

An empirical analysis of incentive structures in German online job advertisements using a topic modeling approach

Michelle Katharina Gassner^{1,2}, Michael Tiemann ^{2,3}, and Jens Dörpinghaus ^{2,3,4}

Abstract: In light of the imminent labor shortage, companies find themselves compelled to enhance their standing as desirable employers in the competitive pursuit of talent. This endeavor can be facilitated through the medium of online job advertisements (OJAs), which serve as a means to subtly communicate the merits of an organization to prospective employees. The objective of this study is to develop and evaluate a topic-modeling approach, known as Latent Dirichlet Allocation (LDA), for analyzing online data. The study will also discuss the strengths and weaknesses of the approach. The work will also include considerations on social sciences theory, signaling theory, and methods will be evaluated.

Keywords: Topic Modeling, Online Job Advertisements, Computational Social Sciences

1 Introduction

Job advertisements have been a subject of study in various specialist disciplines in Germany for an extended period, with scholars examining them for specific entities related to digitalization or sustainability, see [BS22; HR24]. In the context of longitudinal analysis of skills and tools, job advertisements have been utilized as a medium for extracting relevant information. Additionally, it has been utilized in the context of longitudinal analysis of skills and tools [Dö24; DWB22; Er23], and more broadly for text mining approaches [Gn22; Sc24].

In the context of the imminent shortage of skilled labor, companies find themselves compelled to fortify their position as appealing employers in the fierce competition for talent, a pursuit often referred to as the "war for talent". A pivotal medium through which this endeavor can be realized is online job advertisements (OJAs), which serve as a means for companies to discreetly communicate their merits to prospective employees. This exploratory analysis examines the incentive structures employed in the early recruitment process for the occupational group of natural scientists. The analysis is based on a dataset comprising 200,000 online job advertisements from 2015 to 2022.

¹ Universität Bonn, Germany

² 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>

³ 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>

⁴ Linnaeus University, Department of Computer Science and Media Technology, Växjö, Sweden,
doerpinghaus@uni-koblenz.de,  <https://orcid.org/0000-0003-0245-7752>

The aim of this study is thus to investigate which incentive structures are formulated in job advertisements from an employer's perspective and to what extent the topics covered can be identified using an automated process. The main research questions are: Is it possible to utilize topic-modeling approaches to analyze incentive structures in OJAs? If yes, can we see whether and to what extent these topics change in their frequency and thus importance in job advertisements over a period of seven years (2015-2022)? Is it possible to embed the results of computational methods within social sciences theory, in particular signaling theory? This approach builds on previous research on the topic of job advertisements and the theoretical background of employer branding and signaling.

In order to accomplish this objective, it is necessary to discuss the special features of the job advertisement as a discrete text genre and subsequently link them to specific benefits and their signaling effects, such as career development opportunities and work-life balance. The signaling effects of the benefits are then classified by examining the generational division, and all elements are placed in an overall context before formulating assumptions from the preliminary theoretical considerations.

The paper is organized into six sections. The first section provides an introduction and a brief overview of the background of the research question. The second section presents related work. In the third section, computational methods and theories from the social sciences used to answer the research question are introduced. The fourth section presents the data. The fifth section is dedicated to experimental results and the evaluation of these methods. In this section, a detailed interpretation of the results is provided. The final section of the paper presents our conclusions and an outlook onto future research.

2 Related Work

From a linguistic perspective, OJAs have already been examined in detail by Ehrenheim [Eh11], among others. She looks at how the job advertisement is characterized in this respect in terms of application and how it is also used across the board in other disciplines such as business administration and corporate communications, especially marketing. She comes to the conclusion that only by looking beyond the textual part (for example, the visual design) can all the characteristics of the text type and its interactions be identified and correctly classified. Gansel [Ga07] also worked on the formal structure of the job advertisement, which is often referred to in other works on job advertisements.

Online job advertisements are also examined from various perspectives diachronic/multimodal, in linguistic and cultural contrast, from a legal and gender-specific perspective and from a perspective of change, see Nielsen; Luttermann; Lévy-Tödter [NLL17]. The object of investigation is print job advertisements as well as online job advertisements. Many contributions in the anthology examine the special features between German and foreign job advertisements from different perspectives, such as Luttermann; Engberg [LE17]. Interestingly, the largest data set examined in this anthology comprises a corpus size of 285

job advertisements [Sc17]. Particularly noteworthy are the works of Stumpf [St17], who examines the extent to which branding strategies are implemented in job and advertising advertisements on the basis of 125 German and Austrian job advertisements. He uses the perspective of the “Unique Employment Proposition” (UEP) and the “Unique Selling Proposition” (USP), whereby he comes to the conclusion that the UEP is expressed less in job advertisements than the USP and therefore employer branding is only implemented to a limited extent through the job advertisement.

Stohr [St19] examined the professional requirements of digitalization with regard to formal, physical and competence-specific aspects in his work. He examined a data set of around 150,000 online job advertisements from 2017 using a topic modeling method to prepare the analysis. He used a GloVe algorithm for this purpose. His analysis focuses on the requirement needs that arise in some occupational fields as a result of digitalization and how these structures can be identified in the job advertisement. The aim of this partial analysis using topic modelling is to generate categories of requirements that have similarities in the way they can be digitized. The result of this partial analysis is that, due to the homogeneity in the area of competence requirements (with a pre-assumption of five topics as fixed model parameters), no five differentiated topics can be identified in the area of the digitization groups identified by him, nor any specific topics in the area of requirements.

Benefits and fringe benefits are primarily examined from the employee’s perspective, particularly in terms of which factors make employers attractive. For example, Firfiray; Mayo [FM17] investigated the extent to which additional benefits can influence the impression of a personal fit between the company and the individual and how the offer of work-life benefits, among other things, affects attractiveness, also with regard to different generations. They came to the conclusion that work-life benefits in particular can trigger a stronger attraction and a higher perceived fit with the company.

It can be said that the literature research carried out did not explicitly identify the actual state of the formulated benefits on the part of the employer. In this paper, the present job advertisement corpus is not considered as a whole, but only the offer of benefits is examined. For this purpose, linguistic and formal characteristics of the job advertisement are taken into account in the analysis. In addition, employer branding is also used as a theoretical basis and expanded to include the signaling perspective. A topic modeling approach is also chosen as the method, but unlike Stohr [St19], an LDA algorithm is applied and the effect of various model parameters is exploratively examined. In addition, it is assumed that the job advertisement corpus in the area of benefits has greater diversity than the formulated competence requirements due to the greater signaling effect in employer branding. The analysis using topic modeling could provide new insights into underlying structures in the formulation of benefits.

In summary, we can find several gaps in research, in particular for the study of German OJAs. This study contributes by investigating which incentive structures are formulated in job advertisements from an employer’s perspective and to what extent the topics covered

can be identified using an automated process. Furthermore, it analysis whether and to what extent these topics change in their frequency and thus importance in job advertisements over a period of seven years (2015-2022). This approach builds on previous research on the topic of job advertisements and the theoretical background of employer branding and signaling. Furthermore, with our proposed combination the methodological approach of the automated, quantitative procedure of topic modeling is also extend.

3 Method

3.1 Signaling as a method of employer branding

The signaling theory, first articulated by Michael Spence in the 1970s, addresses the information asymmetry between employers and employees, wherein signalers possess a greater depth of knowledge than receivers [Sp73]. Employers are considered “insiders”, while potential employees, as “outsiders”, are deficient in crucial information [Co11]. Outsiders are compelled to acquire this data to facilitate informed decision-making, such as the selection of an employer.

Employers, in their capacity as insiders, possess knowledge about job roles and organizational qualities that is inaccessible to job seekers. Conversely, employers lack complete information about candidates’ attributes, rendering recruitment evaluations challenging [WB18]. Spence [Sp74] characterized signals as actions or attributes conveying information in markets, and Wolf; Berg [WB18] noted that they help bridge the aforementioned gap. Employers have the capacity to shape these signals, particularly their employer image, which reflects perceptions of the organization [CT01]. Factors influencing this image include organizational characteristics, job details, and employee experiences. Signals may be intentional, such as advertisements, or independent, like media coverage or peer opinions [CT01]. Early-stage recruitment signals aim to attract attention, while independent sources validate these messages. The effectiveness of signals is contingent upon their observable nature, as it establishes a baseline of attractiveness and fosters deeper engagement from candidates [Co11]. The practice of false signaling, wherein organizations exaggerate their qualities, erodes trust. Candidates misled by such signals are susceptible to dissatisfaction, diminished commitment, and premature turnover [CT01]. Conversely, credible signals, such as third-party certifications, distinguish high-quality firms and ensure competitive advantage [Co11]. Employees’ feedback on platforms like Glassdoor and Kununu provides additional layers of reality, challenging overly idealized portrayals and encouraging transparency.

Conflicting employer websites and reviews, for example, can lead to a reduction in perceived attractiveness. However, credible critiques enhance realism and trust, while offering feedback to companies to refine their internal structures and signals [PLM21]. These adjustments are critical for maintaining competitiveness in a transparent labor market [DKD17].

3.2 The online job advertisement as a signaling medium

Job advertisements frequently serve as the initial substantial and coherent expression of recruitment efforts, functioning not only as appeals but also as representations of the position and as a medium to showcase the company's attractiveness [Je20; NLL17]. The self-expression component in job advertisements has undergone significant growth, thereby linking them more closely with employer branding. The employment sector has adopted common parlance such as "job posting" and "job offer", with the latter implying an invitation to embrace the company's values [Je20].

The transition to online platforms has bestowed upon companies several advantages, including expedited publication, continuous availability, and augmented reach through keyword-based searches [Eh11]. Survey results have indicated a shift in preference toward online portals over print media, with 74% of companies utilizing job boards and only 30% still relying on print media⁵.

Job advertisements typically follow a standardized format, which includes an introduction, job description and requirements, incentives, and a call to action. This segmentation serves both promotional and selective functions, with the objective of attracting qualified candidates while deterring unsuitable ones [Sc17].

While advertisements tend to exhibit uniformity in design and phrasing, particularly when managed by external agencies, companies can differentiate themselves by emphasizing unique benefits and cultural values as part of their employer branding strategy [LH03].

Preliminary theoretical considerations yield assumptions that function as points of reference for the subsequent exploratory data analysis.

Assumption 1: In a competitive environment, particularly within bottleneck sectors, companies strive to differentiate themselves from their competitors. Job advertisements serve as a primary point of reference for potential employees, offering an opportunity to capture their attention. Consequently, companies endeavor to distinguish themselves in this context, particularly in terms of the additional incentives they can offer. They achieve this by employing diverse language.

Assumption 2: The ongoing impact of social megatrends and events, such as the global pandemic, has been a significant factor in reshaping the contemporary workplace. The scarcity of skilled labor has been a persistent challenge in recent years, and it is plausible to hypothesize that disparities in the array of supplementary services will become more pronounced during the period from 2015 to 2022.

⁵ See <https://de.statista.com/statistik/daten/studie/682313/umfrage/umfrage-unter-fuehrungskraeften-zur-nutzung-verschiedener-rekrutierungskanaele/>.

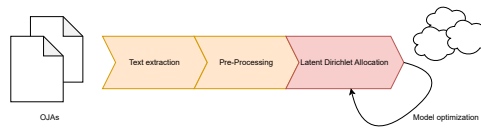


Fig. 1: Workflow

3.3 Topic Modeling

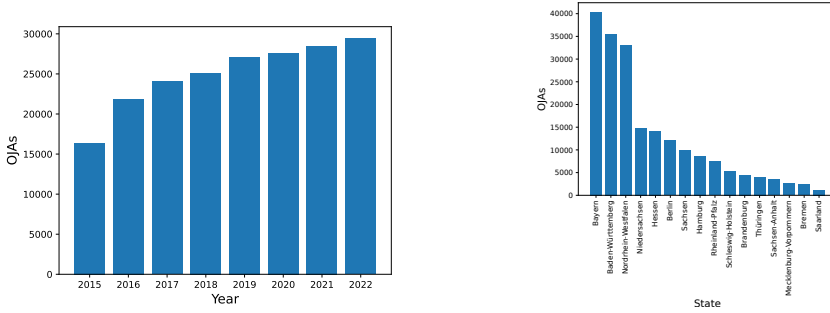
For this empirical study, an inductive-quantitative text analysis approach was selected and implemented through an automated topic modeling algorithm. See Figure 1 for an illustration.

The application of topic modeling, a method originally developed in the field of computer science, has gained traction in interdisciplinary applications within the social and digital humanities for the analysis of large datasets [GH23, p. 530]. The primary advantage of this method lies in its ability to process extensive data computationally, thereby surpassing the limitations of manual annotation [DNB13, p. 577]. In contrast to conventional, manual text analysis that is constrained to specific case studies, topic modeling facilitates systematic processing and adaptable scaling based on the availability of computational resources [A122, p. 259]. It is categorized as an unsupervised machine learning (UML) method, wherein the algorithm autonomously discerns patterns and clusters in data without the need for predefined categorizations or theoretical assumptions [Ho22, p. 161]. This approach contrasts with supervised machine learning (SML), which necessitates targeted training based on researcher-specified inputs [Ho22, 160f.].

In topic modeling, latent thematic structures within texts are inferred without prior knowledge of overarching document topics [BNJ03, p. 8]. The method meets four essential criteria for rigorous text analysis: reproducibility, scalability for large datasets, inductive exploration, and topic-based contextualization [DNB13, p. 577].

This study employs Latent Dirichlet Allocation (LDA), a topic modeling algorithm that has been widely adopted. LDA is based on the assumption that latent topics precede document creation and are represented variably within documents [BNJ03, p. 78]. In this context, topics are defined as mathematical distributions over a fixed vocabulary, with hidden structures revealed through word co-occurrence patterns [BNJ03; GH23]. LDA's output comprises a list of topics with associated word probabilities and the proportional distribution of these topics across documents [GH23, p. 531].

The algorithm strives to achieve a balance between two objectives: the allocation of a limited number of significant words with high probabilities to each topic, while minimizing the overlap between topics and their associated words [DNB13, p. 578]. It operates under the assumption of a "bag-of-words," disregarding word order and syntax to emphasize overarching patterns [BNJ03, p. 82].



(a). Frequency of online job advertisements by year (b). Frequency of online job advertisements by state

Fig. 2: Frequency of online job advertisements

Post hoc, tokens, representing individual words or word forms, are analyzed by the researcher to assign meaning to the topics [GH23, p. 533]. The interpretative process adds semantic relevance to the computational output. Ultimately, LDA functions as a tool, shaped by parameter adjustments and tailored to specific research questions [DNB13, p. 582].

The study's workflow commences with dataset preparation, incorporating linguistic pre-processing to eliminate noise (e.g., punctuation). LDA is then conducted iteratively, with adjustments to parameters and preprocessing steps (hyperparameter tuning). The resulting topics and their token distributions are analyzed to identify dominant topics within documents, enabling a temporal comparison of their prevalence from 2015 to 2022.

4 Data

The data set of OJAs covers a range from 05.01.2015 - 31.12.2022, whereby not every single date in between is shown and several advertisements may have been published on the same date. Figure 2a show the exact distribution of cases per year. It can be seen that more and more job advertisements were scraped in the years after 2015, so that the minimum number of 16,376 is to be found in 2015, while the maximum of this data set is in 2022 with a number of 29,423.

A total of 9,201 different websites are represented in the final data set (in the “source_website” data set), on which the job advertisements were posted online as the original source. Of these, 4,883 (approx. 53%) are represented with only one job advertisement per website. 1.8% of the sources are represented with at least 100 job advertisements or more and less than one percent with 1,000 or more. The sources from which the most job advertisements originate are the websites “arbeitsagentur.de” (26,635 cases), “xing.com” (19,552 cases) and “stepstone.de” (16,627 cases), followed by “jobboerse.arbeitsagentur.de” (7,812 cases), ‘glassdoor.de’ (7,418 cases) and ‘monster.de’ (4,022 cases) as well as ‘careerjet.de’ (3,832

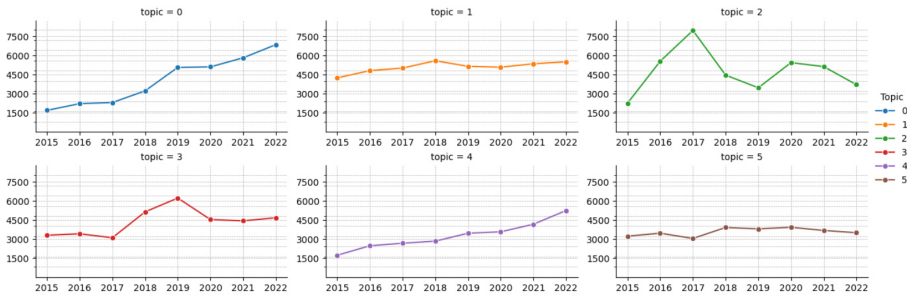


Fig. 3: Distribution of the dominant topics over time 2015-2022

cases), ‘jobfinder.de’ (3,166 cases), ‘bild.de’ (3,117 cases) and ‘neuvoo.de’ (2,774 cases). This shows that 31% of the job advertisements in the final data set were published by the first three websites and 47%, i.e. almost half, by the ten sources listed here.

For around 20% of the dataset, jobs were searched for the Bavaria region (40,393 times), followed by 18% for the Baden-Württemberg region (35,468 times) and 17% for the North Rhine-Westphalia region (33,082 times). The other federal states, on the other hand, fell sharply (see Figure 2b).

5 Experimental Results

First, the topics of the final model were labeled using the “eyeballing” method. This involves determining a topic based on its top tokens (see Table 1) and the associated probability distributions. In a further step, the documents were assigned their respective most dominant topic, broken down by frequency per year and displayed over time (see Figure 3). The following labels were selected for the six topics:

- 0: Social responsibility
- 1: Dynamic workplace
- 2: Working time models
- 3: Modern company
- 4: Typical job advertisement
- 5: Working atmosphere

It is noticeable that many tokens appear several times as top words in the six topics (see Table 2), which is why it is more difficult to use them to determine the main topics. These tokens differ only in terms of their weighting (probability value) in the respective topic. The tokens that occur in all six topics include verguetung, team, attraktiv and unbefristet (see

Topic 0 “Social responsibility”	Topic 1 “Dynamic workplace”	Topic 2 “Working time models”
betrieblich arbeitszeit flexibel verguetung altersvorsorge tarifvertrag attraktiv abwechslungsreich arbeitsumfeld mensch eignung frau team leistung homeoffice unbefristet persoenlich offen schwerbehindert fachlich	verguetung team abwechslungsreich arbeitszeit attraktiv fuehrungsverantwortung leistungsgerecht unbefristet offen tarifvertrag betrieblich interessant dynamisch flexibel modern einarbeitung sicher arbeitsumfeld arbeitsverhaeltnis vollzeit	tarifvertrag verguetung konditionen zusatzleistungen arbeitsverhaeltnis vollzeit offen arbeitszeit unbefristet teilzeit wochenstund dienst oeffentlich entgeltgruppe befristet team interessant attraktiv betrieblich fuehrungsverantwortung
Topic 3 “Modern company”	Topic 4 “Typical job advertisement”	Topic 5 “Working atmosphere”
team abwechslungsreich interessant attraktiv unbefristet verguetung einarbeitung hierarchie flach leistungsgerecht wachsend spannend modern flexibel sicher weiterbildung arbeitsumfeld engagiert kurz verantwortungsvoll	tarifvertrag arbeitszeit verguetung entgeltgruppe vollzeit arbeitsverhaeltnis betrieblich flexibel einarbeitung offen modern unbefristet attraktiv konditionen dienst team befristet sozialleistung interessant oeffentlich	team interessant abwechslungsreich verguetung leistungsgerecht bezahlung flexibel betrieblich angenehm attraktiv arbeitszeit engagiert unbefristet arbeitsumfeld modern betriebsklima verantwortungsvoll motiviert projekt einarbeitung

Tab. 1: Topic modeling benefits with top tokens (highlighting unique tokens)

Benefits	Tokens	Topics	Number
Monetary benefits	altersvorsorge	0	1
	sozialleistung	4	1
	bezahlung	5	1
	entgeltgruppe	2,4	2
	tarifvertrag	0,1,2,4	4
	verguetung	0,1,2,3,4,5	6
Career Goal Opportunities Benefits	weiterbildung	3	1
	fuehrungsverantwortung	1,2	2
Work-Life-Benefits	homeoffice	0	1
	teilzeit	2	1
	wochenstund	2	1
	hierarchie	3	1
	betriebsklima	5	1
	vollzeit	1,2,4	3
	arbeitsumfeld	0,1,3,5	4
	einarbeitung	1,3,4,5	4
	arbeitszeit	0,1,2,4,5	5
	team	0,1,2,3,4,5	6

Tab. 2: Selected categorization of tokens and number of tokens in topics

Table 2). This indicates that the job advertisement corpus is a very homogeneous corpus in which the thematic representation of benefits overlaps. This is confirmed by the fact that models with a higher number of topics (and thus the chance of a more accurate fine-tuning between them) are also affected by an overlap of the tokens.

In addition, the weighting of the corpus was adjusted by the TF-IDF factor in a further test run and words that occur in more than 70% of the documents were removed in order to check the possibility of greater selectivity between the topics. However, many tokens were also found in this run that made up the top words of several topics. For this reason, no further adjustment was made to the TF-IDF value and no further words were removed for the final model. These two alternatives show that the overlap of tokens can accordingly be assumed to be a characteristic of the corpus. It is inherent to the job advertisement corpus that the presentation of benefits in terms of monetary benefits (such as remuneration) career goal opportunities (such as further education) and work life benefits (such as home office), supported by the use of positive adjectives such as attractive, are often expressed together in the “we offer” section in a limited space. Accordingly, some tokens, but not the topics, can be clearly assigned to these benefit categories (see Table 2).

The tokens remuneration and payment can be counted among the monetary, instrumental factors. The tokens retirement benefits and social benefits can also be seen here as additional monetary benefits, although they also cover a social component. The other top tokens,

which fall into instrumental, monetary categories, are tokens that describe the type of employment contract (collective agreement, pay grade) in more detail. The benefits in the area of CGO are assigned to further training and development opportunities (further training, management responsibility). In the area of work-life benefits, the top words include those that can be assigned to the categories workplace (home office, working atmosphere, working environment), interpersonal cooperation (induction, team, hierarchy) and working hours (part-time, weekly, full-time, working hours). The adjectives used to characterize the benefits in more detail and present them positively can be divided into the following categories: Work environment (dynamic, growing, exciting, pleasant, motivated, committed, varied, open, modern, flexible, interesting, attractive), employment contract (fixed-term, open-ended) and other job description (personal, professional, flat, short, public, performance-related, company). Further tokens were assigned to the “Other” category: human, aptitude, woman, performance, severely disabled, additional benefits, project, conditions, service and employment relationship.

For Topic 0: Social responsibility, the model assigned as top 20 tokens those shown in the first column of Table 2. Tokens that can only be found in the respective topic and in no other topic were highlighted in bold, as they are suitable as points of reference for differentiating the topics from one another due to the other homogeneity in the corpus.

The generally positive use of language in job advertisements is also made clear by the use of evaluative adjectives (interesting, attractive, exciting, pleasant). In some cases, they are necessary in order to be able to classify other tokens correctly (e.g. “flat hierarchies”). However, negations, which are intended to indicate that something is not required or not applicable, cannot be recognized directly using the bag-of-words approach. For example, the token *fuehrungsverantwortung* (topics 1 and 2) can be misunderstood, in the sense that leadership responsibility must be assumed, leadership responsibility is promised (for example after further training measures), but also that no leadership responsibility must be assumed for the performance of the job.

The slight lack of selectivity can also be seen as a challenge for comparability over the seven-year period (2015-2022), as the diversity within the topics means that it is not always possible to clearly state which part of the topic is responsible for an increase in frequency in the documents. For example, due to the same probability distribution for Topic 0 (see Figure 3, top left graph), it cannot be clearly stated whether the increased offer of home office has led to an increase in the number of topics or the debate on equal opportunities for women and severely disabled people (e.g. due to legal regulations). This means that the second assumption cannot be confirmed for this research project either, as although changes can be seen over the years, their cause cannot be conclusively determined. One reason for this may also be the small range of years. An investigation over an even longer period (for example, twenty years instead of seven) could possibly reveal differences in the application for jobs even more clearly through topics.

In comparison with Topic 0, Topic 1: Dynamic workplace, has only one token that is

not contained in any other topic: dynamic. For this reason, the topic Dynamic workplace was designated here. The tokens "varied"(represented in four topics in total) and interesting"(represented in five topics in total) as well as "working environment"(represented in four topics in total) (see Table 4) could indicate a similar direction and represent the advertised workplaces primarily as those in which the same processes do not always take place. A gradual rise in the number of documents pertaining to the dominant topic 1 was observed, with a slight decline occurring after a peak in 2018, and the total only reaching the 2018 level again in 2022. However, the trend is not sufficiently clear to draw definitive conclusions about the significance of the topic (see Figure 3, top middle graph).

Topic 2: Working time models, is composed of tokens that, in addition to the first tokens, which are primarily associated with monetary benefits and have the highest probability values compared to their existence in other topics (collective agreement, remuneration, conditions), represent different working time models. For instance, the tokens part-time and weekly hours, which are unique to this topic, delineate the diverse options for structuring the employment relationship in terms of time, in addition to full-time, permanent, and fixed-term positions.

The topic also contains the tokens öffentlich and dienst, which suggests that more jobs were being sought in the public sector and were therefore included as a topic theme. It is noteworthy that there is a sharp increase over time up to 2017, which then falls until 2019, rises This is followed by a decline in 2020 and another rise in 2022 (see Figure 3, top right graph). Given that the focus of this topic is likely on the particulars of job advertisements, providing information on employment contract conditions and organizational work schedules, the fluctuations cannot be readily interpreted through a simple examination of the time course.

In topic 3: Modern Company, the label was selected, as many tokens could indicate self-portrayals of companies that want to emphasize their progressiveness. For instance, the tokens flat and hierarchy, which are unique to this topic, suggest the promotion of a less rigid employee structure characterized by flat hierarchies. Additionally, the tokens team and varied are prioritized, indicating a potential emphasis on teamwork and engaging activities, as also conveyed by the token exciting. The tokens growing and training could also reflect innovation and future orientation, which could be important to young, modern companies.

A close examination of Figure 3 (bottom left graph) reveals a marked increase in the frequency of Topic 3 from 2017 to 2019, followed by a slight decline until 2022.

The fourth topic (4: Conventional Job Advertisements) pertains to a conventional type of job advertisement, to which a single token was allocated. This topic was unique in that it was not assigned to any other topic (social benefit). However, given the heterogeneity inherent in the term, it was deemed non-pertinent for the designation of the topic. As this topic did not demonstrate any distinguishing characteristics, it was classified as a typical job advertisement.

As was the case in Topic 2, the tokens collective agreement, service, public, and pay group

indicate that job advertisements in the public sector were increasingly assigned to this topic. Over time, a continuous increase can be seen from 2015-2022 (see Figure 3, bottom middle graph).

The topic 5: Working Atmosphere, designated by the tokens pleasant, motivated, and working atmosphere, pertains to the workplace environment, which is associated with the WLB. Although the token project is distinctive within this topic, it is primarily employed to describe the method of work (i.e., project work). It is noteworthy that this topic also encompasses the token payment, in addition to the token remuneration. A modest rise in the topic's prevalence is evident from 2017 to 2018, followed by a slight decline from 2020 to 2022, as depicted in Figure 3 (bottom right graph).

The findings indicate that a diversified language design, as outlined in assumption 1, does not apply to the data set under examination. Despite the linguistic preprocessing and the model parameter decisions that enabled the extraction of topics containing vocabulary relevant to the description of ancillary services or benefits, these topics frequently contain numerous tokens that also occur in several other topics, thereby diminishing their distinctiveness. The positive use of language in job advertisements is further highlighted by the employment of evaluative adjectives such as interesting, attractive, exciting, and pleasant.

6 Discussion and Outlook

Regarding the first of the main research question, this study has shown that it is possible to utilize topic-modeling approaches to analyze incentive structures in OJAs. Their changes over time though can be shown in the form of trends in the six topics that were extracted. But the study also found that this period of seven years (2015-2022) is too short to assess causes for these trends. The second question was whether it would be possible to embed the results of computational methods within social sciences theory, in particular signaling theory. This approach builds on previous research on the topic of job advertisements and the theoretical background of employer branding and signaling and the study shows how these results and methods can be fruitfully combined.

As a topic modeling approach, LDA offers a number of advantages, especially when it comes to processing large volumes of data that cannot be captured manually. For example, by processing 200,000 documents, it can be assumed that the model recognizes patterns that cannot be identified individually by randomly evaluating individual documents. In the case of manual classification, the documents may be classified in the same way on the basis of unique keywords (if both documents deal with flexible working hours, for example). However, the topic model is able to assign a different weighting, which may only make sense on a more abstract level (document 1 deals with flexible working hours to 30%, document 2 to 80%) [cf. Al22, 271, own example].

In their research project, Zhao; Jiang; Weng [ZJW11] developed their own topic model, as they suspected that the processing of short documents such as tweets might be less efficient

using an LDA algorithm [ZJW11, p. 341]. In this research project, the average number of characters per document was 500 (and thus has more characters than an average tweet), but this number was reduced to an average of 32 characters per document after linguistic preprocessing. This is due to the individual decision within the linguistic pre-processing, which could be adjusted accordingly in order to generate different results.

The medium of the online job advertisement offers companies in their function as employers the opportunity to send their attractiveness features as signals to the public and potential employees. Even if these have a greater scope for design than print media, they are still strongly bound to their text convention in their formal structure. On the one hand, this is an advantage for the processing of unstructured data, as the thematic focus of the benefits is bundled through the formal structure and can be consciously formulated. On the other hand, this can also result in restrictions. The display could be influenced by a character restriction on the website or other predefined structures such as the use of keywords and/or ellipses.

One reason for concise wording could also be that the shortage or lack of skilled workers in recent years has not yet led to companies feeling compelled to present their benefits in more detail and with greater differentiation than their competitors. The short, concise language formulation in job advertisements and the possible neglect of other grammatical structures are suitable for examining the meaningfulness of the content with a bag-of-words format of the LDA. At the same time, the lack of depth, which can be reinforced by linguistic pre-processing, means that no clear selectivity can be achieved in the topics. The corpus size of 200,000 documents can only partially compensate for effects such as the average number of characters of 32 per document, as the present study has shown.

This analysis has broadened the perspective of the investigation of employer attractiveness and benefits, as it was carried out directly on the basis of process-produced data and examined the current state of the employer side.

However, it is not possible to derive any consequences for the labor market or explain causes for the formation of topics. The 2015-2022 period examined is also too narrow to be able to assume the influence of any social events and megatrends on the formulation of benefits on the employer side. This would require a wider range of years, preferably several decades, to be able to assess these developments, especially with regard to the influence of any generational differences in employer signaling.

As an initial analysis, topic modeling offers the opportunity to explore the structures behind a large data set. For the consideration of benefits in job advertisements, this analysis made it possible to determine the importance of individual words and their assignment to the benefit categories and to create relatively broad topics that can say something about the various signaling effects of employer branding. However, further analyses beyond topic modeling are required in order to be able to make statements beyond the current situation.

References

- [Al22] Althage, M.: Potentiale und Grenzen der Topic-Modellierung mit Latent-Dirichlet Allocation für die Digital History. In (Döring, K. et al., eds.): Digital History. Konzepte, Methoden und Kritiken Digitaler Geschichtswissenschaft. Berlin, Boston, pp. 255–277, 2022.
- [BNJ03] Blei, D. M.; Ng, A. Y.; Jordan, M. I.: Latent dirichlet allocation. *Journal of machine Learning research* 3/Jan, pp. 993–1022, 2003.
- [BS22] Binnewitt, J.; Schnepf, T.: Ökologische Nachhaltigkeit als Attraktivitätsdimension der Berufsausbildung–Extraktion von Aspekten der “Green Economy” aus Online-Stellenanzeigen./, 2022.
- [Co11] Connelly, B. L. et al.: Signaling theory: A review and assessment. *Journal of management* 37/1, pp. 39–67, 2011.
- [CT01] Cable, D. M.; Turban, D. B.: Establishing the dimensions, sources and value of job seekers’ employer knowledge during recruitment. In: *Research in personnel and human resources management*. Emerald Group Publishing Limited, pp. 115–163, 2001.
- [DKD17] Dabirian, A.; Kietzmann, J.; Diba, H.: A great place to work!? Understanding crowdsourced employer branding. *Business horizons* 60/2, pp. 197–205, 2017.
- [DNB13] DiMaggio, P.; Nag, M.; Blei, D.: Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics* 41/6, pp. 570–606, 2013.
- [Dö24] Dörpinghaus, J. et al.: Understanding Informatics in Continuing Vocational Education and Training Data in Germany. *ACM Transactions on Computing Education*/, 2024.
- [DWB22] Dörpinghaus, J.; Weil, V.; Binnewitt, J.: Analyzing longitudinal Data in Knowledge Graphs utilizing shrinking pseudo-triangles. In: *2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS)*. IEEE, pp. 323–327, 2022.
- [Eh11] Ehrenheim, A.: *Das Textdesign der Stellenanzeige: linguistisch und interdisziplinär*. Peter Lang, 2011.
- [Er23] Ernst, F. et al.: Competences for Digital Transformation in Companies–An Analysis of Job Advertisements in Germany. In: *International Conference on Digital Economy*. Springer, pp. 81–96, 2023.
- [FM17] Firfiray, S.; Mayo, M.: The lure of work-life benefits: Perceived person-organization fit as a mechanism explaining job seeker attraction to organizations. *Human Resource Management* 56/4, pp. 629–649, 2017.
- [Ga07] Gansel, C.: Argumentationsstrategie als “Textdesign” in Stellenangeboten. In: *Textdesign und Textwirkung in der massenmedialen Kommunikation*. Pp. 291–305, 2007.

- [GH23] Gillings, M.; Hardie, A.: The interpretation of topic models for scholarly analysis: An evaluation and critique of current practice. *Digital Scholarship in the Humanities* 38/2, pp. 530–543, 2023.
- [Gn22] Gnehm, A.-s. et al.: Fine-grained extraction and classification of skill requirements in German-speaking job ads. In: *Association for Computational Linguistics*, 2022.
- [Ho22] Hodel, T.: Supervised and Unsupervised: Approaches to Machine Learning for Textual Entities. In (Jaillant, L., ed.): *Archives, Access and Artificial Intelligence. Working with Born-Digital and Digitized Archival Collections*. Bielefeld, pp. 157–177, 2022.
- [HR24] Herold, M.; Roedenbeck, M.: Cultural values and the PO fit: comparative NLP analysis of German online job advertisements. In: *Evidence-based HRM: a Global Forum for Empirical Scholarship*. Emerald Publishing Limited, 2024.
- [Je20] Jedrzejczyk, A.: *Internationales employer branding: eine Analyse von Stellenangeboten amerikanischer Unternehmen in Polen und Deutschland*. Springer-Verlag, 2020.
- [LE17] Luttermann, K.; Engberg, J.: Kulturkontrastive deutsch-dänische Textanalyse von sprachlichen Handlungen in juristischen Stellenanzeigen. In (Nielsen, M.; Lévy-Tödter, M.; Luttermann, K., eds.): *Stellenanzeigen als Instrument des Employer Branding in Europa : Interdisziplinäre und kontrastive Perspektiven*. Springer Fachmedien Wiesbaden, Wiesbaden, pp. 107–131, 2017.
- [LH03] Lievens, F.; Highhouse, S.: The relation of instrumental and symbolic attributes to a company's attractiveness as an employer. *Personnel psychology* 56/1, pp. 75–102, 2003.
- [NLL17] Nielsen, M.; Luttermann, K.; Lévy-Tödter, M.: Die Stellenanzeige als Instrument des Employer Branding in Europa – eine Einführung. In (Nielsen, M.; Lévy-Tödter, M.; Luttermann, K., eds.): *Stellenanzeigen als Instrument des Employer Branding in Europa : Interdisziplinäre und kontrastive Perspektiven*. Springer Fachmedien Wiesbaden, Wiesbaden, pp. 11–27, 2017.
- [PLM21] Pernkopf, K.; Latzke, M.; Mayrhofer, W.: Effects of mixed signals on employer attractiveness: A mixed-method study based on signalling and convention theory. *Human Resource Management Journal* 31/2, pp. 392–413, 2021.
- [Sc17] Schmidt, C. M.: Kulturbedingte Konzeptualisierung in Stellenanzeigen. In (Nielsen, M.; Lévy-Tödter, M.; Luttermann, K., eds.): *Stellenanzeigen als Instrument des Employer Branding in Europa : Interdisziplinäre und kontrastive Perspektiven*. Springer Fachmedien Wiesbaden, Wiesbaden, pp. 181–200, 2017.
- [Sc24] Schnepf, T. et al.: *Emphasizing the laborious side of machine learning: annotation guidelines and how to develop them.*, 2024.
- [Sp73] Spence, M.: Job market signaling. *Quarterly Journal of Economics* 87/, pp. 354–374, 1973.

- [Sp74] Spence, M.: *Market Signaling: Informational Transfer in Hiring and Related Screening Processes*. Harvard University Press, 1974.
- [St17] Stumpf, M.: *Employer Branding versus Consumer Branding – (Stellen-)Anzeigen im Vergleich*. In (Nielsen, M.; Lévy-Tödter, M.; Luttermann, K., eds.): *Stellenanzeigen als Instrument des Employer Branding in Europa : Interdisziplinäre und kontrastive Perspektiven*. Springer Fachmedien Wiesbaden, Wiesbaden, pp. 81–103, 2017.
- [St19] Stohr, D. C.: *Die beruflichen Anforderungen der Digitalisierung hinsichtlich formaler, physischer und kompetenzspezifischer Aspekte: eine Analyse von Stellenanzeigen mittels Methoden des Text Minings und Machine Learnings*. Peter Lang, 2019.
- [WB18] Wolf, S.; Berg: *Signaling Family Firm Identity*. Springer, 2018.
- [ZJW11] Zhao, W. X.; Jiang, J.; Weng, J.: *Comparing Twitter and Traditional Media using Topic Models*. In (Clough, P.; Foley, C.; Gurrin, C., eds.): *Advances in Information Retrieval*. Pp. 338–349, 2011.