

Automatic identification of role-specific information in product development: a critical review on large language models

Dominik Ehring[✉], Ismail Menekse, Janosch Luttmer and Arun Nagarajah

University of Duisburg-Essen, Germany

[✉] dominik.ehring@uni-due.de

Abstract

In the era of digitization and the growing flood of information, the automatic, role-specific identification of information is crucial. This research paper aims to investigate whether the adaptation of LLM is suitable for classifying information obtained from standards for corresponding role profiles. This research reveals that with systematic fine-tuning, prediction accuracy can be increased by almost 100%. The validation was carried out using a two-digit number of standards for three predefined roles and demonstrates the significant potential of LM for labelling content with regard to roles.

Keywords: *artificial intelligence (AI), large language model (LLM), fine-tuning, role-specific information, knowledge management*

1. Introduction

In the era of digital transformation, users face a number of challenges in the product development process. One of the central and most challenging tasks is managing the flood of information. The number of internal and external sources of information that need to be considered and used as part of development activities is constantly increasing (Arezki et al., 2004). This is particularly due to the quantitative increase of technical documents, including laws, guidelines, standards, patents and technical documentation (Mahringer and Gabler, 2018).

Today's product development is characterised by a high degree of manual work. Employees involved in the product development process must manually identify the relevant information from a variety of sources and integrate it in a system-oriented manner. A study by the Harvard Business Review shows that 83% of companies surveyed have critical development knowledge in written documents, but finding, understanding and processing this knowledge is time-consuming. Worryingly, employees spend about 16% of their time searching and processing information in these sources (Chirapurath, 2019).

This raises the question of how to reduce or eliminate monotonous information retrieval activities, while encouraging creative and inventive activities in the product development process. This challenge highlights the need for automated, role-specific identification and delivery of relevant information. In this context, the effectiveness of machine learning approaches, in particular large language models, as an enabler for digitalised product development will be investigated in detail. This will involve analysing and evaluating the potential of language models to address the complex challenge of role-specific identification of information in the product development process. When aiming to recognise role-specific information with sufficient forecast quality, it is important to consider how to provide precisely fitting information in target systems, such as requirements management tools or CAD systems in the design environment.

1.1. State of the art

A Large Language Model (LLM) is a powerful machine learning model in the field of Natural Language Processing (NLP) that is trained to understand and generate human-like language patterns. In order to this, LLMs are artificial neural networks trained on large text data to understand, process and generate natural language. They can be used for a variety of natural language processing tasks, including text classification. The strength of LLM lies in its ability to recognise contextual dependencies and complex patterns in text data. This enables precise and flexible adaptation to different classification tasks. LLM is fine-tuned to adapt a pre-trained model to a specific task or set of requirements. There are already a large number of custom language models that generally aim to provide information in a more needs-based way. In the area of standards, the Digital Standards Initiative has already developed a domain-based language model which, through training, has a better technical understanding and leads to optimised extraction results in the context of standards. The quality of results for identifying relevant standards content for products and their characteristics has been significantly improved, with a particular focus on the typical language of standards documents. ([Digital Standards Initiative, 2023](#))

[Arezki et al, 2004](#) present an information retrieval approach to user-specific identification of information. They build a model for information retrieval based on a user profile. The aim of the model is to improve information retrieval by taking into account the user's interests and preferences. The model uses a combination of text analysis and user profiles to find relevant information based on a specific query. The user profile is generated by mapping the content of the search history in the backend, thus characterising the user's context. ([Arezki et al., 2004](#))

In addition to language models and information retrieval approaches, ontology-based approaches can also be used for classification tasks. Here, the first results show a domain-specific rather than a role-specific possibility of information identification, so that domains such as "construction" or "quality" could be automatically assigned with an accuracy of about 65% ([Ehring et al., 2023](#)).

[Kestel's \(2021\)](#) approach presents an assistance system for the development of design-related finite element analyses using ontologies. The system also provides information in a more needs-based manner. A knowledge acquisition module supports the semi-automated transfer of simulation knowledge into the database. Sources used in this study include calculation data from CAD models, unstructured texts from guidelines, specialist journals, and conference papers. In contrast to the approach presented in Chapter 2, the research of [Kestel \(2021\)](#) does not focus on large language models and their potential through fine-tuning. Additionally, the system boundary centres on a technical, delimited context and does not aim to identify information on different roles. The literary analysis of a role-specific identification of information from development documents shows that the user's request is at the centre of today's knowledge management. The solution space is generated depending on the quality of the request. In product development, the user is not always aware of the boundary conditions or the question and its context. As a result, the user may not receive all the information that is relevant to their task. The research shows that there are clear gaps in the identification of information corresponding to a user with an assigned role in the company.

1.2. Research goal

The vision is to realise the automatic identification and provision of information from development documents according to generic profiles of defined roles. This approach should enable individuals to assign themselves to a profile based on their practical role in the company. The relevant information can be identified from all documents used in product development, including e. g. internal and external standards, design guidelines, and technical documentation. The paper examines whether large language models (LLMs) can serve as suitable solutions for the role-specific identification of information. To answer this question, a critical, literary review of current LLMs will be used to select a model that will serve as the basis for further analyses. Particular attention will be paid to identifying models that meet the specific requirements and challenges in the context of role-specific identification of information. Standards are used as validation examples - in the form of training and test data - to determine and evaluate the predictive quality of the selected models. This approach makes it possible to evaluate the performance of the models on the basis of objective criteria and thus to verify their suitability for the intended purpose. Another focus of the research paper is to investigate if and how the prediction quality

of large language models can be optimised with regard to the automatic classification of information according to profiles of defined roles. The aim is to identify and discuss possibilities for improvement in order to contribute to the reorganisation of innovative knowledge management. Chapter 2 presents the methodology, which is derived from existing approaches and serves as a basis for practical implementation. The concept is then elaborated and validated in Chapter 3. The paper concludes with a detailed discussion in Chapter 4, which places the findings in a broader context. Finally, Chapter 5 provides a brief summary of the main findings and an outlook on possible further developments.

2. Methodological approach

In contrast to the approaches presented in Section 1.1 for the subject-specific labelling of information via ontology-based text classification or the provision of information in the form of generated user profiles via information retrieval, a systematic, generally valid procedure for the fine-tuning of large language models is chosen in the following. The fine-tuning of language models is a process in which a pre-trained language model is adapted to specific tasks or domains. The task is to classify information according to pre-defined roles, with the domain focusing on the area of standards. The following steps are performed to adapt language models (see Figure 1). When developing and adapting language models, the fine-tuning process goes through several crucial steps. The first step is to analyse and select a pre-trained language model. This involves a detailed examination of the various models available, taking into account factors such as architecture, size and tasks. The focus is on choosing the model that best meets the specific requirements. Model selection is followed by data pre-processing, a critical step in the process. This involves collecting or accessing relevant data for the specific task and pre-processing it. This includes tasks such as tokenisation, noise removal and formatting into the input format expected by the model. The subdivision of data into training, validation and test data forms the basis of the actual fitting process. Training data is used to adapt the model to the specific task, while validation data is used for fine-tuning and setting hyper-parameters. After training, test data is used to evaluate the performance of the model. Actual model fitting begins with initialisation of the pre-trained model, followed by adaptation to the specific task using the training data. Optimisation involves fine-tuning the weights and bias, and updating the model parameters. Validation is performed by checking the performance of the model against the validation data. Fine-tuning is based on this performance and involves adjusting hyperparameters to improve the generalisation ability of the model. Finally, the model is analysed. Testing on the test data allows the overall performance and generalisation capability to be assessed. If necessary, further optimisation is performed to achieve the desired level of performance. These steps are iterative and can be extended depending on changing requirements or data availability. (Devlin et al., 2018) The data processing procedure is divided into five steps: "Selection of roles in product development", "Definition of profiles with tasks and skills of selected roles", "Analysis and selection of standards", "Role-specific labelling of selected standards", "Noise removal, tokenisation and lemmatization", which are explained in detail in Chapter 3.

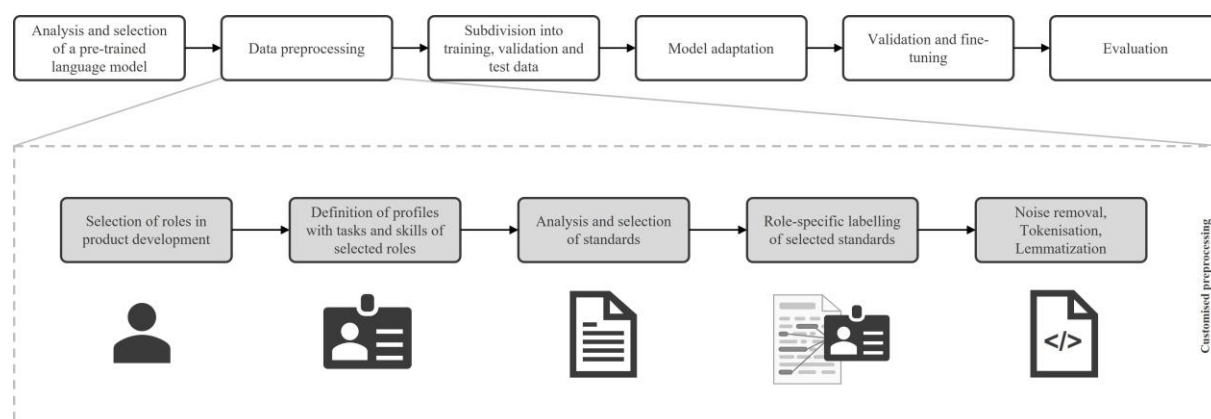


Figure 1. Methodological approach for a role-specific fine-tuning

3. Implementation and evaluation

3.1. Analysis and selection of a pre-trained language model

The analysis and selection of which pre-trained large language model should be considered for fine-tuning - against the background of the role-specific identification and classification of information - is divided into two steps. In the first step, current and popular LLMs and their characteristics are evaluated on the basis of defined basic requirements for the use case under consideration (see Table 1).

Table 1. Requirements for selection of a pre-trained LM

ID	Requirement
Req 1	The model shall be pre-trained with a sufficiently large data set of the German language for the use case.
Req 2	The model shall be open source and fine-tunable.
Req 3	The model shall have an application programming interface (API) from the software company.
Req 4	The model shall be applicable for a limited training data set.

The aim of the research is to analyse the classification performance of language models in the specific use case of assigning roles to content from German language standards. The aim is to avoid the need for new model pre-training, which would require a model that has already undergone sufficient training on the German language. The language model should be open source and, above all, fine-tunable, so that different experiments can be carried out. The API of the software developer should be used to ensure that fine-tuning remains possible for very large models, otherwise local fine-tuning cannot be carried out effectively due to the computing power required. In addition, the model must be applicable to limited training data while still achieving satisfactory prediction quality. Avoid excessively large language models, as they require large amounts of training data. All language models have been pre-trained with very large text corpora in the order of 110 million to 560 billion parameters, with English being the main focus. BLOOM is the only model that does not explicitly consider German, and only includes a small number of German text segments in some of the training documents (see Table 2). All language models except GTP4, PaLM2, Claude2 and Megatron-Turing NLG meet the requirements for fine-tuning and free availability, which means that they cannot be included in the applicability criteria for limited training data. It can be seen that the models "BERT" and "RoBERTa" as well as "LLaMa2" fulfil the defined basic requirements. After the pre-selection, in a second step, the three language models are tested for the central task of text classification and, based on this, one model is selected for validation. For reasons of comparability, literature studies are used in which the models "BERT", "RoBERTa" and "LLaMa2" are analysed on the same training and test data sets. The results of the classification task - based on four different datasets (the datasets differ in terms of content, number of data and number of labels - e.g. the dataset "AGNews" contains 120,000 news texts, of which 7,600 have to be classified with the labels "World", "Sport", "Economy" and "Technology") - show that the prediction quality of the three models has comparable results in the region of 60%. This means that each model classifies about 60% of the texts correctly. Another finding is that the larger LLaMa2 version with 13 billion parameters leads to a decrease in classification performance. This is due to underfitting, as there is not enough training data to optimally adjust the parameters of the model. (Li et al., 2023)

Smaller models, such as BERT and RoBERTa, are particularly suitable for the use case defined here, where only a very small training data set can be used, due to their architecture and the size of the model. According to the studies by Briskilal et al. and Shaheen et al. it can be concluded that RoBERTa has a higher classification accuracy than BERT. In the study by Briskila et al. a classification of idioms is carried out. The results are transferable to standards as a data set, since idioms are text fragments that do not mean what they imply. Standards are also deliberately written implicitly, leaving room for interpretation. Here, the specific language can be seen as a unifying element that causes difficulties in the classification task. 1470 pieces of text were classified and RoBERTa classified them correctly with an accuracy of 88%. In addition, the study by Shaheen et al. presents a pairwise comparison of language

models in relation to the classification of legal texts. For example, legal areas are classified into rubrics such as finance or international agreements (Shaheen et al., 2020).

It can be concluded that RoBERTa has a higher overall classification accuracy than BERT. After a literary analysis, a variant of RoBERTa, which was analysed with 145 gigabits of text in German, was selected as the language model with the most potential for classifying role-specific information from standards (Scheible et al., 2020).

With the selection of the GottBERT model, the first step of the process "Analysis and selection of a pre-trained language model" is completed.

Table 2. LLM evaluation matrix

	Usability for the German language	Fine-tunable / Open-source	Applicability for limited training data
LLaMa2	●	●	◐
Claude2	●	○	○
Megatron-Turing NLG	●	○	○
Falcon	●	●	◐
PaLM2	●	○	○
BERT	●	●	●
RoBERTa	●	●	●
UL2 20B	●	●	◐
BLOOM	◐	●	◐
GTP4	●	○	○

3.2. Data preprocessing

After the LLM selection process, the question arises as to which data are used for fine-tuning and how they are prepared in order to make a generally valid statement about the role-specific classification. The data pre-processing is divided into five steps: "Selection of roles in product development", "Definition of profiles with tasks and skills of selected roles", "Analysis and selection of standards", "Role-specific labelling of selected standards", "Noise removal, tokenisation and lemmatization".

Product development is characterised by a broad division of tasks, where users with a specific role in the company have a concrete task and skill profile. Three roles, namely Quality Manager, Design Engineer and Calculation Engineer with their specific activities are first defined for the basic investigation of the possibility of identifying and providing role-specific information by fine-tuning LLM. The main focus was to ensure that all roles overlap to some extent in terms of content. Two roles (Design engineer, Calculation engineer) are very similar in content, whereas one role (Quality manager) is very different from the others. The language model should therefore be subjected to a higher degree of difficulty when classifying roles. Profiles have been defined for each role in terms of tasks and skills, on the basis of which an assignment to relevant information from standards can be made (see Figure 2). Standards were selected as relevant development documents and datasets for fine-tuning. This has the advantage of providing a broad, cross-cutting range of knowledge that can be applied across roles. On the other hand, by taking standards into account, only one 'type' of language is relevant for the adaptation of the language model, thus reducing the amount of data required. A total of 13 standards were selected, some of which can be clearly assigned to individual roles and/or capture cross-role content. The process of role-specific tagging of selected standards content was carried out by an expert panel of 20 engineers from different disciplines. The resulting dataset was then tokenised, lemmatised and noise removed in order to begin formatting the dataset for input into the language model.

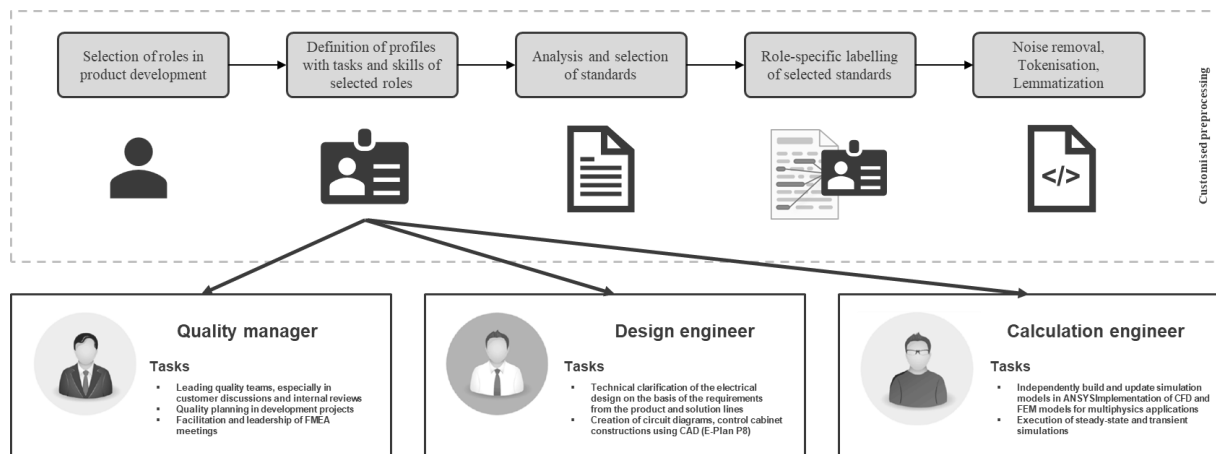


Figure 2. Examples of roles and their profiles

An example dataset is taken from DIN 743-1 "Calculation of load capacity of shafts and axles". Individual text passages were manually analysed and assigned to corresponding activities of previously defined roles with the names "calculation engineer", "quality manager" and "design engineer". The activities were taken from a preliminary analysis of job descriptions from publicly accessible job portals. In the context of the chapter "Proof of avoidance of permanent deformation, cracking or forced fracture under maximum load", for example, the phrase "Even in shafts with a hard surface layer [...] component deformation can occur before cracking [...]". Due to the lack of plastic deformation capacity of the surface layer, the verification of the avoidance of permanent deformation below the surface layer and the verification of [...] must be carried out in every case" can be assigned to a calculation engineer.

3.3. Subdivision into training, validation and test data

Once the data has been prepared, it is divided into training, validation and test data. A specific link between training and test data should be ensured. This link has been established through the normative references in the standards. This means that the test data come from standards that have been referenced in the training documents and have not yet been fed into the model via the training. The collected texts were divided into 128 token sections to ensure an even amount of information and to stabilise the training. This results in a total of 1006 labelled phrases with a sequence length of 128 tokens, which are fed into the training. Care is taken to ensure that the labelled roles are balanced (see Figure 3).

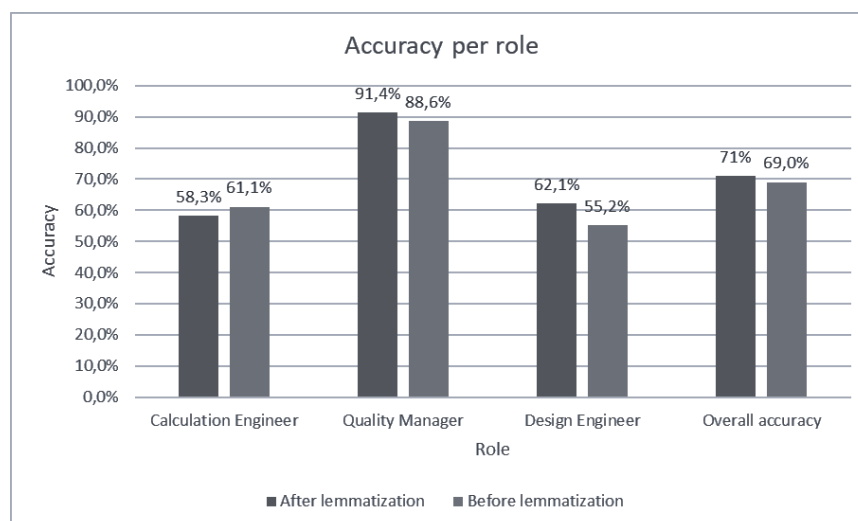


Figure 3. Number of training samples before and after lemmatization

3.4. Model adaption, validation and fine-tuning

After preparing the dataset and partitioning the data according to Section 3.3, the hyperparameters are determined before fine-tuning the GottBERT model. Hyperparameters are parameters that need to be defined before training begins, as they can directly affect the performance of the model (Claesen et al., 2015). Examples of parameters are the number of epochs, the batch size, the maximum sequence length, the learning rate or the optimiser. For example, epochs describe how often a model is run through the data during training. This means that each time a dataset is processed by a model, an epoch is completed (Mansour et al., 2023).

Since each dataset is different and therefore requires a different optimum, it is advisable to experiment with different hyperparameter settings before performing validation on the test dataset. Sobhanam et al. determined the optimal hyperparameters for different datasets in the area of text classification for RoBERTa models. One example dataset contains 836 training examples and 449 test texts. This dataset is therefore comparable to the number generated for the use case considered here. A learning rate of 1×10^{-4} and 1×10^{-5} is defined as the optimal hyperparameter. The optimal batch size, i.e. the number of examples fed into training in one run, was determined to be 8, 16, 32 and 64. To determine the optimal number of epochs, the model was deliberately run in the direction of overfitting. The optimal number of epochs is reached when the training loss, i.e. the result of the loss function, does not change significantly. After a detailed analysis, the hyperparameters for the validation results are determined with a learning rate of 4×10^{-5} , a batch size of 32, an epoch number of 8 and a maximum sequence length of 128 tokens.

3.5. Evaluation

In order to be able to make a statement about the possibility and effectiveness of LLM for role-specific identification and classification of information from standards, a comparison of the prediction quality before and after fine-tuning must be made. In a first step, the GottBERT model is applied to the test dataset without prior tuning with the training dataset. The results show that the overall accuracy of 36% is close to a random value. In the role of Quality engineer the model achieves an above average value of 71.43%, while in the role of Design engineer it achieves only 37.93%. The role of the Calculation engineer is not recognised at all by the model (see Figure 4).

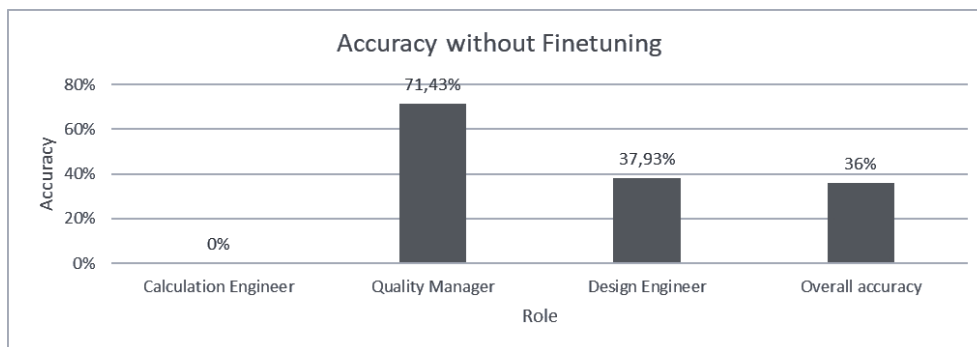


Figure 4. Accuracy without fine-tuning

After fine-tuning, 61.1% of the information is correctly identified by the Calculation engineer, 88.6% by the Quality manager and 55.2% by the Design engineer. This increases the overall accuracy to 69%, so it can be concluded that fine-tuning LLM can be a powerful method for identifying role-specific information (see Figure 5). In addition, the difficulty of differentiation becomes apparent when looking at the analysis in detail. A total of 13 phrases were misclassified in the role of Design engineer. Nine of these misclassifications were for the role of Calculation engineer, which means that the proportion of Calculation engineer misclassifications is around 70%. A closer look at the misclassifications of the Quality engineer shows that the model classified the Calculation engineer in all misclassified sentences. This is due to the fact that the roles of the Calculation engineer and the Quality engineer are similar.

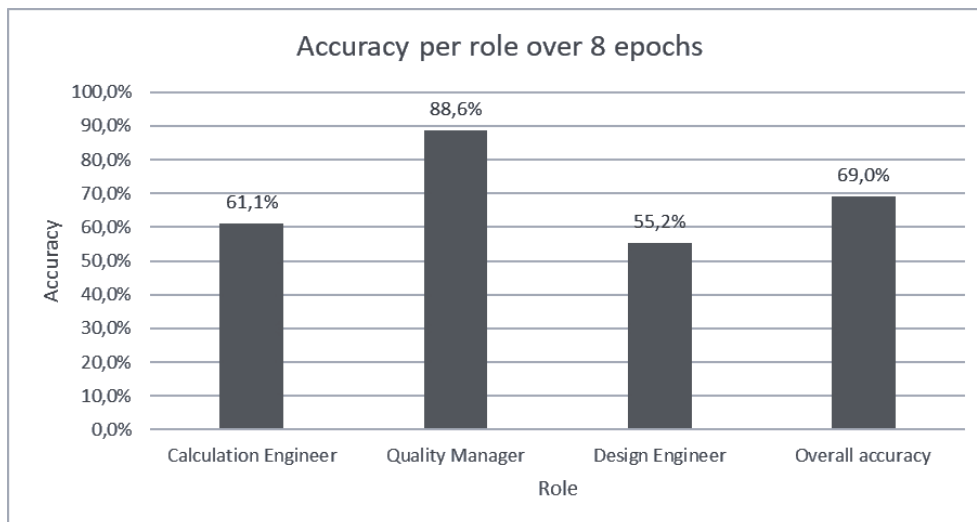


Figure 5. Accuracy per role over eight epochs

When lemmatization is taken into account during fine-tuning, the number of labelled text segments decreases to 714 (see Figure 3), but the overall accuracy increases to 71%. This effect is not surprising, as less but higher quality data is preferable. Consequently, to improve the quality of the results, lemmatization should be carried out.

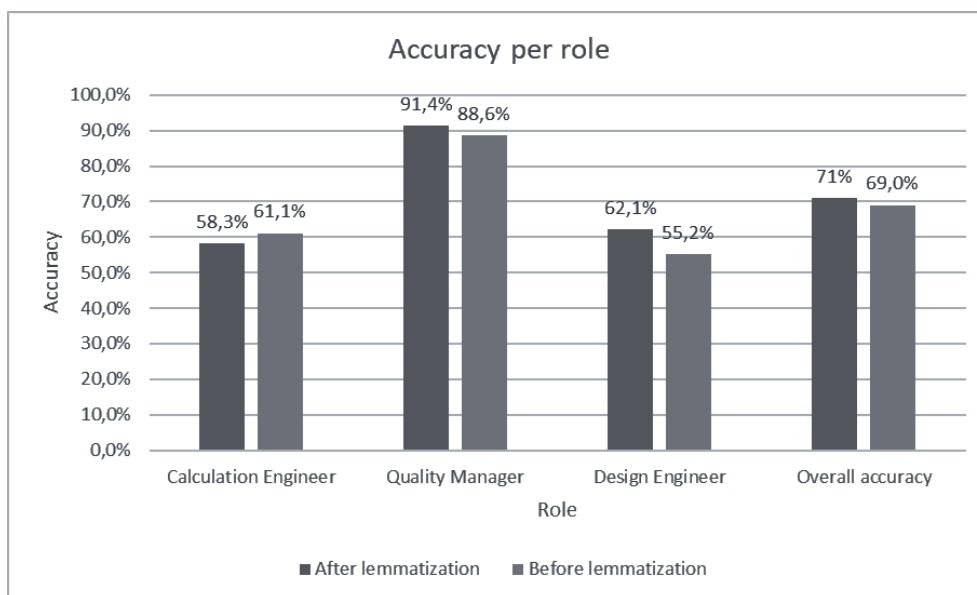


Figure 6. Accuracy per role before and after lemmatization

An interesting research question concerns the behaviour of a model when the training data set is significantly expanded. This question, together with the investigation of model behaviour when using different and non-uniform data sources, motivates the expansion of the database to include several sources. Data from freely available textbooks and Wikipedia articles were added. After this expansion and lemmatization, the training database is six times larger than before. The number of training texts per role is 1524 for the Calculation engineer, 1227 for the Quality engineer and 1469 for the Design engineer. The input texts were labelled also manually according to the activities. If the hyperparameters remain the same, the prediction quality is 75% (see Figure 7). It is evident that performance improves as the amount of data increases. Moreover, there is no negative trend in the different source documents.

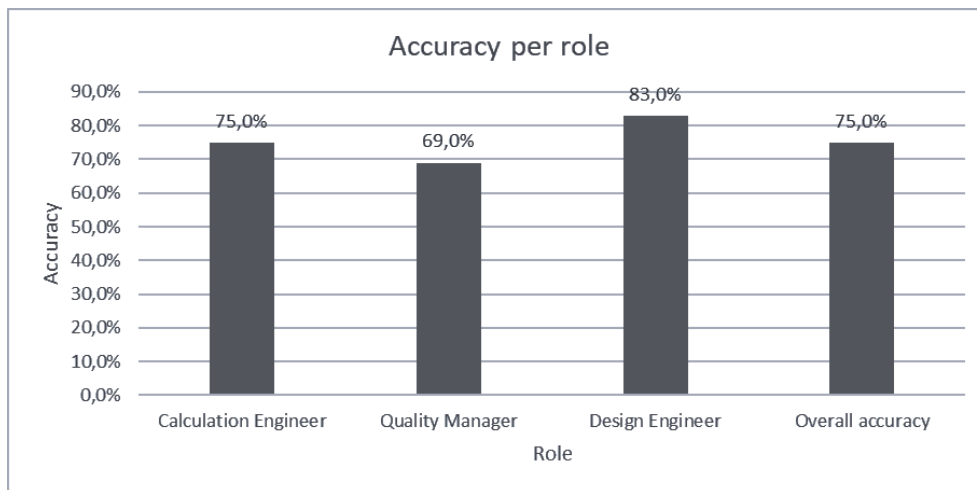


Figure 7. Accuracy per role after training with an extended dataset

4. Discussion of results

Without fine-tuning, today's models are unable to classify information in a role-specific way. The models lack too much industry-specific knowledge. This can be seen in the analysis of the selected model which, despite being pre-trained in German, only achieved an overall accuracy of 36%.

The analysis results show that fine-tuning the GottBERT model leads to a significant increase in prediction accuracy compared to an untrained model, and that lemmatization and expansion of the training data set can double the prediction quality. The adapted language model can effectively distinguish roles and assign content. It shows a high generalisation ability despite a small amount of training data, as the model performance could be increased by fine-tuning. The approach presented here is limited by the fact that only one role can be assigned to a text segment at a time. In reality, text segments may be relevant to different roles. This would have required a larger training data set. It would be interesting to investigate how the language model behaves when more than one role is assigned to the same text segment. In addition, an investigation of more than three roles is necessary in order to make a general statement about predictability and added value. In future research, the profiles need to be specified and scientifically derived. Despite all the limitations, the potential of encoder models for role-specific classification of information has been recognised. After considering the potential for recognising role-specific information, the question arises how to provide precisely fitting information, such as in requirements management systems or in the design environment. It is recommended to refer to the approaches of [Luttmer et al. \(2023\)](#), which analyse extraction techniques that recognise requirements, in general. In the future, requirements could be assigned directly to the people involved in product development and moreover could be directly implemented in CAD, CFD tools. With this approach, developers can use the algorithm to search through important documents. In the simplest version, all the necessary information for a particular activity is visually highlighted in a PDF.

5. Summary and outlook

The current way of working with development documents, especially standards, is primarily characterised by manual identification and processing of information. As an approach to optimising this time-consuming process, this research paper has shown that the use of large language models and their fine-tuning offers a way of classifying the content of standards into roles with specific tasks and skills. After a detailed analysis and selection of a language model adapted to this specific task on the basis of a prepared data set, a prediction quality of 69% was achieved without lemmatisation, 71% with lemmatisation and 75% by expanding the database with additional heterogeneous sources. Compared to the text classification before fine-tuning, this is an increase of 100% and confirms the thesis presented in chapter 1.2. The evaluation shows that LLMs have a great potential for labelling role-specific information. Future research will focus in particular on a role profile consisting of tasks, skills, responsibilities and a reference to an object of interest. One possible approach to be examined is web

scraping of job portals, from which generic domains are to be formed by clustering profiles, in order to link these in turn with suitable data for training and thus come closer to the vision of automatic and role-specific information provision.

References

- Arezki, R., Poncelet, P., Dray, G., & Pearson, D. W. (2004). "Information retrieval model based on user profile." *Artificial Intelligence: Methodology, Systems, and Applications: 11th International Conference, AIMSA 2004, Varna, Bulgaria, September 2-4, 2004. Proceedings 11* (pp. 490-499). Springer Berlin Heidelberg.
- Mahringer, C. A., & Gabler, M. (2018). *Wie können Wissensmanagementsysteme nutzerorientiert gestaltet werden? Die Rolle organisationaler Routinen.* [online] Available at https://c-mahringer.de/Mahringer_2018_Wie_k%C3%B6nnen_Wissensmanagementsysteme_nutzerorientiert_gestaltet_werden_HMD.pdf (accessed 20.09.2023).
- Chirapurath, J. (2019). Knowledge mining. The Next Wave of Artificial Intelligence-Led Transformation. Harward Business Review Analytic Services.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Initiative Digitale Standards (2023), *Optimierung von Rechercheprozessen im deutschen Normungswesen durch Methoden des Natural Language Processing.* [online] Available at: <https://www.dke.de/idis/pilotprojekte/optimierung-von-rechercheprozessen-durch-nlp> (accessed 06.11.2019).
- Ehring, D., Ferraz-Doughty, P., Luttmer, J., & Nagarajah, A. (2023). "A first step towards automatic identification and provision of user-specific knowledge: A verification of the feasibility of automatic text classification using the example of standards." *Procedia CIRP*, 119, 1103-1108.
- Kestel, P. (2021). "Assistenzsystem für den wissensbasierten Aufbau konstruktionsbegleitender Finite-Elemente-Analysen." FAU University Press.
- Briskilal, J.; Subalalitha, C. N. (2022). "An ensemble model for classifying idioms and literal texts using BERT and RoBERTa". *Information Processing & Management* 1/59, S. 102756.
- Li, Z., Li, X., Liu, Y., Xie, H., Li, J., Wang, F. L., ... & Zhong, X. (2023). "Label Supervised LLaMA Finetuning". *arXiv preprint arXiv:2310.01208*.
- Shaheen, Z., Wohlgenannt, G., & Filtz, E. (2020). "Large scale legal text classification using transformer models". *arXiv preprint arXiv:2010.12871*.
- Scheible, R., Thomczyk, F., Tippmann, P., Jaravine, V., & Boeker, M. (2020). "GottBERT: A pure German language model". *arXiv. arXiv preprint arXiv:2012.02110*.
- DIN (2012), DIN 743-1:2000-10: Calculation of load capacity of shafts and axles - Part 1: General. DIN Deutsches Institut für Normung e.V., Berlin.
- Claesen, M., & De Moor, B. (2015). "Hyperparameter search in machine learning". *arXiv preprint arXiv:1502.02127*.
- Mansour, M., Cumak, E. N., Kutlu, M., & Mahmud, S. (2023). "Deep learning based suture training system". *Surgery Open Science*, 15, 1-11.
- Sobhanam, H., & Prakash, J. (2023). "Analysis of fine tuning the hyper parameters in RoBERTa model using genetic algorithm for text classification". *International Journal of Information Technology*, 15(7), 3669-3677.
- Luttmer, J., Prihodko, V., Ehring, D., Nagarajah, A. (2023): "Requirements extraction from engineering standards: systematic evaluation of extraction techniques". The 33rd CIRP Design Conference; 17.-19.06.2023, Sydney, Australia, In: *Procedia CIRP* Jg. 119 (2023) S. 794-799.