

Occupations and Education in X Data: How representative is the data?

Michael Tiemann ^{1,2} and Jens Dörpinghaus ^{1,2,3}

Abstract: Valuable insights can be gained regarding jobs and professions across various sectors of society based on their inherent and acquired traits. Previous studies relied on methods such as action research, surveys, and questionnaires that are time-consuming and resource-intensive. This study examines vocational education and training data on Twitter. Although the data has been utilized in multiple studies, we will examine a vital research inquiry within computational social science: Is it plausible to employ Twitter/X data for analyzing vocational education and training in Germany or does the data display excessive bias? This investigation is infrequently explored since most researchers endeavor to discover representative samplings of larger subsets, and gauging representativeness against a ground truth can prove challenging. However, we will demonstrate that with research inquiry and statistical data, it is feasible to calculate a representative distance d , correction factors κ , and an overall bias γ . This provides a unique technique towards labor market research that makes novel data interoperable, which has not been considered in previous literature. Our approach is versatile and can be readily extended to other data.

Keywords: Computational Social Sciences, Labor market research, social media, Twitter

1 Introduction

The German labor market is dynamic: technical innovations and social changes lead to new skill demands. Within the German education system - in contrast to other countries - vocational education and training (VET), retraining and continuing vocational education and training (CVET), play a pivotal role in responding to these new requirements. This is also resembled by roughly half of the workforce having a (C)VET training as their highest vocational education. [Do19; He16]

The German education system offers different pathways for professionals, see [GL21]. These are negotiated between four different stakeholders (educational institutions, companies and enterprises, employees, and financiers), who all have to react to emerging situations. From a research perspective, the task is to determine which education content is increasingly being offered and demanded in order to draw conclusions about the development needs of the vocational training system. We find information on these in different data sources.

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

² 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>

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

Our primary research questions is: If we want to utilise data from Twitter/X or other social media, how representative are they? In other words: “Can we utilize Twitter/X data as an interoperable data source extending existing resources for analysing vocational education and training in Germany?”

We will present an overview, followed by a segment covering current research and related work, outline the methodology employed in the study, experimental results and assesses the X dataset, relating to research on the German labor market as well as conclusions and an outlook.

2 Related Work

The labor market for IT professionals in Germany depends on academic and vocational education, as well as training, retraining, 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]. 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; SI15], happiness [Gu16], and socioeconomic status [AKF18]. However, existing studies are limited to English texts and no method attempts to produce interoperable data, such as 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.

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]. Biases also arise from inequalities in access to social media, since not every person has comparable chances in

terms of access to internet-providers and online access. While it would be conceivable to construct measures to account for such inequalities, these would be highly complex. In essence, the best way would be to find representative survey data for all times and regions matching the Twitter/X data. Such surveys cannot be done in hindsight. A second best way would be to use some survey data and infer probable correction factors. This would still be complex and computing intensive, but not necessarily increase weighting efficiency in comparison to simpler correction factors like the ones proposed here.

As shown in [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)⁴ 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 nations 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. 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.

While the extraction of skills, tools and other entities from labor market data is known to be hard [DD23; DJ20; DSH23], in the next section we describe the data foundation and methods to analyze their representation of the labor market. In general, context makes these problems hard [Dö20; FDF23].

3 Data and Methods

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 from the database, then transferring the data using a custom UIMA Typesystem to the second part which analyzes 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 dataset includes occupational IDs, corresponding occupational groups, and occupational areas based on 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.

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

Area	# Tweets
0	12,425
1	123,531
2	859,805
3	214,294
4	273,029
5	220,832
6	281,153
7	410,809
8	580,802
9	430,801

Tab. 1: Number of tweets to a particular occupational area.

In general, out of 3,407,481 tweets with duplicates, 88.5% were neutral, 9.3% negative and 2.2% positive. Very few tweets contain geographical information (ca. 3%). The amount of tweets for a given occupational area are presented in Table 1. In [DT23] we showed that our approach makes tweets interoperable with KldB and ESCO.

We can now assume that we have all tweets related to any occupation in KldB⁵. It should be a reflection of reality. A representative sample is defined as a smaller version of the population that captures its features [GS14]. Finding representative subsets of a large data set is well-studied in literature [Bo12; Pa05], but in our case we need to test if a data set actually is a representative sample.

The initial concern is what ground truth we want to compute representativeness for. Given that we are working with categorized data, we will assume classes $\mathbb{C} = C_1^y, \dots, C_n^y$ for a timepoint, e.g., years in $\mathcal{T}$, y and a ground truth sample size of $s : \mathbb{C} \rightarrow \mathbb{N}$, where $s(C_1^y), \dots, s(C_n^y)$ can be compared with the samples in $\sigma : \mathbb{C} \rightarrow \mathbb{N}$, $\sigma(C_i)$. This will enable the computation of the representative distance as $d : \mathbb{C} \rightarrow \mathbb{R}$.

$$d(C_i^y) = \frac{s(C_i^y) - \sigma(C_i^y)}{s(C_i^y)}.$$

We can also describe a correction factor for a class C_i as

$$\kappa(C_i) = \sum_{y \in \mathcal{T}} \frac{d(C_i^y)}{|\mathcal{T}|},$$

and the overall bias by

$$\gamma = \sum_{i \in \{1, \dots, n\}} \frac{\kappa(C_i)}{n}.$$

⁵ Although this assumption is not true due to spelling mistakes and indirect references, we have however a common problem in data science

However, the ground truth is determined by the research question. Therefore, we will compare the data under the following circumstances:

We will begin with a spatial analysis, which is significant because only a small fraction of the data includes geographic information. This analysis will determine whether Twitter data can be used to analyze regional phenomena.

The second and third analyses will focus on Occupational analysis, both in general and concerning vocational education. Here, we will first analyze the complete dataset to determine over- and underrepresented occupations in a given year. We will then limit our analysis to tweets mentioning vocational education and examine whether this over- or underrepresents novel training contracts in that year. Every analysis can be carried out on different levels of KldB in GLMO, for example 9 occupational areas or occupational groups.

Overall, our analytical stance resembles a marginal weighting, where a given sample would be compared to a known population. The difference is that we cannot easily compute matrices for comparison. Most commonly samples would be compared according to educational attainment, age, regions, or occupations. We cannot easily use such information. The problem of authenticity would also complicate matters since it is not established whether self-representation in social media is authentic (in favor: [KMR22]) or not (e.g. [To17]). Self-representation is also hindered by undisclosed preselections between front- and back-ends [Re14], where back-ends superimpose structures and opportunities for self-representation (e.g. by only allowing for giving age-ranges). On top of this researchers cannot be sure to what aim social media data have been gathered and produced (both by providers or firms and users), the best guess being commercial uses. Before this backdrop of different sources for variation and biases (access to the internet and social media, questionable self-representation and authenticity, structural problems of preselection and intent) we propose that the best way forward is a correction factor as simple and at the same time as fitting to the research at hand as possible. Notably we find our approach also differs from most found in an extensive meta-study on sampling with Twitter data [Vi23]. While we do apply the “API search followed by secondary data” approach, usually additional metadata on users is collected for weighting, whereas we focus on the aggregate level of occupations.

4 Analysis

4.1 Spatial analysis

In general, out of 2,440,626 non-duplicated tweets, 1,947,744 were neutral, 400,170 negative, and 92,712 positive. This gives a broad basis for testing collocational prestige. And even though only very few tweets contain geographical information (ca. 3%), we can find regional clusters for particular occupations. When presenting the poster, we will not only show that our approach makes tweets interoperable with KldB and ESCO, but also present those cases where we have enough data to process job prestige analysis on a geographical basis.

For this we will compare the standard sociological prestige assignment via occupational classifications to a) tweet sentiments and b) most frequent tokens (words), but also c), applying ideas of embodiment of prestige, co-occurrences of bodily concepts with different prestige values.

Immanent and ascribed job characteristics hold valuable information. They are usually surveyed with action research, questionnaires, interviews, . . . - all intensive in time and resources. Our project shows how social media data, in particular Twitter data, can be used to supplement and enrich such „traditional“ data.

Sentiments and occupations 1 meaningful results (e.g. Fournier et al. 2023) During pandemic occ. In healthcare viewed slightly more positive, construction and logistics slightly more negative

Sentiments and locations 2 For smaller federal states not enough cases for detailed information Number of tweets largely increasing for last decade making them better suited for more current phenomena

4.2 Employee analysis (Mikrozensus)

We will now analyze if the Twitter data represents the total number of workers in an occupational area or group. Statistical data comes from Mikrozensus and BIBB [LR97]. This data is frequently used for labor market analysis and forecasting future trends [He12; Ma14]. Data for occupational groups are only represented at occupational areas and groups. Table 2 denotes the distance d between the proportion of workers in a given year and the proportion of tweets linked to an occupational areas. Positive values for occupations 9 and 4, as well as for 2 and 6, are evidently present in some years.

However, in Figure 3 we display a heatmap representation of both occupational groups and areas. Here, Occupation 9, which includes creative occupations in publishing and media management along with philology. However, we mainly find a peak in field 4 and a large plus in field 2. A closer look shows that it is mainly specific occupational groups that are important for this result:

- 232 Occupations in technical media design
- 233 Occupations in photography and photographic technology
- 733 Occupations in media, documentation and information services
- 911 Occupations in philology
- 923 Occupations in publishing and media management

If we display the data in a box plot, see Figure 4, we can immediately see that occupational area 9 is the main area, with much more tweets about vocational training than other areas, which usually have fewer tweets.

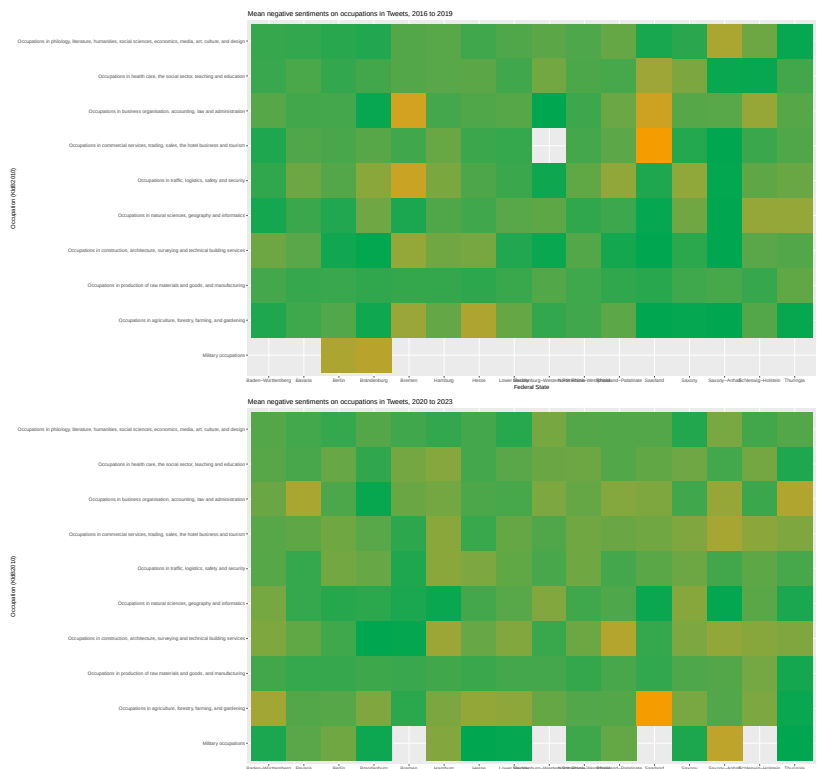


Fig. 1: Sentiments and occupations meaningful results (e.g. Fournier et al. 2023) During pandemic occ. In healthcare viewed slightly more positive, construction and logistics slightly more negative

To calculate the correction factor κ , we simply take the mean of d over time. For area 9, this value is 1.995, with all other values falling within a range of -0.56 to 1.24. In this instance, $\gamma = 0.249$.

When examining occupational groups, some may have larger correction factors.

- 232 26.200023
- 233 88.909980
- 281 6.877897
- 733 21.700635
- 923 50.572560

An overall bias of $\gamma = 1.81$ means that there is a significant disparity. Therefore, we can conclude that the information on occupational areas can be utilized with a correction factor,



Fig. 2: Sentiments and locations For smaller federal states not enough cases for detailed information
Number of tweets largely increasing for last decade making them better suited for more current phenomena

particularly for certain fields. However, this may not be feasible for occupational groups without further limitations.

4.3 Vocational training analysis (Dazubi)

Data from the vocational training statistics of the Federal and State Statistical Offices can be accessed through DAZUBI. The annual total survey incorporates data on vocational training in accordance with the Vocational Training Act (BBiG) and the Crafts Code (HwO), and covers trainee, contract, and examination data. Table 3 denotes the distance d between the proportion of new training contracts in a given year and the proportion of tweets linked to vocational training. Positive values for occupations 2 and 9, as well as for 5 and 8, are evidently present in some years.

	2010	2011	2012	2013	2014	2015
Area						
1	-0.057806	-0.610543	-0.272339	-0.675933	-0.245685	-0.681493
2	1.507389	0.752382	0.149469	0.626881	0.806626	0.657555
3	-0.475278	-0.391903	-0.187396	-0.222087	-0.371572	-0.120870
4	0.140821	0.969498	1.184526	0.784938	1.015686	2.673778
5	-0.638295	-0.222046	-0.557111	-0.494134	-0.536415	-0.515042
6	-0.168371	0.240505	0.723188	0.373695	-0.061697	0.220233
7	-0.448906	-0.505607	-0.200808	-0.322521	-0.229641	-0.421173
8	-0.684885	-0.653580	-0.483452	-0.516870	-0.462670	-0.526139
9	2.526545	3.210485	2.463338	2.529874	2.309713	1.650335
	2016	2017	2018	2019	2020	2021
Area						
1	-0.575837	-0.449622	-0.383116	0.092686	0.187024	1.026898
2	0.605672	0.572571	1.157838	1.404790	1.855483	1.853060
3	-0.443410	-0.569739	-0.729240	-0.544924	-0.191452	-0.402928
4	2.591697	3.633144	0.773455	0.397584	0.423998	0.338281
5	-0.456436	-0.423194	-0.398711	-0.699558	-0.776967	-0.814641
6	-0.099535	0.100120	0.320189	0.454824	-0.004805	-0.033165
7	-0.284219	-0.398783	-0.537804	-0.553030	-0.741410	-0.731737
8	-0.361465	-0.475068	-0.621229	-0.707618	-0.662505	-0.601879
9	1.762657	1.469142	1.816641	1.482075	1.228522	1.495865

Tab. 2: The distance d for different occupational areas (1-9) and the size of X samples per year.

However, in Figure 5 we display a heatmap representation of both occupational groups and areas. Here, Occupation 2, which includes creative occupations such as photographers and media designers, is important. However, we mainly find a peak in field 5 and a large plus in field 9. A closer look shows that it is mainly specific occupational groups that are important for this result:

- 933 Occupations in artisan craftwork and fine arts
- 913 Occupations in the social sciences
- 281 Occupations in textile making
- 311 Occupations in construction scheduling and supervision, and architecture
- 611 Occupations in purchasing and sales.

If we display the data in a box plot, see Figure 6, we can immediately see that occupational area 9 is the main area, with much more tweets about vocational training than other areas, which usually have fewer tweets than new contracts and no more than twice as many tweets.

This approach is reflective of actual importance, as some occupations have a negligible

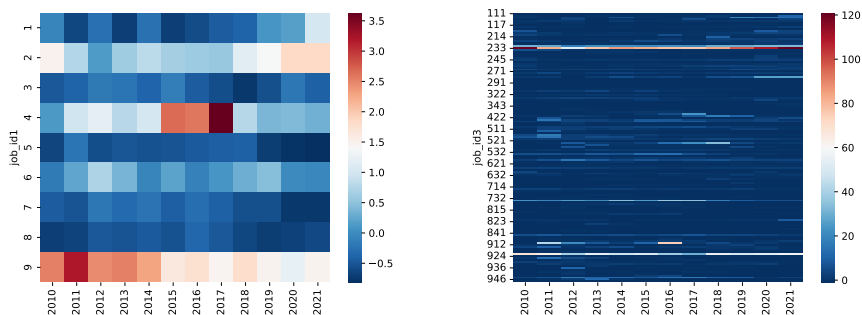


Fig. 3: The distance d for different occupational areas (1-9, left), occupational groups (right) and the size of X samples per year.

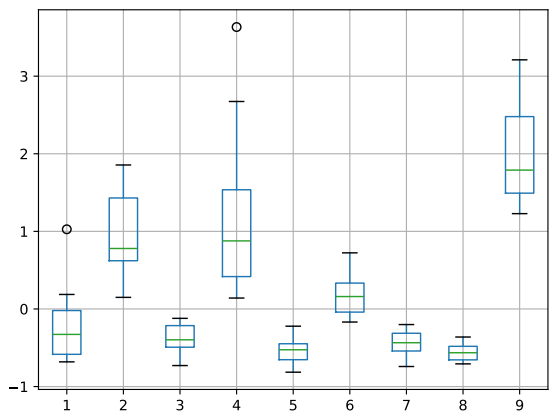


Fig. 4: A box plot depicting the distance d for various occupational areas (1-9) and X samples per year.

share of new contracts. To calculate the correction factor κ , we simply take the mean of d over time. For area 9, this value is 5.175203, with all other values falling within a range of -0.63 to 0.35. In this instance, $\gamma = 0.454$.

When examining occupational groups, some may have larger correction factors.

- 232 18.325647
- 233 69.444706
- 281 56.907285
- 311 42.570592

	2010	2011	2012	2013	2014	2015	2016
Area							
1	1.262076	-0.254492	0.233715	-0.692853	-0.509986	-0.082682	-0.572713
2	-0.581133	-0.119071	-0.276161	0.384254	1.198552	0.645691	0.906384
3	-0.639111	-0.792622	-0.456121	-0.538606	-0.741887	-0.590211	-0.609194
4	0.145466	-0.198800	0.581954	-0.630980	-0.606135	-0.611569	-0.576467
5	1.449583	2.472200	0.291577	-0.508918	-0.828630	-0.511640	-0.546336
6	0.254521	-0.583069	-0.266035	0.382154	-0.688291	-0.463526	-0.513942
7	-0.595705	-0.000391	0.200119	-0.553938	-0.685450	-0.803155	-0.746415
8	1.342296	-0.164697	0.626277	-0.072432	-0.149516	0.170831	-0.361241
9	1.210632	7.032732	2.508180	1.902418	5.560484	8.737524	8.147497

	2017	2018	2019	2020	2021	2022
Area						
1	-0.257034	-0.525342	-0.452591	0.320748	-0.197276	-0.328932
2	0.671837	0.342820	1.098826	0.244159	0.130930	0.017536
3	-0.734076	-0.598787	-0.592948	-0.756362	-0.591968	-0.554190
4	-0.643672	0.061197	0.191085	1.115724	0.573212	0.260722
5	-0.176176	0.444593	-0.507342	-0.316500	-0.412498	-0.426818
6	-0.239203	-0.096526	-0.619100	-0.349356	0.333653	0.019371
7	-0.801745	-0.646549	-0.663125	-0.364208	-0.181107	-0.231574
8	-0.296260	-0.174626	-0.437058	-0.238563	-0.332810	0.293543
9	7.845721	5.087217	2.540952	8.686764	2.471263	5.546260

Tab. 3: The distance d for different occupational areas (1-9) and the size of X samples per year.

An overall bias of $\gamma = 4.05$ means that there is a significant disparity. Therefore, we can conclude that the information on occupational areas can be utilized with a correction factor, particularly for certain fields. However, this may not be feasible for occupational groups without further limitations.

5 Conclusions and Outlook

The objective of this study was to assess the representativeness of Twitter/X data for labor market analysis, with a particular emphasis on vocational education and training (VET) in Germany. Through spatial, occupational, and training-related analyses, we demonstrated that despite inherent limitations, social media data can offer valuable insights when aligned with established classification systems such as KldB and ESCO.

The findings of this study demonstrate the particular utility of Twitter/X data at the level of occupational areas. While strong deviations from reference statistics are evident for specific occupational groups, particularly those related to media, arts, and social sciences, the application of correction factors enables the meaningful integration of these data. This

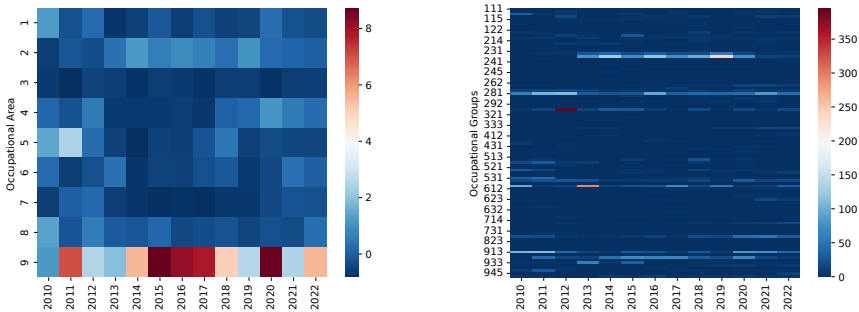


Fig. 5: The distance d for different occupational areas (1-9, left), occupational groups (right) and the size of X samples per year.

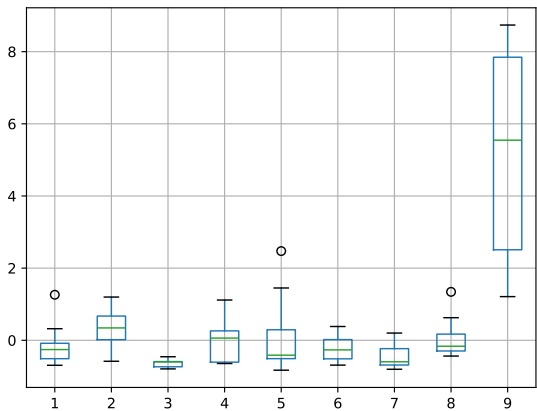


Fig. 6: A box plot depicting the distance d for various occupational areas (1-9) and X samples per year.

approach offers a promising method to bridge traditional labor market statistics and real-time, user-generated social media content. However, the significant overall biases observed at the group level indicate the necessity of exercising caution and avoiding simple extrapolations without further adjustments.

In the future, further research should expand the application of correction factors to additional domains and examine the possibility of adapting similar approaches to other areas, such as regional analyses or the study of qualification trends. Furthermore, the incorporation of multimodal data sources, such as the integration of social media with job advertisements or educational databases, has the potential to enhance the robustness of the model and address existing coverage gaps. The development of more sophisticated methods for sentiment and

semantic analysis, particularly for German-language texts, is another critical step in ensuring the reliability of social media as a component in labor market monitoring.

This study contributes to a growing body of work exploring the use of novel data sources in computational social science. It underlines both the potential and the boundaries of social media as a data source for labor and education research.

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.
- [Bo12] Borovicka, T. et al.: Selecting representative data sets. *Advances in data mining knowledge discovery and applications* 12/, pp. 43–70, 2012.
- [DD23] Derksen, F.; Dörpinghaus, J.: Digitalization and Sustainability in German Continuing Education. *Lecture Notes in Informatics (LNI)/*, 2023.
- [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: *SEMAN-TiCS (Posters & Demos)*. 2019.
- [DJ20] Dörpinghaus, J.; Jacobs, M.: Knowledge detection and discovery using semantic graph embeddings on large knowledge graphs generated on text mining results. In: 2020 15th conference on computer science and information systems (fedcsis). IEEE, pp. 169–178, 2020.
- [Do19] Dobischat, R. et al.: Digitalisierung und die Folgen: Hype oder Revolution? *Bildung 2.1 für Arbeit 4.0?*, pp. 9–24, 2019.
- [Dö20] Dörpinghaus, J. et al.: Towards context in large scale biomedical knowledge graphs, 2020, arXiv: 2001.08392 [cs.DB], URL: <http://arxiv.org/abs/2001.08392>.
- [DSH23] Dörpinghaus, J.; Samray, D.; Helmrich, R.: Challenges of Automated Identification of Access to Education and Training in Germany. *Information* 14/10, p. 524, 2023.
- [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.

- [FDF23] Fechner, R.; Dörpinghaus, J.; Firll, A.: Classifying Industrial Sectors from German Textual Data with a Domain Adapted Transformer. In (Ganzha, M. et al., eds.): Proceedings of the 18th Conference on Computer Science and Intelligence Systems. Vol. 35. Annals of Computer Science and Information Systems, IEEE, 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.
- [GL21] Graf, L.; Lohse, A. P.: Advanced skill formation between vocationalization and academization: The governance of professional schools and dual study programmes in Germany. *Governance Revisited. Challenges and Opportunities for Vocational Education and Training*; Gonon, P., Bürgi, R., Eds/, 2021.
- [GS14] Grafström, A.; Schelin, L.: How to select representative samples. *Scandinavian Journal of Statistics* 41/2, pp. 277–290, 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.
- [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, URL: <https://doi.org/10.15439/2023F6718>.
- [He12] Helmrich, R. et al.: Engpässe auf dem Arbeitsmarkt: Geändertes Bildungs- und Erwerbsverhalten mildert Fachkräftemangel. *Neue Ergebnisse der BIBB-IAB-Qualifikations- und Berufsfeldprojektionen bis zum Jahr 2030/*, pp. 1–14, 2012.
- [He16] Helmrich, R. et al.: Digitalisierung der Arbeitslandschaften: keine Polarisierung der Arbeitswelt, aber beschleunigter Strukturwandel und Arbeitsplatzwechsel. *Wissenschaftliche Diskussionspapiere*, 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.
- [KMR22] Kreling, R.; Meier, A.; Reinecke, L.: Feeling Authentic on Social Media: Subjective Authenticity Across Instagram Stories and Posts. *Social Media + Society* 8/1, 2022, ISSN: 2056-3051.
- [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.

- [LR97] Lüttinger, P.; Riede, T.: Der Mikrozensus: Amtliche Daten für die Sozialforschung. ZUMA Nachrichten 21/41, pp. 19–43, 1997.
- [Ma14] Maier, T. et al.: Löhne und berufliche Flexibilitäten als Determinanten des interaktiven QuBe-Arbeitsmarktmodells: ein Methodenbericht zur Basisprojektion der dritten Welle der BIBB-IAB Qualifikations- und Berufsfeldprojektionen. Wissenschaftliche Diskussionspapiere, 2014.
- [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.
- [Pa05] Pan, F. et al.: Finding representative set from massive data. In: *Fifth IEEE International Conference on Data Mining (ICDM'05)*. IEEE, 8–pp, 2005.
- [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.
- [Re14] Reichert, R.: Facebook und das Regime der Big Data. *Österreichische Zeitschrift für Soziologie* 39/, pp. 163–179, 2014.
- [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.
- [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.
- [To17] Toma, C. L.: Taking the good with the bad. Effects of Facebook Self-Presentation on Emotional Well-Being. In (Reinecke, L.; Oliver, M. B., eds.): *The Routledge Handbook of Media Use and Well-Being. International Perspectives on Theory and Research on Positive Media Effects*. Routledge, pp. 170–182, 2017.
- [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, pp. 505–514, 2014.
- [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.

- [Vi23] Vicente, P.: Sampling Twitter users for social science research: evidence from a systematic review of the literature. *Quality & Quantity* 57/6, pp. 5449–5489, 2023, ISSN: 1573-7845.
- [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.
- [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.