Research Proposal

December 1, 2021

Student Name: Rahkakavee Baskaran

Supervisor: Prof. Dr. Susumu Shikano

Supervisor: JunProf. Dr. Juhi Kulshrestha

1 Introduction

1.1 Motivation

Job titles are key information within the labor market. They are useful for job seekers to find jobs (Marinescu and Wolthoff, 2020). They are an important component of job search engines (Slamet et al., 2018; Javed et al., 2015, 2016) and job recommendation systems (Malherbe et al., 2014). And lastly, they serve as a valuable data source for various analyses, such as job market trend (Martin-Caughey, 2021; Li et al., 2021), job perception (Smith et al., 1989; Boydston and Hirst, 2019) or social science analyses (Martin-Caughey, 2021). However, since job titles are not normalized, it is challenging to structure them in an appropriate way for downstream tasks. Various institutions developed job taxonomies in order to structure and generalize job titles. Established taxonomies are, for example, the "International Standard Classification of Occupation" (ISCO) for the European job market or the "Klassifikation der Berufe 2010" (KLdB) for the German job market (Uter, 2020). Matching job titles from job postings with classes from those taxonomies is inevitable in order to improve job search engines or recommendation systems as well as analyzing the labor market. In Natural Language Processing (NLP), this process of matching is known as text classification.

1.2 Related Work

Text classification, a highly researched area, is the process of classifying text documents or text segments into a set of predefined classes. During the last decades, researches developed a various number of classifiers. As Kowsari et al. (2019) summarize in their survey of classifiers, we can group the approaches mainly into three groups. The first group contains traditional methods like Naive Bayes (NV), Support Vector Machines (SVM), K-nearest neighbors (KNN), Logistic Regressions (LR) or Decision Trees (DT) (Vijayan et al., 2017; Colas and Brazdil, 2006; Kowsari et al., 2019; Sebastiani, 2001). Deep learning methods like Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN), which are currently the most advanced algorithms for NLP, form the second group. The last group consists of ensemble learning techniques like Boosting and Bagging.

Each text classification task presents different challenges. One challenge is that domain-specific problems may arise. There is some work that deals with job classification in the English speaking job market. In terms of classifiers, the corresponding work can be categorised into traditional classifiers or deep learning methods. Zhu et al. (2017) for example, use a KNN classifier in combination with document embedding

as feature selection strategy. Javed et al. (2015) rely on traditional methods as well, by combining a SVM classifier and a KNN classifier for their job recommendation system. In contrast, the approaches of Decorte et al. (2021), Wang et al. (2019) and Neculoiu et al. (2016) are based on Deep Learning methods. From a higher perspective, there is another dividing line between the approaches. As mentioned earlier, job title normalization can be considered as a typical text classification task (Wang et al., 2019; Javed et al., 2015; Zhu et al., 2017). Decorte et al. (2021) and Neculoiu et al. (2016), however, formulate the task as a string representation approach of similar job titles.

Another potential issue is the length of input documents for classification. Job titles are clearly short text with often not more than 50 characters. Short texts suffer from sparseness, few word co-occurrences, missing shared context, noisiness and ambiguity. Traditional methods, however, are based on word frequency, high word co-occurrence and context, which is why they often fail to achieve high accuracy for short texts (Song et al., 2014; Wang et al., 2017b, 2014). In their overview, Song et al. (2014) present three approaches to solve this. First, since short text data often suffers from unlabeled data in the context of online text data, such as Twitter postings, they suggest using semi-supervised approaches. Second, they recommend to use ensemble learning methods, which focus on the sparseness of the data. Third, Song et al. (2014) propose feature dimensionality reduction and extraction of semantic relationship methods. Based on the latter more recent work on short text classification criticizes the use of the "Bag of Word" concept for feature representation as it only reflects the appearance of words in the text. Instead, they represent short texts with semantically similar and conceptual information (Bouaziz et al., 2014; Wang et al., 2014; Chen et al., 2019). Another question concerning the representation of short texts is whether to represent them as dense or sparse vectors. In their comparison between tf-idf/counter vectorizer and the dense vectorizer word2vec and doc2vec, Wang et al. (2017b) conclude that among the classifiers Naive Bayes, Logisitic Regression and SVM, the sparse vectorizers achieve the highest accuracy. Chen et al. (2019), conversely, see limitations in sparse representation as it cannot capture the context information. In their work, they integrate sparse and dense representation into a deep neural network with Knowledge powered Attention, which outperform state-of-art deep learning methods, like CNN, for Chinese short texts. Concerning the classifiers, there is no consensus approach for short text classification. For traditional approaches Wang et al. (2017b)'s results indicate that logistic regression and SVM perform best, while KNN seems to achieve best accuracy in Khamar (2013)'s work. Similar to job title specific work, more recent work prefers deep learning methods, mostly CNN (Chen et al., 2019).

A last challenge of text classification tasks comes with the number of classes. As Li et al. (2004) show in their classification of tissue, multiclass classification is more difficult than binary classification problems. Partly, because most of classification algorithms were designed for binary problems (Aly, 2005). Approaches for multiclassification can be grouped into two types. Binary algorithms can handle multiclassification naturally. This is, for example, the case for Regression, DT, SVM, KNN and NV. The second type is the decomposition of the problem into binary classification tasks (for the different subtypes see Aly (2005)). The literature so far does not have a clear answer to solve multiclassification problems. Different approaches, like

Boosting (Schapire and Singer, 2000) or CNN (Farooq et al., 2017) are applied. It is noticeable, however, that many works use variations of SVM (Guo and Wang, 2015; Tomar and Agarwal, 2015; Tang et al., 2019).

2 Goal of the Thesis

2.1 Problem Statement

The presented work on job classification has several gaps. First, while there is extensive work on job title classification for the English speaking job market, as far as I know, there have not been any classification attempts for the German job market. However, an accurate classification of job titles with the German taxonomy would facilitate several downstream tasks for the German job market. With the KldB 2010, an occupational classification was created for Germany that reflects the current trends in the labor market based on empirical and theoretical foundations. It contains 5 hierarchical levels, each of them reflecting different aspects. Especially level 3, which represents the professionalism of occupations, and level 5, which gives information on the requirements for a job are powerful tools for job market analyses (Paulus and Matthes, 2013). In addition, the KldB 2010 is based on ISCO, which makes it easier to link the two taxonomies. Further, since job portals often use different categories for job titles, a classification according to the standardized KldB 2010 facilitates comparisons across job portals. The strengths of the KldB clearly show that an accurate classification of job titles according to KldB 2010 opens up new possibilities for labor market analyses. Furthermore, the classification could improve job search engines and recommendation systems.

Second, the brevity of the input texts and the large number of classes are the main challenges in job title classification. In the previous approaches to job title classification, the challenges are partly mentioned and considered, but the frameworks are not built on the basis of these challenges. A clear focus on methods from the two problem areas, could help to improve classification.

Finally, most of the work about job title classification suffers from solid databases. Therefore, Decorte et al. (2021), for example, use skills to understand the meaning of job titles and to avoid manually labelling them. Javed et al. (2015) rely on a weakly supervised approach to get enough labelled data. The advantage of classifying for the German job market is that the Federal Employment Agency of Germany provides a data set with job titles and the possibility of linking them with die KldB classes, which offers a huge and powerful training data set which in turn allows for more flexibility in which algorithms are applicable.

2.2 Research Objectives

Based on the research gaps of job classification and the relevance of a classification algorithm for the KldB Taxonomy for the German job market, I define the following research objective:

Develop and implement a classification algorithm based on the challenges of short text and multiclass classification in order to match job titles of German job postings with the Taxonomy KldB 2010. Then, the

algorithm should be evaluated against current state-of-the-art methods.

2.3 Expected Results

The project shall contribute to improve downstream tasks for the German labor market. The goal of this project is to develop a classification algorithm that assigns job titles from job advertisements to the predefined classes of the KldB 2010 Taxonomy with a high precision and recall. In this way, job titles, originally non-standardized pieces of information, are normalized to job classes, which can then be used for analyses, search engines and recommendation systems. The thesis includes recent studies on short text classification and multi class classification, as well as job title classification approaches.

3 Methodology

In order to reach the goals of this Thesis, several tasks have to be fulfilled.

First exploration

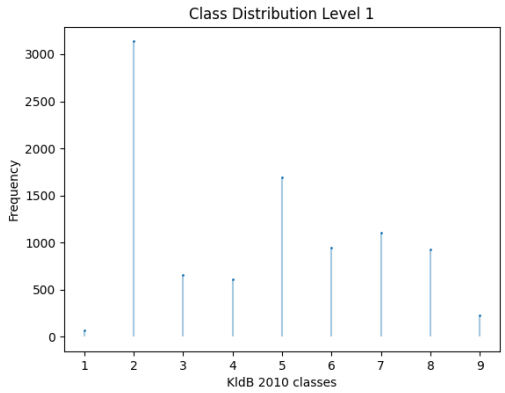
In a first exploration, the data should be analysed with a simple algorithm in order to check how they would perform for the classification task. The analysis consists of three steps. In the first step, I look at the first ten main groups from level 1 of KldB 2010. Then, I apply the algorithm to the level 3 class and lastly to the level 5 classes, thereby extending the number of classes. I choose level 1, level 3 and level 5 since they have the most valuable information of all levels. Results with a preliminary training data set show the following results:

Figure 1 (a), (b) and (c) show the distribution of the classes for each level. In figure 1a we can see that most of the classes are well balanced. The classes for level 3 and level 5 are not well distributed. Figure 1 (b) and Figure 1 (c) show that the subclasses of class 5 and class 7 are strongly unbalanced. In addition some of the subclasses have only few or no training examples, which is not sufficient for training. Besides, there is no example for class zero. Thus, there is a need to rescrape data for certain classes.

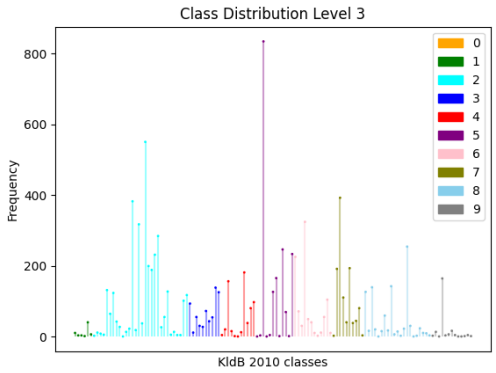
Table 1 to 3 show the evaluation reports with different classification metrics for the levels 1, 3 and 5. For the first exploration, I used a NB classifier, since it is a quite simple, but usually good performing algorithm. Results indicate for level 1 a moderate precision and recall. With increasing number of classes, precision and recall clearly decrease.

Training data

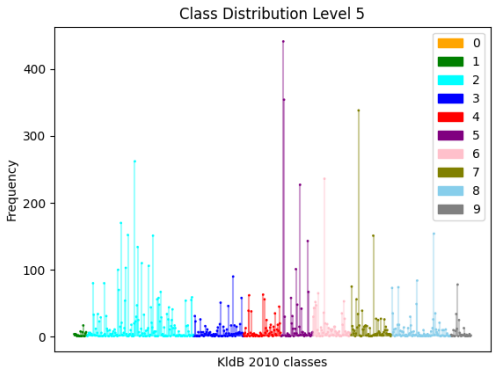
The training data is scraped from the “Bundesagentur für Arbeit” (BA), which provides current job postings. The advantage of this data set is, that it is already structured, with each entry containing the job title and the “Dokumentationskennziffer” (Dkz). The Dkz is a internal key, which can be easily mapped to the KldB



(a)



(b)



(c)

Table 1 Classification Report Level 1

	precision	recall	f1-score	support
1	0.050633	0.727273	0.094675	11
2	0.884615	0.498724	0.637847	784
3	0.328947	0.709220	0.449438	141
4	0.443182	0.726708	0.550588	161
5	0.818681	0.683486	0.745000	436
6	0.529801	0.730594	0.614203	219
7	0.813559	0.491468	0.612766	293
8	0.809091	0.744770	0.775599	239
9	0.336283	0.633333	0.439306	60
macro avg	0.557199	0.660620	0.546603	2344
weighted avg	0.740923	0.611775	0.641521	2344
accuracy			0.611775	2344

Table 2 Classification Report Level 3

	precision	recall	f1-score	support
macro avg	0.330428	0.399629	0.321017	2344
weighted avg	0.544320	0.422782	0.446718	2344
accuracy			0.422782	23144

Table 3 Classification Report Level 5

	precision	recall	f1-score	support
macro avg	0.260878	0.282669	0.240905	2344
weighted avg	0.474158	0.342577	0.367748	2344
accuracy			0.342577	2344

2010 IDs (Paulus and Matthes, 2013). Hence, the training data can be constructed from BAs structured data, together with the KldB 2010 classes and the job title.

Most of the classification approaches demand the data to be balanced. In real world application this is often not possible (Japkowicz, 2000). Especially for level 5 it is difficult to get enough data for all classes. The descriptive analysis of the first exploration clearly shows that the data is unbalanced for level 3 and level 5, having zero observations in some classes. This has to be handled in some way in the classification process itself or the data must be post-scraped for specific classes. Since the BA updates the data daily, the second way should be enough for most of the classes. In order to do so, I will scrape data regularly to have a solid data base.

Baseline Algorithms

The developed algorithm should be compared against the current state-of-the art methods in order to check the improvements. As mentioned in the literature overview, state-of the art methods are Deep Neural Networks, especially CNN. Thus, CNN provides a strong comparison, which is why I will use a simple CNN as the baseline. However, since traditional methods like SVM performed well in some approaches, especially for multi-class handling, it is convenient to also use a traditional approach as baseline. As mentioned in the literature review, often different versions of the SVM algorithm are used. It is therefore a reasonable option to use a basic version of the SVM algorithm.

Developing of own approach

As the literature on short text classification shows, the representation of short texts with “Bag of Words” is not feasible. Instead, including semantically and conceptual information often leads to better results. The KldB 2010 taxonomy includes for each level 5 class a job description and search words, which allows for the inclusion of a knowledge base into the classification process. In a first step, I will create a knowledge base using the job description and search words. I will use the knowledge base for three approaches. In the first

approach, I will test the performance of an entity linker, which requires a knowledge base and might already improve the performance. Evaluation will be checked against the named baseline algorithms using precision and recall.

The second approach follows the method of Wang et al. (2017a) by conceptualizing the short text first and then apply a CNN algorithm.

For the reasons already stated about the baseline algorithms for the SVM classifier, I will also incorporate the knowledge base in an SVM classifier as my third approach. For this, I follow the method suggested by Le and Smola (2006). The advantage of this approach is, that the authors have a specific method for multiclass SVM with knowledge bases.

After implementing and evaluating these approaches, depending on the results, further improvements or other approaches might be necessary, which will be evident during the development process.

4 Timeline

Each implementation and evaluation step involves taking notes of related literature, theory, implementation steps and results.

Date	**Milestone**
21 Nov 2021	Implement baseline algorithms, scrape further data and analyze them
30 Nov 2021	Implement CNN and entity linker approach and evaluate
15 Dec 2021	Implement SVM approach and evaluate
31 Dec 2021	Write and finish the first version of the MA
20 Jan 2022	Further improvement implementations
31 Jan 2022	Finish final version of MA
15 Feb 2022	**Submission**

References

Aly, M. (2005). Survey on multiclass classification methods. *Neural Netw*, 19:1–9.

Bouaziz, A., Dartigues-Pallez, C., da Costa Pereira, C., Precioso, F., and Lloret, P. (2014). Short text classification using semantic random forest. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8646 LNCS:288–299.

Boydston, P. S. and Hirst, E. S. J. (2019). Public perceptions and understanding of job titles related to behavior analysis. *Behavior Analysis in Practice 2019 13:2*, 13:394–401.

Chen, J., Hu, Y., Liu, J., Xiao, Y., and Jiang, H. (2019). Deep short text classification with knowledge powered attention. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 6252–6259.

Colas, F. and Brazdil, P. (2006). Comparison of svm and some older classification algorithms in text classification tasks. *IFIP International Federation for Information Processing*, 217:169–178.

Decorte, J.-J., Hautte, J. V., Demeester, T., and Develder, C. (2021). Jobbert: Understanding job titles through skills. *arXiv preprint arXiv*, pages 1–9.

Farooq, A., Anwar, S., Awais, M., and Rehman, S. (2017). A deep cnn based multi-class classification of alzheimer's disease using mri. *IST 2017 - IEEE International Conference on Imaging Systems and Techniques, Proceedings*, 2018-January:1–6.

Guo, H. and Wang, W. (2015). An active learning-based svm multi-class classification model. *Pattern Recognition*, 48:1577–1597.

Japkowicz, N. (2000). Learning from imbalanced data sets: A comparison of various strategies. *AAAI Technical Report WS-00-05*.

Javed, F., Luo, Q., McNair, M., Jacob, F., Zhao, M., and Kang, T. S. (2015). Carotene: A job title classification system for the online recruitment domain. *Proceedings - 2015 IEEE 1st International Conference on Big Data Computing Service and Applications, BigDataService 2015*, pages 286–293.

Javed, F., McNair, M., Jacob, F., and Zhao, M. (2016). Towards a job title classification system. *arXiv preprint arXiv*, pages 1–4.

Khamar, K. (2013). Short text classification using knn based on distance function. *International Journal of Advanced Research in Computer and Communication Engineering*, 2:1916–1919.

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., and Brown, D. (2019). Text classification algorithms: A survey. *Information 2019*, 10:150.

Le, Q. V. and Smola, A. J. (2006). Simpler knowledge-based support vector machines. *Proceedings of the 23rd international conference on Machine learning - ICML '06*.

Li, L., Peltsverger, S., Zheng, J., Le, L., and Handlin, M. (2021). Retrieving and classifying linkedin job titles for alumni career analysis. *SIGITE '21: Proceedings of the 22st Annual Conference on Information Technology Education*, pages 85–90.

Li, T., Zhang, C., and Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, 20:2429–2437.

Malherbe, E., Diaby, M., Cataldi, M., Viennet, E., and Aufaure, M. A. (2014). Field selection for job categorization and recommendation to social network users. *ASONAM 2014 - Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 588–595.

Marinescu, I. and Wolthoff, R. (2020). Opening the black box of the matching function: The power of words. *Journal of Labor Economics*, 38:535–568.

Martin-Caughey, A. (2021). What's in an occupation? investigating within-occupation variation and gender segregation using job titles and task descriptions:. *American Social Review*, 86:960–999.

Neculoiu, P., Versteegh, M., Rotaru, M., and Amsterdam, T. B. V. (2016). Learning text similarity with siamese recurrent networks. *Association for Computational Linguistics*, pages 148–157.

Paulus, W. and Matthes, B. (2013). Klassifikation der berufe : Struktur, codierung und umsteigeschlüssel. *FDZ Methodenreport*.

Schapire, R. E. and Singer, Y. (2000). Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39:135–168.

Sebastiani, F. (2001). Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1–47.

Slamet, C., Andrian, R., Maylawati, D. S., Suhendar, Darmalaksana, W., and Ramdhani, M. A. (2018). Web scraping and naïve bayes classification for job search engine. *IOP Conference Series: Materials Science and Engineering*, 288:012038.

Smith, B. N., Hornsby, J. S., Benson, P. G., and Wesolowski, M. (1989). What is in a name: The impact of job titles on job evaluation results. *Journal of Business and Psychology*, 3:341–351.

Song, G., Ye, Y., Du, X., Huang, X., and Bie, S. (2014). Short text classification: A survey. *Journal of Multimedia*, 9:635–643.

Tang, L., Tian, Y., and Pardalos, P. M. (2019). A novel perspective on multiclass classification: Regular simplex support vector machine. *Information Sciences*, 480:324–338.

Tomar, D. and Agarwal, S. (2015). A comparison on multi-class classification methods based on least squares twin support vector machine. *Knowledge-Based Systems*, 81:131–147.

Uter, W. (2020). Classification of occupations. *Kanerva's Occupational Dermatology*, pages 61–67.

Vijayan, V. K., Bindu, K. R., and Parameswaran, L. (2017). A comprehensive study of text classification algorithms. pages 1109–1113. Institute of Electrical and Electronics Engineers Inc.

Wang, F., Wang, Z., Li, Z., and Wen, J. R. (2014). Concept-based short text classification and ranking. *CIKM 2014 - Proceedings of the 2014 ACM International Conference on Information and Knowledge Management*, pages 1069–1078.

Wang, J., Abdelfatah, K., Korayem, M., and Balaji, J. (2019). Deepcarotene -job title classification with multi-stream convolutional neural network. *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, pages 1953–1961.

Wang, J., Wang, Z., Zhang, D., and Yan, J. (2017a). Combining knowledge with deep convolutional neural networks for short text classification. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*.

Wang, Y., Zhou, Z., Jin, S., Liu, D., and Lu, M. (2017b). Comparisons and selections of features and classifiers for short text classification. *IOP Conference Series: Materials Science and Engineering*, 261:1–8.

Zhu, Y., Javed, F., and Ozturk, O. (2017). Document embedding strategies for job title classification. *The Thirtieth International Flairs Conference*.