Proceedings of the 18th Conference on Computer Science and Intelligence Systems pp. 463–470
DOI: 10.15439/2023F6694
ISSN 2300-5963 ACSIS, Vol. 35

Classifying Industrial Sectors from German Textual Data with a Domain Adapted Transformer

Richard Fechner*†, Jens Dörpinghaus*‡, Anja Firll*
* Federal Institute for Vocational Education and Training (BIBB), Bonn, Germany,
Correspondence: richard.fechner.nr@gmail.com,
jens.doerpinghaus@bibb.de, https://orcid.org/0000-0003-0245-7752, anja.firll@bibb.de
† University of Tübingen, Germany
‡ University Koblenz, Koblenz, Germany

***Abstract*—For economics and sociological research, lists of industries and their branches are widely used in research to categorize data and get an overview on different types of industries. However, many different taxonomies and ordering schema exist, due to different research focus but also due to different national scenarios and interests. In this paper, we will focus without loss of generality on regional data from Germany. Manual annotation of textual data is time-consuming and tedious, naturally giving rise to our initial research question, also highly inspired by questions from computational social sciences: How can we automatically categorize textual data, e.g. job advertisements or business profiles, by industrial sectors? We will present an approach towards classification using a pre-trained domain-adapted Transformer model. We find that domain-adapted models generalize better and outperform state of the art non domain-adapted Transformer models on Out-Of-Distribution data. Additionally, we open source two novel data-sets mapping textual data to WZ2008 sections and divisions, enabling further research.**

I. Introduction

FOR economics and sociological research, lists of industries and their branches are widely used in research to categorize data and get an overview on different types of industries. However, many different taxonomies and ordering schema exist, due to different research focus but also due to different national scenarios and interests.

The manual annotation of a diverse set of textual data may not only require an equally diverse set of human experts, but also homogeneity in the ruling of annotation with respect to the underlying taxonomies thereof. Additionally, the process of manual annotation can be time consuming and cumbersome, requiring constant calibration of the annotators ruling. Naturally, one may desire a method to automatically annotate a diverse set of textual data. In this paper, we will focus without loss of generality on annotating German textual data with their respective WZ2008 key (a multi-class classification problem) and provide further details in Section III.

The applications of automated industrial sector recognition are many and varied:

- The categorization of companies. Here, the most important question is on which data a classification should operate. In this paper we will focus on textual data, but other data (e.g. economic data) are also available and could help to improve automated methods.
- The categorization of advertisements, e.g. job advertisements or training advertisements. However, the main question here is whether we want to classify the job position (e.g. a miner under "Mining and Quarrying") or the occupation being recruited. Obviously the two approaches are not interchangeable.
- However, we can also apply this to other textual data: Which industries are mentioned in political speeches or in newspapers? Approaches to literature are even more challenging.

Very limited work has been carried out in this field as we will discuss in the next section. According to our knowledge, no work on German texts has been carried out. Data on companies is usually collected and sold by commercial providers like statista.

This paper is divided into eight 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 section six, we discuss our findings and give a detailed interpretation of the results. After we briefly discuss possible bias in the penultimate section, our conclusions and outlook onto further research are drawn in the final section.

The contributions of this paper include the following:

(i) We fine-tune and compare an openly available domain-adapted BERT model with a standard BERT model. We find that the domain-adapted model shows an increased ability to generalize over the vanilla model on Out-Of-Distribution data.
(ii) We evaluate the models on two novel data-sets, one mapping Wikipedia paragraphs to WZ2008 keys, the other mapping job ads to WZ2008 keys. Both data-sets are open sourced for further research [1].
(iii) We discuss shortcomings of our approach and gain insight on how to improve the current methods. We conjecture that a more diverse mixture of training data will

drastically improve a domain-adapted models ability to generalize.

II. Related Work

Very little work has been done in this area. There are several applications for the given research question: For example, Pejic et al. state the need to analyse Industry 4.0 skills, but do not present a generic categorization approach, but rather pre-select job advertisements according to their needs [2]. Chaisricharoen et al. noted the importance of industrial sectors for legal categories. However, their work is limited to industry-standard keywords [3]. For the generic categorization of English texts, some work has been done by McCallum [4] and Kibriya et al. [5]. However, the data and industrial sectors are mainly for marketing purposes and cannot be used in economic and sociological research. Several other works rely on these data-sets, see for example [6], [7], which underlines the general need for publicly available training and evaluation data.

Text mining on labor market data is a widely considered topic. For an automated analysis of labor-market related texts, the situation in German-speaking countries like Germany, Austria and Switzerland is not much different to English-speaking countries: "Catalogs play a valuable role in providing a standardized language for the activities that people perform in the labor market" [8]. However, while these catalogs are widely used for creating and computing statical values, for managing labor market and educational needs or for recommending trainings and jobs, there is no single ground truth. According to Rodrigues et al., one reason for this could be the fact that labor market concepts are modeled by multiple disciplines, each with a different perspective on the labor market [9]. For German texts, in particular job advertisements, Gnehm et al.[10] introduced transfer learning and domain adaptation approaches with jobBERT-de and jobGBERT. This model was also used for the detection of skill requirements in German job advertisements [11], [12].

For regional data, especially in German-speaking countries, industrial sectors are widely used as a basis for economic and labour market research, see for example [13], [14], [15], they are particularly important for future skills and qualifications [16]. Although classification is a key issue for industrial sectors, see [17], little research has been carried out using computational methods. Examples are mainly limited to regional industries [18] or agriculture and green economy [19].

To our knowledge, no work has been done on German texts. Company data are usually collected and sold by commercial providers such as statista. There is also an online guide from the Federal Office of Economic Affairs and Export Control (BAFA) ("Merkblatt Kurzanleitung Wirtschaftszweigklassifikation"[1]), but this is only a short version of the data available from the Federal Statistical Office. Therefore, we will now discuss the available data.

[1]See https://www.bafa.de/SharedDocs/Downloads/DE/Wirtschaft/unb_kurzanleitung_wirtschaftszweigklassifikation.pdf.

III. Data

As discussed above, several classifications of industrial sectors exist. In our case, we rely on the official German statistics using WZ08. We will describe this taxonomy in the first subsection. However, also a rich variety of possible applications exists. Thus, in the next subsection we will describe several textual data for training and evaluation.

A. Classification of industrial sectors

The so-called "Klassifikation der Wirtschaftszweige" (Classification of branches of industry, short: WZ) is used in Germany, in particular for official statistics by the "Statistische Bundesamt" (Federal Statistical Office), to classify economic activities of employers. The most recent version is WZ 2008, making WZ 2003 and 1993 deprecated. It is compatible to the European "Nomenclature statistique des activités économiques dans la Communauté européenne" (NACE) Rev. 2, but adds more detailed data. For more details we refer to [20]. All data is available in English and German at https://www.destatis.de/DE/Methoden/Klassifikationen/Gueter-Wirtschaftsklassifikationen/Downloads/klassifikation-wz-2008-englisch.html. In this text, for the description of examples we usually rely on the official English translation, while the work itself is carried out on German data.

Similar to NACE, WZ 2008 provides several hierarchical levels. A first level describes 21 sections (letters A-U), a second divisions, a third groups, a fourth classes. In contrast to NACE, WZ 2008 adds subgroups as fifth level, which is, however, only added to particular classes. See Figure 1 for an illustration of a particular hierarchy in sector C. Thus, with examples we find the following hierarchical elements:

- Sections (21), A-U, e.g., "B MINING AND QUARRYING"
- Divisions (88), 01-99, e.g., "05 Mining of coal and lignite"
- Groups (272), 01.1-99.0, e.g., "05.1 Mining of hard coal"
- Classes (615), 01.11-99.00, e.g., "05.10 Mining of hard coal"
- Sub-classes (839), 01.11.0-99.00.0, e.g., "05.10.0 Mining of hard coal"

While sectors are very broad and specific, for example A (Agriculture, Forestry and Fishing) and B (Mining and Quarrying), others are not clearly defined at this level, for example S (Other Service Activities). On the other hand, classes and groups often do not differ and the naming of divisions and groups usually does not provide much more information (e.g. 77 "Rental and leasing activities" towards 77.1 "Renting and leasing of motor vehicles"). In addition, a company might well belong to two or even more industrial sectors, e.g. to several manufacturing divisions. However, the official guidelines recommend to label the most dominant sector. Thus, while the taxonomy of industrial sectors is well-defined by WZ08, we rely on external data to train and evaluate our approaches. In addition, we need to discuss on

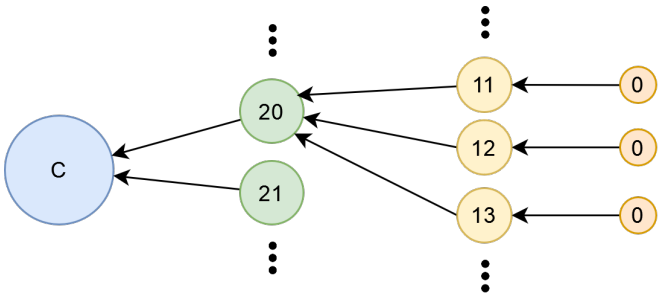

Fig. 1. An example subset of WZ08: Sector C (Manufacturing), division 20 (Manufacture of chemicals and chemical products), group 20.12 and class 20.12.0 (Manufacture of dyes and pigments). Note that for context, other groups (20.11 or 20.13) are displayed as well.

Stichwort	Schlüssel WZ 2008
3D-Druck, Binder Jetting (nur wenn Werkstoff aus keramischem Pulver besteht)	23.44.0
3D-Druck, Binder Jetting (nur wenn Werkstoffpulver aus Kunststoff besteht)	22.29.0
3D-Druck, Laser Sintern	25.50.5
3D-Druck, Multi-Jet Modeling	22.29.0
3D-Druck, Stereolithografie	22.29.0
Aalräuchereien	10.20.0
Abaca, Anbau	01.16.0
Abbau von Bohranlagen (Dienstleistungen im Rahmen der Erdöl- und Erdgasgewinnung)	09.10.0
Abbau von Gerüsten	43.99.1
Abbau von Grauwacke	08.11.0
Abbau von Messeständen	43.32.0
Abbau von Sand (Sandgrube)	08.12.0
Abbauhämmer (handgeführte Druckluftwerkzeuge), Großhandel	46.62.0
Abbauhämmer (handgeführte Druckluftwerkzeuge), Handelsvermittlung	46.14.1
Abbauhämmer (handgeführte Druckluftwerkzeuge), Herstellung	28.24.0
Abbaumaschinen (Bergwerksmaschinen), Herstellung	28.92.1
Abbeizmittel (Zubereitungen zum Abbeizen von Metallen), Großhandel	46.75.0
Abbeizmittel (Zubereitungen zum Abbeizen von Metallen), Handelsvermittlung	46.13.2
Abbeizmittel (Zubereitungen zum Abbeizen von Metallen), Herstellung	20.59.0

Fig. 2. An example subset of keywords or descriptive texts for WZ08 offered by the German Federal Statistical Office.

which level, e.g. sectors or divisions, the categorization can be carried out.

B. Training and evaluation data

1) Official Data: The German Federal Statistical Office (see above) provides a list with 33,945 keywords or descriptive texts of up to 30 words, hence subsequently called snippets, covering all classes in WZ08[2], see Figure 2 for an illustration. It clearly separates between different industries, for example for barrel-locks ("Zylinderschloss") we find entries for retail (47.52.1), whole sale (46.74.1), trade agency (46.15.4) and production (25.72.0). However, this underlines the complexity of this data-set containing not only single keywords but also activities and often even more information, e.g. technical information ("Zylinderschleifereien für Kraftwagen von 3,5 t und weniger"). Thus, we will carefully evaluate how and in which cases we can use this data for training.

The data-set is very imbalanced, as approx. 44% of all snippets map onto a single WZ08-group, namely G 46 Wholesale (excluding trade in motor vehicles) ("Großhandel (ohne Handel mit Kraftfahrzeugen)"). Similarly, about 66% of snippets map onto one of five out of the overall 40 groups.

2) Wikipedia: Covering free text descriptions, we collected and manually annotated 1122 entries of German Wikipedia by industrial sector, division and group. This list contains companies (e.g. "Vereinigte Margarine-Werke Nürnberg"), brands ("Whiskas"), concepts ("Tabak", "Flachglas") and activities or tools ("Weben", "Hammer", "Werkzeug"). Thus, this data highlights how broad industrial sectors are and the questions remains what really characterises them. However, having a cross section of these entities at hand might result in better accuracy.

As a first example, consider 10.9 "Manufacture of prepared animal feeds": here the entries refer to pet food ("food"), but also to brands such as Whiskas. Other industries are even more fuzzy, such as 61.1 "Wired telecommunications activities": Here we collected mainly technical entries (e.g. network connection, mobile phone network), as there are no entries for industries with such a limited focus in the German Wikipedia.

Similarly to the WZ08 snippets, the wikipedia data-set is imbalanced, but towards a different section, namely "C" Manufacturing Industry, containing WZ08-groups like mechanical engineering and so on. About 41 % of the text descriptions belong to this section, please refer to Figure III-B3 for more details.

3) Job Advertisement: As a third data collection, we have 635 manually annotated job advertisements. Here, it is crucial to differentiate, whether the job itself or the company ought to be categorized, as the data may contain information about both domains. This is especially challenging, as the information for and requirements about the position in question may lead to a misclassification in case the inference model cannot distinguish between company-specific or job-specific information. For the annotation process, we decided to categorize companies or businesses, because in most cases job advertisements contain a section with information on them. We excluded advertisements without this profile. In addition, we excluded all advertisements from temporary-employment agencies, because they allow no conclusion about the real company searching for the particular job profile.

We collected data from the BA's official job search, "Jobsuche", which also classifies advertisements by industry, see figure 3. The adverts provide a free text field describing the job and the requirements, see Figure 4 (left). There is also information about the employee, although not all the information seems to be mandatory, see Figure 4 (centre, right). Some companies add extensive promotional texts and descriptions of their profile.

The data-set is also quite imbalanced, although towards completely different sections than the other two data-sets. The most represented section is "C" (Manufacturing Industry) with about 20%, followed by "M" (Provision of freelance, scientific and technical Services) with about 12%.

We make all data-sets openly available at [1].

[2]See https://www.destatis.de/DE/Methoden/Klassifikationen/Gueter-Wirtschaftsklassifikationen/Downloads/klassifikation-wz-2008-3100100089004-aktuell.pdf.

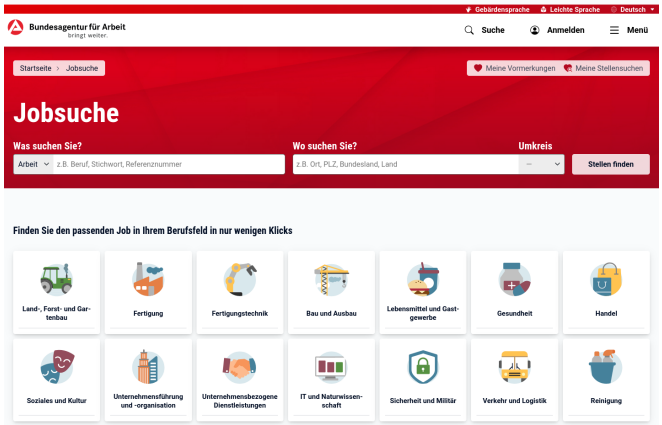

Fig. 3. The landing page of BA "Jobsuche" with job advertisements sorted by industrial sectors (bottom).

C. Manual curation

In this section, we present some information about the manual curation of the data and how we created a gold standard to evaluate the methods presented in this paper. The process was carried out by five domain experts.

During the annotation of various job advertisements from the data-set of the Federal Employment Agency, it was found that certain occupations could not be clearly assigned to one economic sector, but belonged to several industries. Thus, uniqueness was disproved. Furthermore, it can be concluded that occupations can be assigned individually to the economic sectors depending on the job description in the job advertisement and are dependent on this description. Thus, depending on the description, the industry or the frequency of economic categories changes.

The same challenges apply to Wikipedia data: Similar to companies, skills and tools may belong to different industries, leading to either ambiguity in the assignments or missing data. Therefore, we decided to annotate the data only up to the group level, choosing the most typical representations and avoiding generic lemmas.

In the near future, we will provide a more detailed qualitative evaluation of the annotation process, such as the inter-annotator agreement.

IV. Method

A. Preprocessing

The training subset contains data from the 10 most frequent sections. Only classes with more than 100 samples were selected. To reiterate on the issue of imbalanced data, about 44% of the data points of the training data-set map onto one class (G 46 Wholesale). As we didn't want to introduce an intrinsic bias of our classifier towards one class, we had to re-balance the data-set s.t. the samples constitute a uniform distribution over the WZ08-groups. Naturally, one may consider the threshold for samples per class to be the sample size for the lowest represented class (minority class). This approach might work, but limits the effective usage of a large portion of the training data for classes, which are over-represented (majority class). We chose to set a sample threshold of 4000, under-sample the majority classes, i.e. draw from the pool of samples for the respective class without replacement, and over-sample, i.e. sample with replacement from the respective samples, for the minority-classes to arrive at precisely 4000 samples for each of the classes. To optimize the diversity of the data of the minority classes, we made sure that all of the samples originally contained within the subset were present in the over-sampled data-set. This allowed for a balanced training data-set, trading in loss of generalization of the model on minority classes for a gain of generalization on the majority classes. In practice, we observed that for our training data-set the over- and under-sampling yielded very minor performance improvements on the evaluation data-sets.

B. Model Architecture

We used the `spacy` Python framework for Natural Language Processing (NLP) to fine-tune pre-trained transformer models to the task of text classification. Since it's great influx in popularity following the original publication by authors at Google [21], the transformer has arrived as a commonly used Neural Network Architecture. It has become the de-facto-standard for applications in NLP, given it's highly preferable ability to work on sequential data in parallel, making use of today's large amount of available compute resources as well as enabling the processing of even larger data-sets. The breakthrough included the introduction of a so called "Self-Attention-Layer", a Neural Network component (Figure 6) able to introduce the importance of the relationship between words within a sequence into an embedding used for subsequent processing. Each "Attention-Head" within the Self-Attention-Layer learns to attend to different semantic relationships during training, allowing for enough capacity to find crucial structural and semantic information in the data. We used the encoder part of the transformer to create sequence embeddings, which were then fed into a fully connected block, followed by classification head. This allowed us to fine-tune the pre-trained models on the variable length snippets. We fine-tuned two transformer models `bert-base-german-cased` and `agne/jobBERT-de`, evaluating them respectively on the evaluation data-sets described in Section (III-B). The base BERT (Bidirectional Encoder Representations from Transformers) model is a transformer based model trained by authors at Deepset, available at [22] and was pre-trained on German Wikipedia dumps, OpenLegalData dumps and news articles. The jobBERT-de model [23] is based on the previously mentioned base BERT model and adapted to the domain of job advertisements through continued in-domain pre-training on approx. 4 million German-speaking job advertisements from Switzerland from the years of 1990 to 2020.

We trained our models on a V100 GPU for approximately one hour each. We make the training configurations as well

Fig. 4. Several anonymized parts of job advertisements at BA "Jobsuche": A description (left), and two information on the employee.

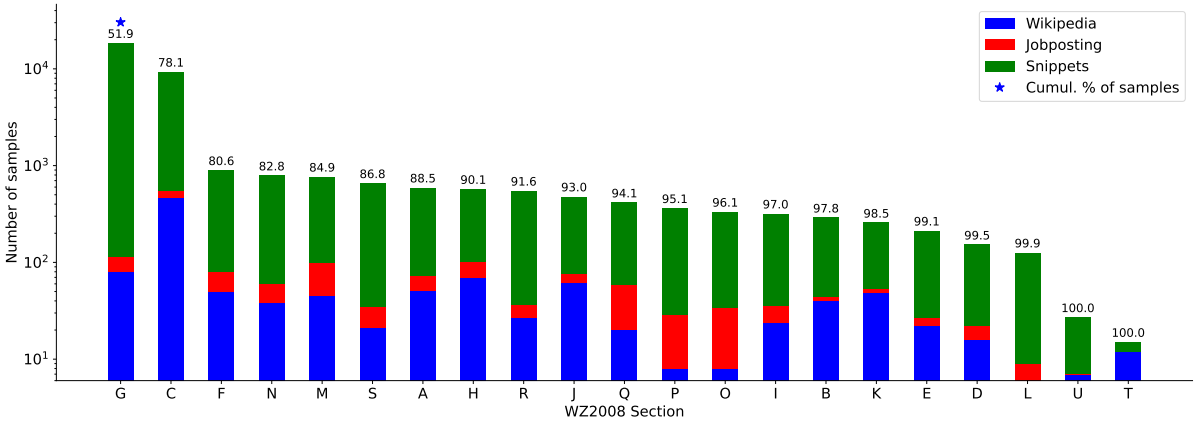

Fig. 5. The distribution of samples training and evaluation data across sections of WZ2008. Data are very unbalanced. Some sections are strongly underrepresented. Training data (green) cover only part of the evaluation data.

as all hyperparameters used in the preprocessing, training and evaluation of our experiments openly available at [1].

V. Evaluation

In the following we present the performance of the two fine-tuned models, for simplicity we will address them by the name of the pre-trained transformer model they are based on, `bert-base-german-cased` (BERT) and `agne/jobBERT-de` (jobBERT). We evaluated both models on three different data-sets. For each pre-trained model, we fine-tuned two classifiers, one to classify Sections and one to classify Divisions. As portrayed in Figure 1, a Sector may contain multiple Divisions. Naturally, the task to classify Divisions is harder, as there are more classes to be classified than in sector-classification.

A. Sections

We can see from Table I, that the BERT model has generally demonstrated a higher capability to capture information from the training set (Snippets) and archives better results on the Snippets evaluation data-set, which comes from the same data distribution as the training set. The domain-adapted jobBERT model although, demonstrates it's ability to generalize better across data distributions, as we can see that the metrics for the non-training data distributions (i) Wikipedia and (ii) Job-postings are increased in comparison to the vanilla BERT model.

TABLE I
(MACRO-) F_1-SCORE, PRECISION, RECALL AND TOP-5-ACCURACY FOR CLASSIFIERS TRAINED FROM BERT (a) AND JOBBERT (b) TO CLASSIFY SECTIONS (19 CLASSES).

Data-set	Wikipedia		Job-postings		Snippets	
Model	(a)	(b)	(a)	(b)	(a)	(b)
F_1	0.62	0.64	0.20	0.22	0.95	0.88
Precision	0.62	0.66	0.29	0.43	0.94	0.87
Recall	0.72	0.71	0.25	0.23	0.97	0.90
Top 5 Accuracy	0.92	**0.95**	0.58	**0.72**	0.99	**0.99**

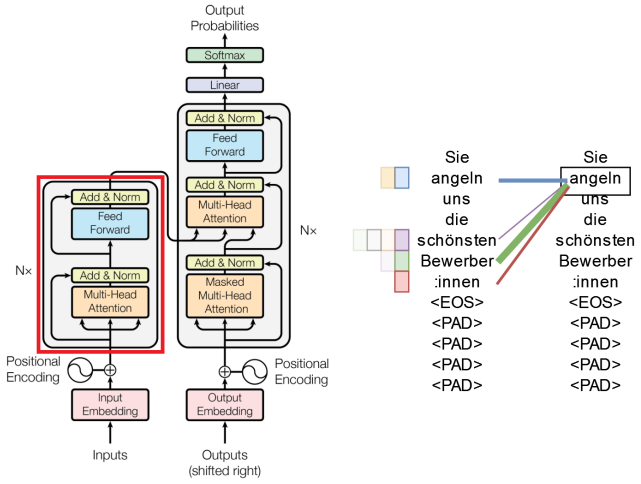

Fig. 6. *Left*: Encoder- (red box) and Decoder-components of the Transformer architecture presented in [21]. For our experiments, the encoder part of the architecture was used.
Right: The Self-Attention mechanism. Different attention-heads (here indicated by different colors) attend to different words in the sequence. Note, that for clarity only the attention for the word "angeln" is displayed. This example highlights the importance of context-sensitive methods. The word "angeln" (fishing) can have multiple meanings in different contexts.

B. Divisions

Our experiments for the classification task on divisions, which partition the sections into subsections, is in contrast to our findings for the section-classification, as we see in Table II that the difference in the predictive capability on the training set data distribution for both classifiers is very similar. The advantages of the domain-adapted jobBERT model on the Job-postings data-set is marginal and on the Wikipedia data-set, the BERT model even outperforms the jobBERT model in terms of the macro-Precision metric. Nonetheless, the jobBERT model still achieves higher top 5 accuracy on the non-training data distributed evaluation data-sets.

TABLE II
(MACRO-) F_1-SCORE, PRECISION, RECALL AND TOP-5-ACCURACY FOR CLASSIFIERS TRAINED FROM BERT (a) AND JOBBERT (b) TO CLASSIFY DIVISIONS (40 CLASSES).

Data-set	Wikipedia		Job-postings		Snippets	
Model	(a)	(b)	(a)	(b)	(a)	(b)
F_1	0.46	0.43	0.11	0.13	0.86	0.86
Precision	0.70	0.51	0.17	0.16	0.87	0.87
Recall	0.45	0.46	0.13	0.14	0.86	0.87
Top 5 Accuracy	0.80	**0.85**	0.41	**0.51**	0.98	**0.98**

VI. DISCUSSION

With our experiments, which we presented in the previous section, we provided empirical evidence, that the domain-adapted transformer model jobBERT generalizes better on non-training data distributions whereas the vanilla BERT model outperforms jobBERT on the training data distribution for a text classification task. We think that the results presented in Table I are comparatively more representative, as the lowered granularity increased the overall training data size. The precision of these results may be improved further by collecting a more diverse data-set, containing more samples from the under-represented minority classes. In our experiments, we counteracted the imbalance of the data-set with oversampling but although this method was mainly introduced to avoid an intrinsic bias of the model and still be able to use most of the training data for the majority classes, it cannot improve predictive capability on the minority classes. If anything, it even lowers the predictive capability for minority classes as our model is prone to over-fitting. An additional concern is, that since the number of tokens the BERT models can process is limited to 512 tokens, some of the valuable information contained in the latter parts of a job-posting or Wikipedia article might be lost, as the underlying `spacy` textcat-model may truncate the input to fit its maximal token length. An approach would be to split the input sequence into sub-sequences that individually conform to the token length constraints of the model, then feed the sub-sequences to the model individually, obtaining multiple classifications from one partitioned sequence. This approach would then introduce the obvious problem, that the model cannot include potentially crucial context across sub-sequences. Another approach would be to make use of a summarization-model or repeated prompting of a Large Language Model (LLM), in order to condense or extract important information. Simple tests, which we've conducted using state of the art LLMs have shown that this approach yields underwhelming summarization or data-extraction performance for our data. However, using more powerful models or different prompting techniques like Chain-Of-Thought [24] or Tree-Of-Thought [25], one might be able to improve upon the simple pre-processing used in this work.

Most importantly though, we would like to bring attention to the dominant problem we are faced with in classifying sectors, namely the problem of mixed domains. As already outlined in the latter parts of subsection III-B, job-postings might contain information on the job itself (e.g. mechanic) *and* the company searching for said talent (e.g. an agricultural firm looking for a mechanic). It is challenging to make the context sensitive distinction between both pieces of information. Staying in the domain of job-postings, text-segmentation could be used to first partition the text into sections (e.g. Address of company, Job description, Requirements etc.), then feeding the segmented text into a subsequent network.

A point of discussion should also be the plausibility of the choice of granularity for the classification, meaning whether it is plausible to attempt to differentiate between Divisions, as for some cases the distinction of inter-section Divisions is challenging even for a human expert. It is for this reason we decided to include the Top-N Accuracy metric into our analysis. We show that even with a straight forward approach like fine-tuning domain-adapted transformer models, we are able to reach 85% Top-5 Accuracy on non-training data

distributions (Wikipedia data-set, Table II), i.e. data from a distribution the model has never seen before and hence has had no chance of adapting to. To this end, our findings should be interpreted as a proof of concept, as we have demonstrated that with simple data-sets and straight forward methods we are able to generalize across different data distributions. Naturally, more work is to be done, as we currently lack big annotated data-sets from different data-distributions (one may consider Twitter data, YouTube descriptions, reddit posts etc.) in order to train a model on a balanced mixture data-set.

VII. Bias

We would like to touch on the topic of bias, which is in part introduced by the fine-tuned models themselves a priori. It is likely, that BERT has seen some collection of job-postings, for which the data distribution is unknown. Similarly, the domain-adapted jobBERT was pre-trained on a corpus of job-postings, for which we also don't know the data-distribution. Additionally, it is crucial to mention, that the job-postings on which jobBERT was pre-trained on are of Swiss origin which introduces additional bias. It remains a topic of discussion, whether a corpus of Swiss job-advertisements suffices as a pre-training data-set, if one is trying to fine-tune for classification onto German industrial sectors. As for the evaluation data-sets, the Wikipedia data-set and the Job-postings data-set may have introduced human bias of unknown form. As outlined in subsection (III-C) we will supplement the open sourced datasets provided at [1] with a detailed qualitative evaluation of the annotation process, such as the inter-annotator agreement.

VIII. Conclusion and Outlook

In this paper, we presented our novel approach to classifying general textual data onto German industry sectors using pre-trained transformer models. In the second sections, we introduced the WZ08, a taxonomy of the German industrial sectors, and subsequently discussed the imbalance of both the training- and evaluation-data. We sourced two novel data-sets (a) Wikipedia articles and (b) Job-postings mapping to WZ08 Divisions, to be included in our analysis and discussed the respective details in section III. In the evaluation, we showed that in spite of the difficult challenge of mixed domains and the imbalance of the data available, the domain-adapted transformer model jobBERT was able to generalize better across different data distributions than the regular BERT model in a text classification task. This hints that even with simple methods like fine-tuning a domain-adapted transformer model, one is able to generalize relatively well across unknown data-distributions given a good mixture of data-sets.

Our initial research question was whether one can automatically categorize textual data, such as job ads or company profiles, by industry. We presented and discussed several approaches and showed that this categorization is possible. However, its quality depends on both training and evaluation data. Thus, it also depends on the application and the research question.

All approaches failed for job advertisements. Here we need to redefine a precise research question and in particular provide more feasible information about what data is available (e.g. which metadata could help to improve the quality) and what the expected result should be.

However, our approach provides a reasonable recall of Wikipedia data. Thus, it could help to recommend and provide suggestions for manual curation and annotation on similar textual data. Further research and quality control could help to improve the model. While the presented approach works and provides meaningful results, it is far from being ready for productive use, but shows the significant impact of research in this area.

The initial research question was difficult not only because of the diversity of data and expected outcomes, but also because of the interdisciplinary nature of the research. The social sciences and the example use cases for labor market research have a different perspective on industrial sectors than, for example, economics. Thus, understanding the correct classification depends not only on the research questions, but also on the perspective of different scientific domains. It is very unlikely that a single generic solution could be developed to cover all these different needs. However, more interdisciplinary exchange could help to clarify and guide computer science research in this area.

Acknowledgments

We would like to thank our colleagues Katharina Bär, Nicolai Bör, Lisa Fournier and Jan-Philipp Schroer for helping us out with data annotation.

References

[1] R. Fechner, D. J. Dörpinghaus, and A. Firll, "FedCSIS 2023 Classifying Industrial Sectors with a Domain Adapted Transformer - Datasets and Configuration files," Jul. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.8192546

[2] M. Pejic-Bach, T. Bertoncel, M. Meško, and Ž. Krstić, "Text mining of industry 4.0 job advertisements," *International journal of information management*, vol. 50, pp. 416–431, 2020.

[3] R. Chaisricharoen, W. Srimaharaj, S. Chaising, and K. Pamanee, "Classification approach for industry standards categorization," in *2022 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)*. IEEE, 2022, pp. 308–313.

[4] A. McCallum, K. Nigam *et al.*, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752, no. 1. Madison, WI, 1998, pp. 41–48.

[5] A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, "Multinomial naive bayes for text categorization revisited," in *AI 2004: Advances in Artificial Intelligence: 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia, December 4-6, 2004. Proceedings 17*. Springer, 2005, pp. 488–499.

[6] H. Hayashi and Q. Zhao, "Quick induction of nntrees for text categorization based on discriminative multiple centroid approach," in *2010 IEEE International Conference on Systems, Man and Cybernetics*. IEEE, 2010, pp. 705–712.

[7] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, 2019.

[8] C. Ospino, "Occupations: Labor market classifications, taxonomies, and ontologies in the 21st century," *Inter-American Development Bank*, 2018.

[9] M. Rodrigues, Fernández-Macías, and Enrique, Sostero, Matteo, "A unified conceptual framework of tasks, skills and competences," Seville, 2021. [Online]. Available: https://joint-research-centre.ec.europa.eu/publications/unified-conceptual-framework-tasks-skills-and-competences_en

[10] A.-S. Gnehm, E. Bühlmann, and S. Clematide, "Evaluation of transfer learning and domain adaptation for analyzing german-speaking job advertisements," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 3892–3901.

[11] A.-S. Gnehm, E. Bühlmann, H. Buchs, and S. Clematide, "Fine-grained extraction and classification of skill requirements in german-speaking job ads." Association for Computational Linguistics, 2022.

[12] J. Büchel, J. Engler, and A. Mertens, "The demand for data skills in german companies: Evidence from online job advertisements," *How to Reconstruct Ukraine? Challenges, Plans and the Role of the EU*, p. 56, 2023.

[13] B. Gehrke, H. Legler, M. Leidmann, and K. Hippe, "Forschungs- und wissensintensive wirtschaftszweige: Produktion, wertschöpfung und beschäftigung in Deutschland sowie qualifikationserfordernisse im europäischen vergleich," Studien zum deutschen Innovationssystem, Tech. Rep., 2009.

[14] N. Gillmann and V. Hassler, "Coronabetroffenheit der wirtschaftszweige in gesamt-und ostdeutschland," *ifo Dresden berichtet*, vol. 27, no. 04, pp. 03–05, 2020.

[15] U. Kies, D. Klein, and A. Schulte, "Cluster wald und holz deutschland: Makroökonomische bedeutung, regionale zentren und strukturwandel der beschäftigung in holzbasierten wirtschaftszweigen," *Cluster in Mitteldeutschland–Strukturen, Potenziale, Förderung*, p. 103, 2012.

[16] V.-P. Niitamo, "Berufs-und qualifikationsanforderungen im ikt-bereich in europa erkennen und messen," *Schmidt, SL; Strietska-Ilina, O.; Dworschak, B*, pp. 194–201, 2005.

[17] J. Hartmann and G. Schütz, "Die klassifizierung der berufe und der wirtschaftszweige im sozio-oekonomischen panel-neuvercodung der daten 1984-2001," SOEP Survey Papers, Tech. Rep., 2017.

[18] M. Titze, M. Brachert, and A. Kubis, "The identification of regional industrial clusters using qualitative input–output analysis (qioa)," *Regional Studies*, vol. 45, no. 1, pp. 89–102, 2011.

[19] U. Kies, T. Mrosek, and A. Schulte, "Spatial analysis of regional industrial clusters in the german forest sector," *International Forestry Review*, vol. 11, no. 1, pp. 38–51, 2009.

[20] Statistisches Bundesamt, "Klassifikation der Wirtschaftszweige," Wiesbaden, 2008. [Online]. Available: https://www.destatis.de/static/DE/dokumente/klassifikation-wz-2008-3100100089004.pdf

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[22] B. Chan, T. Möller, M. Pietsch, and T. Soni. (2019) bert-base-german-cased transformer model. [Online]. Available: https://huggingface.co/bert-base-german-cased

[23] A.-S. Gnehm, E. Bühlmann, and S. Clematide, "Evaluation of transfer learning and domain adaptation for analyzing german-speaking job advertisements," in *Proceedings of the 13th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2022.

[24] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[25] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," *arXiv preprint arXiv:2305.10601*, 2023.