Optimizing Medical Service Request Processes through Language Modeling and Semantic Search

Daniel Schlör
schloer@informatik.uni-wuerzburg.de
University of Würzburg
Würzburg, Germany

Jan Pfister
pfister@informatik.uni-wuerzburg.de
University of Würzburg
Würzburg, Germany

Andreas Hotho
hotho@informatik.uni-wuerzburg.de
University of Würzburg
Würzburg, Germany

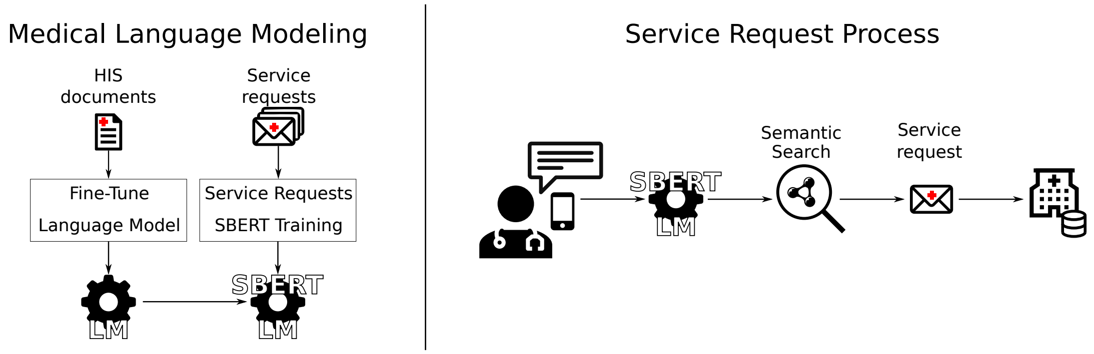

Figure 1: Medical language model training and its application to the service request process.

ABSTRACT

Medical service requests are a crucial part of the workflow in hospitals and healthcare organizations. However, the process of requesting medical services can be time consuming and can require physicians and medical personnel to navigate complex interfaces and enter detailed information about the requested service. In this paper, we propose a system that uses machine learning techniques such as large language models and semantic search to optimize the process of requesting medical services. Our approach enables physicians to request medical services using natural language rather than navigating complex interfaces, allowing for more efficient and flexible interactions with hospital information systems. We evaluate our approach on real-world data and discuss the implications of our work for the future of digital health care. Our results suggest that our approach has the potential to streamline the process of requesting medical services and reduce the time and manual effort required in the daily hospital routine.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**; **Neural networks**; • **Social and professional topics → Medical technologies**.

KEYWORDS

language modeling, semantic search, medical service optimization

ACM Reference Format:

Daniel Schlör, Jan Pfister, and Andreas Hotho. 2023. Optimizing Medical Service Request Processes through Language Modeling and Semantic Search. In *2023 the 7th International Conference on Medical and Health Informatics (ICMHI) (ICMHI 2023), May 12–14, 2023, Kyoto, Japan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3608298.3608324

1 INTRODUCTION

The early development and deployment of hospital and healthcare information systems have encouraged the ongoing digitization of hospital processes, many of which previously required paperwork and telephone arrangements. These processes are now often integrated into IT solutions and require physicians and medical staff to interact with complex interfaces and tools. Although the shift to digital data management and process support has benefited patient care in many ways, it also requires physicians to spend time navigating these interfaces and accurately capturing all relevant information as required by the system. Especially in the current era, marked by events such as the COVID-19 pandemic, the optimization of medical service request processes and the advancement of digital health solutions that organize interactions and workflows more efficiently matching the physician's daily work with patients are particularly important.

In this paper, we address the challenge of optimizing medical service request processes through the use of large neural language models and semantic search. Our goal is to improve the efficiency of interactions with hospital information systems by allowing physicians to request medical services using natural language rather than navigate complex interfaces. To achieve this, we fine-tune large

ICMHI 2023, May 12–14, 2023, Kyoto, Japan

ACM ISBN 979-8-4007-0071-2/23/05.
https://doi.org/10.1145/3608298.3608324

language models for specific medical language and adapt them to incorporate semantic search. This allows us to match freely dictated natural language requests with structured request codes used in the hospital information system, streamlining the process of requesting medical services and reducing the time and effort required to navigate complex interfaces.

The purpose of this study is to evaluate the effectiveness of our approach in improving the efficiency of medical service request processes. Our research question is as follows. Can advanced language modeling and semantic search be used to streamline the process of requesting medical services? We evaluate our approach using a dataset of natural language requests and structured request codes from a hospital information system.

Our results demonstrate that the use of advanced language modeling and semantic search allows an accurate match of natural language requests with structured request codes. This suggests that our approach is an effective tool for combining the natural interaction of physicians with the requirements of structured and well-defined request data required for hospital information systems to overall improve the efficiency of the medical service request process.

In the following sections, we first provide a more detailed description of the current service request process. We then introduce our approach including the language modeling methodology and our algorithm to match natural language requests with associated service codes. Finally, we evaluate our approach on real-world data and conclude with a discussion of the implications of our work for the future of digital health care.

2 SERVICE REQUEST PROCESS

In hospitals, several departments specialize in specific diagnostic services and offer these services to other departments. When a department needs a diagnostic service from another department, it submits a request for a specific examination. The requested examination is then performed by the department that provides the diagnostic service, and the results are reported back to the requesting department.

In the past, these requests for diagnostic services were organized mainly by phone. However, today many hospitals have digitized this process using IT solutions. In our case, for example, physicians can request diagnostic services from the radiology department using a mobile phone application. To improve the request process, the hospital has implemented a system that allows physicians to request diagnostic services using unstructured natural language, either by typing the request or by dictating it to a speech-to-text service.

However, for execution, documentation, and billing, the precise diagnostic procedure must be recorded in a structured format that matches the requirements of the hospital information system. To meet these demands and reduce the workload of practicing physicians, the hospital employs medical assistants to refine the request and extract structured information, such as the exact service code, from this unstructured text before it is forwarded to the appropriate department.

For example, the physician dictates the following text to the system: "*X-ray chest PA in standing position after serial fracture on the left for follow-up. Appointment today, please*". The medical assistant then has to extract relevant information from this request and select the correct service code from a list of over 2 000 possibilities.

This approach has two disadvantages: it takes time, which can be a problem in urgent cases, and is limited by the work capacity of the medical assistant staff, who could be doing more valuable tasks, including patient care. These disadvantages can cause problems, especially in times of unexpected high demand, such as observed during the COVID-19 pandemic. This emphasizes the need for technical solutions that can help hospitals and healthcare organizations adapt to the changing landscape of medical service requests in the post-pandemic era, which motivated the development of the system proposed in this paper.

From another point of view, the implemented process offers good opportunities for developing technical solutions based on machine learning (ML) or artificial intelligence (AI), since training data is generated throughout the process in daily use.

3 PROPOSED SYSTEM

To overcome the drawbacks of the current process, it is important to ensure that medical assistants are not the bottleneck in the process. This can be achieved in a number of ways, such as allowing requesting physicians to refine requests themselves or reducing the workload per request to make refinement more scalable for the assistants. However, all options involve making refinement more efficient while maintaining the intuitive and accessible interface of unstructured dictation for the physician.

Mapping the natural language request text to structured information needed for the process can be formulated as information extraction and ML task, for example, as predicting the service codes for a given request text. Such an automated approach has the advantage to be scalable, effectively reducing the manual workload for refinement of the requests. Although this approach allows fully automated refinements from a technical point of view, considering the principles of responsible medical AI [16], the proposed use case focuses on supporting the medically trained worker, i.e., the physician or the assistant, to efficiently refine requests.

From a conceptional point of view, addressing this task with ML and AI involves two challenging aspects: First, understanding the unstructured request texts, which contain highly specialized medical (and in our case German) language, and second, predicting the precise service codes that best match the requests, which can be one or more codes per request with potentially few and ambiguous indicator words given in the request favoring one service over the other. Accordingly, we propose a system as depicted in fig. 1 consisting of both steps, language modeling, and semantic search, which we describe in detail in the following sections.

3.1 Language Modeling

In recent years, transformer-based large language models have significantly improved natural language processing tasks [4, 22]. For this reason, we decided to use such models for our language-modeling approach. For language modeling of German texts, several multilingual and specifically German models have been proposed, such as gbert [3]. However, our use-case involves highly specific medical language, which requires a more specialized approach.

Algorithm 1 Medical Language Modeling for Service Requests

```
his_dataset = prepare_HIS_dataset()                          ▷ Provide doctor's letters from HIS
model = load_pretrained_gbert()
model = model.fine_tune(his_dataset)                         ▷ Fine-tune model on medical texts
req_dataset = (requests, request_codes)                      ▷ Provide service requests and associated codes
req_test = sample(req_dataset)                               ▷ Sample 300 requests for evaluation
req_train = req_dataset \ req_test                           ▷ Exclude test samples from training
sim_matrix = calculate_sim_matrix(req_train)                 ▷ Pairwise Jaccard similarity between all request-code sets
contrast_dataset = ∅
for request, codes ∈ req_train do
    pos_samples = sim_matrix.most_similar(codes)             ▷ Sample 50 requests with most similar codes
    neg_samples = sim_matrix.least_similar(codes)            ▷ Sample 50 requests with dissimilar codes
    samples = pos_samples ∪ neg_samples
    for sample_request, sample_codes ∈ samples do
        sim_score = jaccard(codes, sample_codes)             ▷ Similarity between both code-sets
        contrast_dataset = contrast_dataset ∪ (request, sample_request, sim_score)
    end for
end for
model = model.contrastive_training(contrast_dataset)         ▷ Perform contrastive training
return model
```

To address this, we first train a large German pre-trained language model, gbert [3], on 11 658 005 German medical papers (gbert-medpaper) that we crawled from the web and on 4 700 752 medical texts (gbert-med) that we extracted from the hospital information system, such as doctors' letters. This allows us to gain a general understanding of German medical language. Next, we focus on the specific language used in service requests, including phrases, technical terms, and indicator words that are semantically relevant to these requests. To do this, we adopt the SBERT training objective [11], which is designed to learn semantically meaningful representations of complete sentences or, in our case, complete requests.

To further fine-tune our language models to this objective, we compose a dataset containing pairwise similar and dissimilar service requests. We quantify the similarity of two request texts using the Jaccard index between the sets of services that were actually performed after each request was made. This is motivated by the observation that requests that resulted in the same examinations are semantically similar. We also sample negative examples of dissimilar requests and use this dataset to perform contrastive SBERT training to further fine-tune the language model.

In detail, for each request in the dataset, we selected 50 positive and negative samples for contrastive training. For positive samples, the similarity value of the 50th most similar request was taken as the threshold to define the positive subset to sample from. All samples with a similarity higher than the threshold were included as positive samples, while from the set of samples at the threshold similarity, the remaining candidates to match 50 samples were randomly selected. Negative samples were randomly drawn from the large set of requests mostly without overlapping service codes, resulting in a dataset consisting of 375 100 samples leaving out each 1 000 positive and negative samples for evaluation. The language modeling approach is summarized in algorithm 1.

3.2 Semantic Search

With our LMs fine-tuned to reflect semantic similarity of request texts with respect to the service codes that the requests are associated with, semantic search can be applied to find requests in a search corpus semantically similar to the query request to be revised and structured. The benefit of such an approach based on a search corpus is that with the ongoing request revision process, the search corpus can be extended by new samples. This has the advantage that such a system will automatically adapt to novel service demands, such as those that may arise in epidemiological situations or during other major health-threatening events, thereby solving the previously identified challenge of scalability for service request refinements.

From a technical point of view, our system holds request embeddings, i.e. the semantic representations of all requests in the search corpus learned by the fine-tuned language model, which are inferred once the system is started. These are stored together with the service codes collected in the refinement process up to the time of the current request. Once a new request is queried, the system calculates the embeddings of the request for this new request using the fine-tuned language model. Then the cosine similarity between the query request and the requests in the search corpus are calculated, indicating the semantic similarity. Sorted by this similarity score, the most similar requests from the search corpus are retrieved along with the service codes they have been associated with. Note that by reporting this similarity value to the medical assistant or physician, an indicator $0 \leq i \leq 1$ is provided of how reliable the suggestion of the service code can be considered. Along with the subsequent refinement process, this can be used to define thresholds or learn models in the future. This allows to prioritize the refinement process, opening opportunities to distinguish between cases which are most certainly correct and can be, for example, directly be approved by the requesting physician, and more complicated cases or cases presumably predicted incorrectly, which are more time-consuming or complex to refine, and thus assigned to medical assistants or expert physicians.

4 EVALUATION

As each request can correspond to one or more service codes, the ML task of predicting the correct codes can be seen as a multi-label classification problem. The Jaccard index can be used to evaluate the performance of a system that predicts multiple labels for a given input, as detailed in eq. (1). It is a widely used evaluation metric in this context [17, 24], as it allows the comparison of the predicted labels with the true labels.

$$Jaccard(Y, \hat{Y}) = \frac{1}{|Y|} \sum_{i=1}^{|Y|} \frac{Y_i \cap \hat{Y}_i}{Y_i \cup \hat{Y}_i} \tag{1}$$

Therefore, Y denotes the list of true label sets $Y_i = \{l_1, \ldots, l_{n_i}\}$ in the test dataset, while $\hat{Y}$ denotes the list of predicted label sets $\hat{Y}_i = \{\hat{l}_1, \ldots, \hat{l}_{\hat{n}_i}\}$ according to our system.

To evaluate the performance of our system, we randomly selected a subset of 300 requests, which were excluded in contrastive training and also removed from the search corpus to guarantee a fair evaluation. These requests and their respective service codes assigned by the medical assistants were additionally manually reviewed to meet the criteria for correctly worded requests to rule out artifacts that are sometimes present due to speech-to-text errors, and to ensure the correct assignment of service codes. The final test dataset consists of 279 requests, which are used in this evaluation.

4.1 Results

In this study, we first report quantitative evaluation in the service code prediction setting, and second we evaluate the correlation of the similarity score as an indicator of prediction quality.

4.1.1 Request Prediction. For this study we evaluate both contributions separately, our fine-tuned medical LMs (gbert-med, gbert-medpaper) as well as our SBERT request similarity ($\text{SBERT}_{\text{requests}}$) training to learn semantically meaningful sentence embeddings. Therefore, we first compare our gbert-med and gbert-medpaper models with the underlying baseline, gbert-base and several other domain specific language models, such as German-MedBERT [18], BioBERT [9], SciBERT [2] and ClinicalBERT [1]. For all approaches, we use average pooling over the word embeddings to generate request embeddings and select the most similar request according to these average-pooled request embeddings along with their corresponding request codes. The results, as depicted in the first row block of table 1 show that our LM fine-tuned to the medical language from hospital information systems outperforms all other models by a large margin with a Jaccard score of 0.659, followed by our model, fine-tuned on German medical papers with a Jaccard score of 0.589.

Next, we evaluated the benefit of LMs pre-trained with the SBERT objective[1] on various non-medical sentence similarity datasets (noted $\text{SBERT}_{\text{general}}$). The results shown in the second block of table 1 suggest, that some LMs generally tuned to sentence similarity are able to solve the task better than the gbert-med LM, which has not been pre-trained with SBERT. This suggests, that the LM benefit from introducing a sentence-similarity-based training, although the fine-tuning to medical language is able to outperform five models specifically trained for sentence similarity.

Finally, we introduce our request-specific SBERT training to tailor the models to our medical service request domain semantics, training each model for 5 epochs over our contrastive learning dataset as detailed in section 3.1. The results depicted in the third block of table 1 show, that our $\text{SBERT}_{\text{requests}}$ training objective improves the performance on the task for all models by large margin although the performance gain varies over different models: The best result was achieved by a multilingual distilBERT model [12, 14], which was originally trained on Quora questions for sentence similarity. Without our service request SBERT training, the model achieved performance of 0.645 which is outperformed by our gbert-med model. Training the model in our $\text{SBERT}_{\text{requests}}$ task, however, improved its performance by large margin yielding a Jaccard score of 0.841. The second best model is based on MPNet [19] and was already among the best models without our SBERT fine-tuning improving from 0.712 to 0.839 directly followed by our gbert-medpaper model and the sciBERT [2] model yielding 0.837, both. These results show that our $\text{SBERT}_{\text{requests}}$ training is highly effective to improve the semantic modeling of service request similarity over a large variety of different LMs.

[1] All models obtained from https://www.sbert.net/docs/pretrained_models.html

Table 1: Language modeling results. The *w/o SBERT* models are language models without SBERT training. *$\text{SBERT}_{general}$* refers to pre-trained general-purpose SBERT LMs, *$\text{SBERT}_{requests}$* refers to SBERT contrastive training on our request data as described in algorithm 1.

	Approach	Jaccard
w/o SBERT	German-MedBERT	0.550
	biobert-base-cased-v1.2	0.551
	scibert_scivocab_uncased	0.555
	gbert-base	0.578
	ClinicalBERT	0.586
	gbert-medpaper (ours)	0.589
	gbert-med (ours)	0.659
$\text{SBERT}_{general}$	cross-en-de-roberta-sentence-transformer	0.594
	quora-distilbert-multilingual	0.645
	distilbert-multilingual-nli-stsb-quora-ranking	0.645
	paraphrase-multilingual-MiniLM-L12-v2	0.649
	distiluse-base-multilingual-cased-v1	0.652
	paraphrase-multilingual-mpnet-base-v2	0.674
	all-mpnet-base-v2	0.683
	multi-qa-mpnet-base-dot-v1	0.712
	msmarco-distilbert-multilingual-en-de-v2	0.736
$\text{SBERT}_{requests}$	paraphrase-multilingual-mpnet-base-v2	0.674
	all-mpnet-base-v2	0.705
	German-MedBERT	0.818
	distiluse-base-multilingual-cased-v1	0.825
	msmarco-distilbert-multilingual-en-de-v2	0.826
	cross-en-de-roberta-sentence-transformer	0.827
	gbert-base	0.827
	gbert-med (ours)	0.828
	biobert-base-cased-v1.2	0.829
	ClinicalBERT	0.831
	paraphrase-multilingual-MiniLM-L12-v2	0.832
	distilbert-multilingual-nli-stsb-quora-ranking	0.834
	scibert_scivocab_uncased	0.837
	gbert-medpaper (ours)	0.837
	multi-qa-mpnet-base-dot-v1	0.839
	quora-distilbert-multilingual	0.841

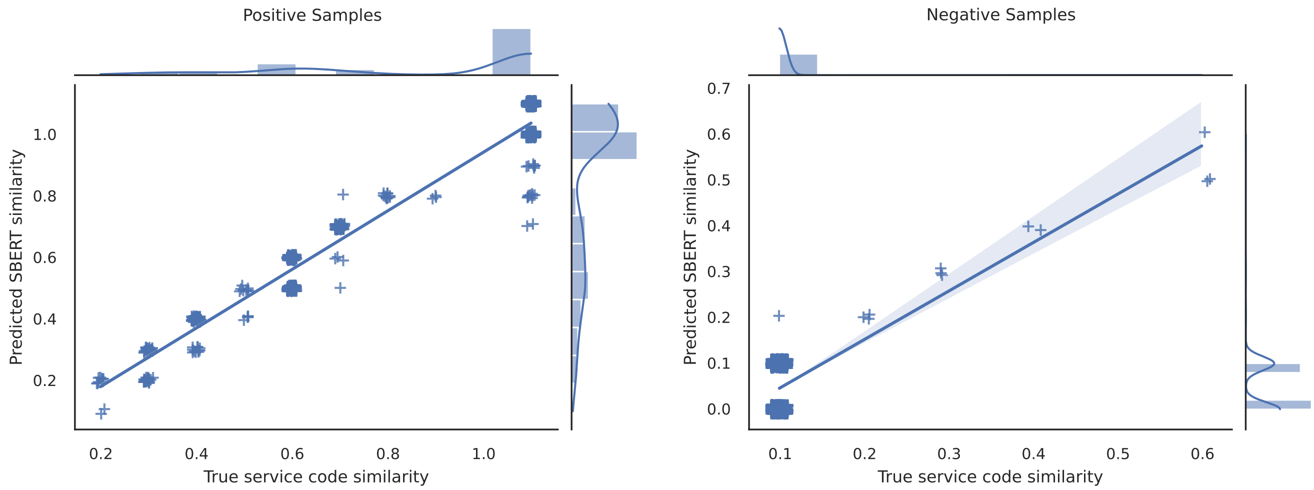

Figure 2: Comparison of SBERT-prediction-based request text similarity and true service code similarity.

4.1.2 Semantic Similarity. Besides the direct use of semantic search in the downstream task of predicting medical service requests, in section 3.2, we also discussed incorporating the semantic similarity score as an indicator of how reliable the prediction is. To validate the assumption that the cosine similarity score of two request text corresponds to the Jaccard similarity of their request codes, we used the semantic similarity test dataset consisting of 1 000 positive and 1 000 negative samples held out of the $\text{SBERT}_{\text{request}}$ training to evaluate the correlation between both scores. We first retrieve the request embeddings applying our $\text{SBERT}_{\text{requests}}$ gbert-med LM to all sample pairs. We then composed two scores lists, one list containing the Jaccard indices between the service codes of all pairs, and another list containing the cosine similarities between the embeddings predicted by the LM. The distributions for positive and negative samples are shown in fig. 2. The Pearson correlation coefficient of 0.998 suggests that the cosine similarity of the texts and the Jaccard similarity of the codes are highly correlated, confirming our initial assumption and motivating the use of the similarity score as indicator in our system.

5 RELATED WORK

Recent large-scale language models based on a transformer architecture have significantly improved natural language processing tasks. Although, as Starlinger et al. [21] discuss, there are challenges in processing German medical texts, particularly the lack of available resources and tools, several models have been proposed for different language modeling tasks in the German medical domain. Zesch and Bewersdorff [23] provide an overview of available data sources and natural language processing models for the German medical domain, as well as strategies to overcome data scarcity. With GERNERMED++ [6], Frei et al. present a statistical model for named entity recognition in German medical texts and demonstrate the effectiveness of transfer learning on pre-trained deep language models, word alignment, and neural machine translation. For information extraction tasks, Roller et al. present mEx [13], an information extraction system for German medical texts specifically targeted to the field of nephrology with the ability to perform several NLP tasks.

In addition to language modeling, research has also been conducted in the area of semantic search in the medical domain. Langnickel et al. presents preVIEW [8], a semantic search engine to explore COVID-19 research preprints based on indexing of identifiers of relevant biomedical concepts. Similarly, López-García et al. introduce SEMCARE [10], a multilingual platform for semantic search in clinical texts in English, German and Dutch, which is also based on traditional NLP and concept mapping. Koopman et al. also propose a concept-based approach to search in electronic medical records that uses the SNOMED-CT ontology to transform queries and documents into medical concepts [7]. Soto et al. describe Thalia [20], a semantic search engine for biomedical abstracts that recognizes eight different types of concepts based on named-entity recognition and ontology concept matching. In contrast to our system, these approaches use ontology-based semantics, which require a mapping between ontologies and named entities or tokens in general. For this fine-grained ontologies covering the language are required, as well as the tools to extract those entities from the text introducing additional complexity and errors, especially for non-English source languages. In contrast, our approach does not rely on hand-crafted ontologies, but semantically matches current and previous text snippets directly. Recently, similar approaches have been applied, for example, to retrieve scientific information related to COVID-19 [5] or for the detection of dementia [15].

6 CONCLUSION

In this work, we presented a novel approach to optimize the medical service request process in hospitals that incorporates state-of-the-art large language models as well as semantic search.

We fine-tuned a language model to improve text understanding of German medical language and constructed a contrastive learning task to focus on semantic similarity relevant for our task of service request prediction. In a comparison, our medical language model outperformed the underlying base model and other state-of-the-art models specifically developed for the medical domain. Further, training in our contrastive $\text{SBERT}_{\text{requests}}$ setting showed very good results improving all language models by a large margin.

Finally, we evaluated our model in terms of practical considerations in the clinical process and showed that the similarity score can serve as a reliable indicator that the prediction is correct. This allows us to implement our approach into the process, finally overcoming the disadvantages of the current process, such as the high manual effort. By this, the proposed system has the potential to improve scalability in the process, which is an important endeavor to transform digital health in the post-pandemic era.

ACKNOWLEDGMENTS

This research was supported by the Bavarian State Ministry for Science and the Arts within the "Digitalisierungszentrum für Präzisions- und Telemedizin" (DZ.PTM) project, as part of "BAYERN DIGITAL II".

REFERENCES

[1] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323* (2019).

[2] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. *arXiv:1903.10676* (2019).

[3] Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's Next Language Model. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 6788–6796.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[5] Andre Esteva, Anuprit Kale, Romain Paulus, Kazuma Hashimoto, Wenpeng Yin, Dragomir Radev, and Richard Socher. 2021. COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. *NPJ digital medicine* 4, 1 (2021), 1–9.

[6] Johann Frei, Ludwig Frei-Stuber, and Frank Kramer. 2022. GERNERMED++: Transfer Learning in German Medical NLP. *arXiv preprint arXiv:2206.14504* (2022).

[7] Bevan Koopman, Peter Bruza, Laurianne Sitbon, and Michael Lawley. 2012. Towards semantic search and inference in electronic medical records: an approach using concept-based information retrieval. *The Australasian medical journal* 5, 9 (2012), 482.

[8] Lisa Langnickel, Roman Baum, Johannes Darms, Sumit Madan, and Juliane Fluck. 2021. COVID-19 preVIEW: Semantic Search to Explore COVID-19 Research Preprints. *Studies in health technology and informatics* 281 (2021), 78–82.

[9] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.

[10] Pablo López-García, Markus Kreuzthaler, Stefan Schulz, Daniel Scherr, Philipp Daumke, Kornél G Markó, Jan A Kors, Erik M Van Mulligen, Xinkai Wang, Hanney Gonna, et al. 2016. SEMCARE: Multilingual Semantic Search in Semi-Structured Clinical Data.. In *eHealth*. 93–99.

[11] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[12] Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813* (2020).

[13] Roland Roller, Christoph Alt, Laura Seiffe, and He Wang. 2018. mEx-an information extraction platform for german medical text. In *Proceedings of the 11th International Conference on Semantic Web Applications and Tools for Healthcare and Life Sciences (SWAT4HCLS'2018). Semantic Web Applications and Tools for Healthcare and Life Sciences (SWAT4HCLS-2018), December*. 3–5.

[14] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv* abs/1910.01108 (2019).

[15] Yamanki Santander-Cruz, Sebastián Salazar-Colores, Wilfrido Jacobo Paredes-García, Humberto Guendulain-Arenas, and Saúl Tovar-Arriaga. 2022. Semantic Feature Extraction Using SBERT for Dementia Detection. *Brain Sciences* 12, 2 (2022), 270.

[16] Daniel Schlör and Andreas Hotho. 2022. Towards Responsible Medical Diagnostics Recommendation Systems. *arXiv preprint arXiv:2209.03760* (2022).

[17] Junyuan Shang, Shenda Hong, Yuxi Zhou, Meng Wu, and Hongyan Li. 2018. Knowledge guided multi-instance multi-label learning via neural networks in medicines prediction. In *Asian Conference on Machine Learning*. PMLR, 831–846.

[18] Manjil Shrestha. 2021. Development of a Language Model for Medical Domain. , 141 pages.

[19] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems* 33 (2020), 16857–16867.

[20] Axel J Soto, Piotr Przybyła, and Sophia Ananiadou. 2019. Thalia: semantic search engine for biomedical abstracts. *Bioinformatics* 35, 10 (2019), 1799–1801.

[21] Johannes Starlinger, Madeleine Kittner, Oliver Blankenstein, and Ulf Leser. 2017. How to improve information extraction from German medical records. *It-Information Technology* 59, 4 (2017), 171–179.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[23] Torsten Zesch and Jeanette Bewersdorff. 2022. German Medical Natural Language Processing–A Data-centric Survey. *Applications in Medicine and Manufacturing* (2022), 137.

[24] Min-Ling Zhang and Zhi-Hua Zhou. 2013. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering* 26, 8 (2013), 1819–1837.