the limitations and opportunities of leveraging social media data for labor market research. The study draws parallels between classical machine learning models and large language models (LLMs), underscoring the strengths and weaknesses of both approaches.

While classical models like Logistic Regression demonstrated robust performance on imbalanced datasets, LLMs exhibited limitations in accuracy despite their capacity to interpret nuanced language patterns. This outcome underscores the necessity for domain-specific fine-tuning when employing LLMs for classification tasks. The annotated dataset revealed a substantial imbalance, with job offers dominating the sample (99.5%).

This distribution poses challenges to the representativeness of the data for broader labor market insights, as categories related to educational aspects or occupational narratives are underrepresented. However, the data's richness in occupational references demonstrates its potential to complement traditional labor market analyses.

The following key findings emerge from our analysis: Firstly, a discernible variation in job advertisement frequencies across various occupational groups is identified, with manufacturing and construction-related occupations exhibiting high representation. These

observations are consistent with established labor market demands within Germany. Secondly, the longitudinal dimension of the dataset offers a distinctive perspective on the evolution of occupational focus and demand over time, particularly in response to societal and technological transformations, such as digitalization. The proposed categorization framework is an effective means of filtering relevant labor market information, thereby laying the groundwork for more granular analyses of specific occupational areas.

In light of these findings, future research should prioritize the following areas: The development of strategies to balance underrepresented categories, potentially through targeted data collection or oversampling, is essential to ensure a more comprehensive representation of labor market discourse. The integration of social media data with other labor market indicators, such as online job advertisements or vocational training statistics, could facilitate a comprehensive understanding of occupational trends and skill demands. Exploration of domain-specific fine-tuning for LLMs holds promise in addressing the performance gap observed in this study. Collaborative efforts to build open-access annotated datasets for German occupational contexts would significantly advance this endeavor. The insights derived from social media data could inform policy decisions in vocational education and training (VET), for example, by identifying emerging skill gaps that could guide curriculum development and resource allocation. The expansion of the methodology to encompass additional languages and regions would facilitate comparative analyses, offering valuable insights into global labor market trends and public perceptions of various occupations.

In conclusion, this research demonstrates the potential of social media as a valuable resource for labor market analysis while highlighting the methodological challenges that must be addressed. By advancing data integration, model development, and policy engagement, future studies can unlock the full potential of digital data in shaping labor market strategies.

## References

- [AKF18] Abitbol, J. L.; Karsai, M.; Fleury, E.: Location, occupation, and semantics based socioeconomic status inference on twitter. In: 2018 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, pp. 1192–1199, 2018.
- [BA11] Ben Abdesslem, W. K.; Amdouni, S.: E-recruiting support system based on text mining methods. *International Journal of Knowledge and Learning* 7 (3-4), pp. 220–232, 2011.
- [DIH17] Dutt, A.; Ismail, M. A.; Herawan, T.: A systematic review on educational data mining. *Ieee Access* 5, pp. 15991–16005, 2017.
- [DJ19] Dörpinghaus, J.; Jacobs, M.: Semantic Knowledge Graph Embeddings for biomedical Research: Data Integration using Linked Open Data. In: SEMANTiCS (Posters & Demos). 2019.
- [Do19] Dobischat, R. et al.: Digitalisierung und die Folgen: Hype oder Revolution? *Bildung 2.1 für Arbeit* 4.0?, pp. 9–24, 2019.
- [DT23] Dörpinghaus, J.; Tiemann, M.: Vocational Education and Training Data in Twitter: Making German Twitter Data Interoperable. *Proceedings of the Association for Information Science and Technology* 60 (1), pp. 946–948, 2023.

- [FAW22] Friehs, M.-T.; Aparicio Lukassowitz, F.; Wagner, U.: Stereotype content of occupational groups in Germany. *Journal of Applied Social Psychology* 52 (6), pp. 459–475, 2022.
- [FD24] 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.
- [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.
- [FW15] Ferguson, H.; Wheat, K. L.: Early career academic mentoring using Twitter: the case of# ECRchat. *Journal of higher education policy and management* 37 (1), pp. 3–13, 2015.
- [GJB17] Gee, L. K.; Jones, J.; Burke, M.: Social networks and labor markets: How strong ties relate to job finding on Facebook’s social network. *Journal of Labor Economics* 35 (2), pp. 485–518, 2017.
- [GT14] García-Albacete, G.; Theocharis, Y.: Opportunities and challenges of analysing Twitter content: A comparison of the occupation movements in Spain, Greece and the United States. *Analyzing social media data and web networks*, pp. 119–153, 2014.
- [Gu16] Guo, W. et al.: Understanding happiness in cities using twitter: Jobs, children, and transport. In: *2016 IEEE International Smart Cities Conference (ISC2)*. IEEE, pp. 1–7, 2016.
- [Gu24] Guo, X. et al.: The evolution of occupational identity in twitter biographies. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 18, pp. 502–514, 2024.
- [Ha20] Hargittai, E.: Potential biases in big data: Omitted voices on social media. *Social Science Computer Review* 38 (1), pp. 10–24, 2020.
- [Ha23] Hakius, B. K.: The reception of holidays in social networks: A case study on Twitter. In (Ganzha, M. et al., eds.): *Proceedings of the 18th Conference on Computer Science and Intelligence Systems, FedCSIS 2023, Warsaw, Poland, September 17–20, 2023*. Vol. 35. *Annals of Computer Science and Information Systems*, pp. 513–522, 2023, <https://doi.org/10.15439/2023F6718>.
- [HBE22] Hooshangi, S.; Buxton, R.; Ellis, M.: Integration of practical computing skills and co-curricular activities in the curriculum. In: *Proceedings of the 27th ACM Conference on on Innovation and Technology in Computer Science Education* Vol. 1. Pp. 61–67, 2022.
- [HD24] Hein, K.; Dörpinghaus, J.: What is said about VET on social media in Germany? Trends, demands, and opinions. In: *NORDYRK 2024, BOOK OF ABSTRACTS*. P. 109, 2024.
- [He16] Helmrich, R. et al.: Digitalisierung der Arbeitslandschaften: keine Polarisierung der Arbeitswelt, aber beschleunigter Strukturwandel und Arbeitsplatzwechsel. *Wissenschaftliche Diskussionspapiere*, 2016.
- [Hu16] Hu, T. et al.: What the language you tweet says about your occupation. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 10 1. 1, pp. 181–190, 2016.
- [KKM20] Kovalev, S.; Kolodenkova, A.; Muntyan, E.: Educational data mining: current problems and solutions. In: *2020 V International Conference on Information Technologies in Engineering Education (Inforino)*. IEEE, pp. 1–5, 2020.

- [Ko10] Kopparapu, S. K.: Automatic extraction of usable information from unstructured resumes to aid search. In: 2010 IEEE International Conference on Progress in Informatics and Computing. Vol. 1, IEEE, pp. 99–103, 2010.
- [KWP16] Kim, S. M.; Wan, S.; Paris, C.: Occupational representativeness in twitter. In: Proceedings of the 21st Australasian Document Computing Symposium. Pp. 57–64, 2016.
- [LL16] Lansley, G.; Longley, P. A.: The geography of Twitter topics in London. *Computers, Environment and Urban Systems* 58, pp. 85–96, 2016.
- [Mi24] Miyazaki, K. et al.: Public perception of generative AI on Twitter: an empirical study based on occupation and usage. *EPJ Data Science* 13 (1), p. 2, 2024.
- [MLG21] Maares, P.; Lind, F.; Greussing, E.: Showing off your social capital: Homophily of professional reputation and gender in journalistic networks on Twitter. *Digital Journalism* 9 (4), pp. 500–517, 2021.
- [MP17] Mellon, J.; Prosser, C.: Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics* 4 (3), p. 2053168017720008, 2017.
- [MT13] Mohamad, S. K.; Tasir, Z.: Educational data mining: A review. *Procedia-Social and Behavioral Sciences* 97, pp. 320–324, 2013.
- [Pa19a] Pan, J. et al.: Twitter homophily: Network based prediction of user's occupation. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Pp. 2633–2638, 2019.
- [Pa19b] Papoutsoglou, M. et al.: Extracting knowledge from on-line sources for software engineering labor market: A mapping study. *IEEE Access* 7, pp. 157595–157613, 2019.
- [Pe11] Pedregosa, F. et al.: Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12, pp. 2825–2830, 2011.
- [PLA15] Preoțiuc-Pietro, D.; Lampos, V.; Aletras, N.: An analysis of the user occupational class through Twitter content. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Pp. 1754–1764, 2015.
- [Ro13] Rost, M. et al.: Representation and communication: Challenges in interpreting large social media datasets. In: Proceedings of the 2013 conference on Computer supported cooperative work. Pp. 357–362, 2013.
- [RV07] Romero, C.; Ventura, S.: Educational data mining: A survey from 1995 to 2005. *Expert systems with applications* 33 (1), pp. 135–146, 2007.
- [Sc22] Schiersmann, C.: Weiterbildungsberatung im Kontext der Nationalen Weiterbildungsstrategie: Finanzielle und strukturelle Aspekte. *Hessische Blätter für Volksbildung* 72 (1), pp. 43–53, 2022.
- [SHL22] Sarsa, S.; Hellas, A.; Leinonen, J.: Who Continues in a Series of Lifelong Learning Courses? In: Proceedings of the 27th ACM Conference on on Innovation and Technology in Computer Science Education Vol. 1. Pp. 47–53, 2022.
- [SK10] Smith, W. P.; Kidder, D. L.: You've been tagged!(Then again, maybe not): Employers and Facebook. *Business Horizons* 53 (5), pp. 491–499, 2010.
- [SI15] Sloan, L. et al.: Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PloS one* 10 (3), e0115545, 2015.
- [Th13] Theocharis, Y.: The wealth of (occupation) networks? Communication patterns and information distribution in a Twitter protest network. *Journal of Information Technology & Politics* 10 (1), pp. 35–56, 2013.

- [Tu14] Tufekci, Z.: Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In: Proceedings of the international AAAI conference on web and social media. Vol. 8 1. 1, pp. 505–514, 2014.
- [TUF23] Tiemann, M.; Udelhofen, S.; Fournier, L.: What social media can tell us about essential occupations. 2023.
- [TV18] Tifferet, S.; Vilnai-Yavetz, I.: Self-presentation in LinkedIn portraits: common features, gender, and occupational differences. *Computers in Human Behavior* 80, pp. 33–48, 2018.
- [VLV21] Verma, A.; Lamsal, K.; Verma, P.: An investigation of skill requirements in artificial intelligence and machine learning job advertisements. *Industry and Higher Education* 36 (1), pp. 63–73, 2021.
- [VP23] Vahdatnia, A.; Peachkah, D.: Monitoring of Digitization and Sustainability on Twitter. 2023.
- [Wa19] Wang, Z. et al.: Demographic inference and representative population estimates from multilingual social media data. In: The world wide web conference. Pp. 2056–2067, 2019.
- [WW17] Warning, A.; Weber, E.: *Wirtschaft 4.0: Digitalisierung verändert die betriebliche Personalpolitik*, tech. rep., IAB-Kurzbericht, 2017.
- [YZD22] Yan, S.; Zhao, T.; Deng, J.: Predicting Social Media User Occupation with Content-aware Hierarchical Neural Networks. In: 2022 8th International Conference on Big Data and Information Analytics (BigDIA). IEEE, pp. 388–395, 2022.
- [ZSM21] Zainab, K.; Srivastava, G.; Mago, V.: Identifying health related occupations of Twitter users through word embedding and deep neural networks. *BMC bioinformatics* 22 (Suppl 10), p. 630, 2021.
- [ZXZ21] Zhao, Y.; Xi, H.; Zhang, C.: Exploring occupation differences in reactions to COVID-19 pandemic on twitter. *Data and Information Management* 5 (1), pp. 110–118, 2021.