Transformer-Based Medical Statement Classification in Doctor-Patient Dialogues

Farnod Bahrololloomi and **Johannes Luderschmidt** and **Biying Fu**
Faculty DCSM
RheinMain University of Applied Sciences
Wiesbaden, Germany
{farnod.bahrololloomi, johannes.luderschmidt, biying.fu}@hs-rm.de

Abstract

The classification of medical statements in German doctor-patient interactions presents significant challenges for automated medical information extraction, particularly due to complex domain-specific terminology and the limited availability of specialized training data. To address this, we introduce a manually annotated dataset specifically designed for distinguishing medical from non-medical statements. This dataset incorporates the nuances of German medical terminology and provides a valuable foundation for further research in this domain. We systematically evaluate Transformer-based models and multimodal embedding techniques, comparing them against traditional embedding-based machine learning (ML) approaches and domain-specific models such as medBERT.de. Our empirical results show that Transformer-based architectures, such as the Sentence-BERT model combined with a support vector machine (SVM), achieve the highest accuracy of 79.58% and a weighted F1-Score of 78.81%, demonstrating an average performance improvement of up to 10% over domain-specific counterparts. Additionally, we highlight the potential of lightweight ML-models for resource-efficient deployment on mobile devices, enabling real-time medical information processing in practical settings. These findings emphasize the importance of embedding selection for optimizing classification performance in the medical domain and establish a robust foundation for the development of advanced, domain-adapted German language models.

1 Introduction

With the introduction of the Transformer architecture by Vaswani et al. (2017), substantial progress was achieved in many application areas, including general natural language processing (NLP) tasks and also in the field of medicine. However, models based on the Bidirectional Encoder Representations from Transformers (BERT) architecture (Devlin et al., 2018), initially trained on large-scale, general-purpose datasets such as Wikipedia, have struggled to accurately classify medical information in German datasets due to the complex and specialized vocabulary of medical language and the scarcity of labeled domain-specific datasets (Idrissi-Yaghir et al., 2024). To address these challenges, specialized models for the medical domain have been developed. An example is the German model medBERT.de, which has been fine-tuned with medical data and achieves an average Area Under the Receiver Operating Characteristic (AUROC) score of approximately 88% on various evaluated medical benchmarks (Bressem et al., 2024). Domain-specific models like medBERT.de can, for instance, detect whether medically relevant information is discussed in dialogues between doctors and patients. This capability is critical for extracting relevant data for patient documentation and improving the Electronic Health Record (EHR) system. Medical documentation is a cornerstone of healthcare, supporting patient care, legal accountability, and research. Yet, the processing of German medical texts remains challenging due to the inherent linguistic complexity and the limited availability of annotated datasets. As our contribution in this paper, we compare different Transformer-based models fine-tuned for medical data with traditional embedding-based methods. In particular, we focus on the analysis of German doctor-patient interviews to determine the most effective approach for classifying medical statements. Furthermore, we introduce a manually labeled dataset of medical statements to support future research in the processing of German medical texts. In doing so, we address two research questions:

- **RQ1:** How does the performance of Transformer models fine-tuned on medical data compare to traditional embedding-based approaches in classifying German doctor-patient

Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP 2025), pages 63–73
August 1, 2025 ©2025 Association for Computational Linguistics

interviews?

- **RQ2:** How does the performance of fine-tuned Transformer and machine learning (ML)-models improve when evaluated on dataset of medical statements for domain-specific German medical texts?

2 Related Work

The classification of text in a medical context represents a fundamental challenge in the field of NLP, particularly in the medical domain. Accurate categorization of medical documents can significantly improve information extraction and decision-making processes (Kesiku et al., 2022). The complex and specialized terminology in medical texts poses a particular difficulty. Managing synonyms, polysemy, and multi-word terms is essential, as these can distort the true meaning of a text (Shanavas et al., 2020). In addition, medical text data often shows low density and high dimensionality due to its special linguistic characteristics, making its classification more challenging compared to other domains (Zhou et al., 2021).

Several studies have shown that ML-models may achieve high accuracy in medical text classification when adapted to the specific language and structure of medical texts. These techniques include support vector machines (SVMs), naive Bayes, logistic regression, and k-nearest neighbors (k-NNs). These methods are often combined with word representation models, such as term frequency-inverse document frequency (TF-IDF) and Word2Vec, to improve classification performance. (Mascio et al., 2020; Almazaydeh et al., 2023)

Almazaydeh et al. (2023) used the mtsamples.com dataset (MTSamples, 2025) to train ML-models using TF-IDF, Bag-of-Words (BOW), and Word2Vec as word representations. They were able to classify 20 medical categories. The Word2Vec-based k-NN classifier achieved an average accuracy of 92%. However, the performance on German medical datasets is unknown due to the challenges posed by the strict regulatory framework of the General Data Protection Regulation (GDPR).

Transformer-based models are gaining importance in medical NLP research. Idrissi-Yaghir et al. (2024) compared different German BERT architectures on medical datasets and evaluated them on different downstream tasks such as named entity recognition (NER), multi-label classification, and extractive question answering. The results show that models with medical or translation-based pre-training typically outperform generic language models, as they are better at capturing complex medical terminology and medical context. The language models achieved the following average F1-Scores: CLEF eHealth 2019 (Neves et al., 2019): 0.820, RadQA (Dada et al., 2023): 0.816, GraSCCo (Modersohn et al., 2022): 0.673, BRONCO150 (Kittner et al., 2021): 0.844, and GGPONC 2.0 (Borchert et al., 2022): 0.779. Idrissi-Yaghir et al. (2024) showed that continued pretraining can match or even surpass the performance of medical models trained from scratch. Furthermore, pretraining on medical data or leveraging translated texts has proven to be an effective approach for domain adaptation in medical NLP tasks. In addition to medBERT.de, there is also BioGottBERT by Lentzen et al. (2022), which was fine-tuned specifically on medical data. They conducted a comprehensive analysis of the suitability of existing and new transformer-based models for the German biomedical and clinical domain by systematically comparing 8 general-purpose language models and 3 newly trained models, including BioGottBERT and two BioELECTRA versions. The study showed that General-Purpose Language Models (GPLMs) performed surprisingly well on clinical NLP tasks, with a German variation of BERT called GBERT (Chan et al., 2020) performing particularly well on document classification tasks and BioGottBERT on NER tasks. Domain adaptation of existing models proved to be more effective than training new models from scratch, which was mainly attributed to the limited size of the pre-training corpus.

In recent years, several German medical datasets have been published, such as GGPONC (Borchert et al., 2020) and BRONCO150 (Kittner et al., 2021), which include annotation information for NER and part-of-speech (POS) tagging. Other German datasets, such as those from Makowski and Simko (2018) and Suominen et al. (2020), lack such annotation. Datasets like CLEF eHealth 2019 (Neves et al., 2019) offer German medical queries and documents for information retrieval and question–answering (QA); RadQA (Dada et al., 2023) comprises German radiology reports with questions to support radiological reasoning and GraSCCo (Modersohn et al., 2022) offers annotated social-care correspondence for entity and relation extraction. A specific German dataset for intent recognition in doctor-patient interviews was developed by Rojowiec et al. (2020), consisting of 63 classes. These classes represent various categories or intentions of questions and statements that can occur during doctor-patient conversations. The dataset supports medical students in taking medical histories by interacting with virtual patients, and the doctors' intentions were detected using BERT (Rojowiec et al., 2020). Section 3 provides further details on this dataset and its application in the context of this paper.

While it has been shown that Transformer-based models can perform well with domain adaptation, their performance in German dialog-based context recognition is not as well studied, and there is no high-quality medical dataset available to classify whether a statement contains medically relevant information or not.

3 Data Acquisition

To develop a German contextualized ML-model for classifying medical and non-medical statements, we used the publicly available "Intent Recognition in Doctor-Patient Interview" (IntRec) dataset (Rojowiec et al., 2020). This dataset consists of German transcriptions of live doctor-patient interviews conducted during university training sessions, in which medical students interviewed actors portraying patients, transcribing only the doctors' statements. 80% of the entries in the dialogue sequence consist of statements in the form of questions directed at the patient, such as "When was the surgery?" while 20% are normal statements, such as "I think so, yes.". For each entry, the corresponding class, its position within the sequence, the previous statement, and the class of the preceding statement are also provided. Table 1 shows the corresponding metadata about the original dataset before preprocessing.

Attribute	Statistic
Total number of samples	2,397
Number of classes	63 (62 + "OTHER")
Classes with $\leq$ 10 samples	50%
Largest class ("OTHER")	1,169 samples
Second-largest class ("AM02")	$>$ 85 samples
Annotated with two classes	101 (4%)
Average utterance length	10 words
Utterance	Previous utterances, intention

Table 1: Overview of the dataset for intent recognition in doctor-patient interviews.

The dataset consists of a total of 2,397 samples with multiple dialogue-label pairs, where 101 of these pairs have two label assignments. Each label consists of a symptom category and a question ID. The symptom category defines the symptom area, and the question ID specifies the intent within that area. For example, the label (PH10) belongs to the "Prior History" category (PH) and refers with question ID 10 to questions about "heart diseases". The dataset includes seven symptom categories (see Table 2).

We developed a preprocessing pipeline in which we divided the samples into individual dialogues and their associated labels. Each utterance and its corresponding labels, as well as the preceding utterance and its labels, were assigned individually to each target utterance and label. In the next step, duplicates in the utterance column were removed, resulting in a normalized dataset of $1,418$ dialogue-class pairs.

Symptom Category	Code
Main Symptoms	MS
Prior History	PH
Allergies and Medication	AM
Social and Family History	SF
System Review	SR
Inquiry	IQ
Other Questions	OQ

Table 2: Symptoms categories and code, with "IQ" + "OQ" summarized under the category "OTHER".

To develop a classification model for detecting medical statements, we transformed the multiclass problem into a binary problem. The dataset was transformed by grouping all categories unrelated to "IQ" or "OQ" under the class "MEDICAL", while "IQ" and "OQ" were combined into the class "OTHER". Following the categorization described by Rojowiec et al. (2020), the symptom category "Inquiry", although referring to previously posed questions, was not considered to contain medically relevant information. In addition, redundant punctuation, such as quotation marks ("), was removed from the documents using regular expressions as an additional preprocessing step to improve data quality. The normalized dataset was split into training and test data in an $80/20$ ratio (see Table 3). To address potential data bias, the dataset was randomized prior to splitting.

In addition, a second test dataset was developed using the publicly available Berlin-Tübingen-Oncology Corpus (BRONCO150) by Kittner et al. (2021). This German-language corpus consists of 150 discharge summaries from cancer patients treated at the Charité-Berlin University of Medicine or the University Hospital of Tübingen. To prevent the reconstruction of discharge summaries and patient identities, Kittner et al. (2021) shuffled the summaries and anonymized them at the sentence level. The dataset, originally intended for information extraction from German medical texts, comprises $8,976$ sentences with POS annotations and includes medical entities along with relevant attributes like negation and speculation.

Since the BRONCO150 dataset contains not only complete sentences but also other information from discharge summaries, we manually labeled the data to extract only complete sentences or medically accurate statements. For a realistic evaluation of the models trained on the IntRec dataset, the BRONCO150 dataset was manually labeled based on specific criteria, categorizing statements as either medical or non-medical:

1. The sentence contains a medical claim.
2. Punctuation at the end is not mandatory if the content conveys a medical statement.
3. The sentence cannot be used as a title.
4. The sentence begins with an uppercase letter.
5. A sentence must not be a list or contain a colon ":" unless it begins with a date and a statement.

Manual labeling was conducted using the publicly available tool LabelStudio[1] (Tkachenko et al., 2020-2025). Annotation was performed by a Computer Science PhD student with expertise in NLP. Of the $8,976$ records, $6,863$ medical statements remain after labeling and duplicate removal. Approximately 60.15% of the data received the label 0 because many sentences contained formatting information such as date values or document headers, "Dear Sir or Madam" or document lines such as "Line ID. from document". This resulted in a reduction, leaving 39.85% with a value of 1. In addition, the dataset included partial sentences that were not standalone statements, but related to the previous line. Furthermore, enumerations were not considered because they were not independent sentences with statements. The following examples from the BRONCO150 dataset are English translations of original German texts published in the work of Kittner et al. (2021). To demonstrate these criteria, we present the following examples from the BRONCO150 dataset. Statements labeled as "MEDICAL" satisfy these conditions by expressing clear clinical information. For instance, the direct quotes "On 07/04/2134, the patient received an uneventful nivolumab infusion." (Kittner et al., 2021, Fig. 1) and "A highly suspicious HCC lesion was observed in liver segment VI on CT." (Kittner et al., 2021, Fig. 1) reflect medical events and fulfill criteria (1) to (5). In contrast, direct quotes such as "Start of chemotherapy according to the GeT protocol cycle 1." (Kittner et al., 2021, Fig. 1) or "Diagnoses: RA: choroidal melanoma (ED 07/2023)" (Kittner et al., 2021, Fig. 3) are often abbreviated, context-dependent, or formatted as titles or lists, thereby violating criteria (3) and (5), and are classified as "OTHER". The resulting German-language dataset can be used not only for our case, but also for fine-tuning German models on medical data, with the aim of supporting medical data extraction and improving semi-automatic methods for annotating medical documents. We use this dataset to evaluate how well the transfer learning of all trained models performs on unseen data, to understand whether the models can understand not only previous medical queries but also complex medical language and derive correct classifications for medical statements. The fully labeled dataset by Bahrololloomi (2025), consisting of the $8,976$ sentence_ids and labels is publicly available in the form of a CSV file. This dataset acts as a mapping and can be combined one-to-one with the original dataset by Kittner et al. (2021).

Dataset	OTHER	MEDICAL
Train/Validation 1134 (80%)	688	446
Test 284 (20%)	160	124
Test BRONCO150 6863 (100%)	4127	2736

Table 3: Data distribution and class distribution for IntRec and the normalized BRONCO150 data with class 0 as "OTHER" and class 1 as "MEDICAL".

4 Model Engineering

We extracted embeddings from four different Transformer models based on the BERT architecture to classify medical statements within sentences. These embeddings were then combined with five traditional ML-models for classification. The advantage of pure embedding extraction, as opposed to training the entire Transformer model, is evident

[1] https://labelstud.io/

in the decreased training duration and the capability to efficiently adapt these models into a mobile variant. This adaptation facilitates their use for local predictions, such as in smartwatches.

During the model selection process, we ensured the use of a German, a multilingual, and a medically specialized English model to systematically evaluate the transfer performance in the classification process. The multilingual model is a variant of the Sentence Transformer (Sentence-BERT)[2] from Reimers and Gurevych (2019). Additionally, we used a general German BERT model (BERT_{ger})[3] (Bavarian State Library, 2025) to evaluate the transfer performance of the BERT architecture on medical data.

We also selected the BioBERT model (BioBERT)[4] for the classification of medical documents. This model was developed by Deka et al. (2022) and specifically trained on English scientific publications related to medical trials. Furthermore, the German model medBERT.de (MedBERT)[5], created by Bressem et al. (2024), was used. This model was trained on a comprehensive collection of German medical documents, including medical reports and patient records. Due to its optimization for longer texts, MedBERT is particularly suitable for the analysis and classification of medical information and outperformed other German-language models in NLP tasks such as NER.

The following ML-models have been used: CatBoost (Dorogush et al., 2018), RandomForest (RF) (Pedregosa et al., 2011a), XGBoost (Chen and Guestrin, 2016), SVM (Pedregosa et al., 2011a), and LightGBM (Ke et al., 2017). In order to extract the best possible embedding, we compared different extraction strategies by calculating the average of the last hidden states over the sequence dimension (mean pooling), extracting the maximum value over all tokens (max pooling), and using the hidden state of the first token (CLS token) as a representation of the entire sequence.

The overall architecture of our approach is as follows. In the first step, the cleaned and shuffled $1,134$ sentences from the training and validation IntRec dataset are passed to the four Transformer models. Simultaneously, the mentioned extraction strategies are applied to the vanilla variants of the models to extract the required embeddings. These embeddings are then fed to the five ML-models. The same seed was used on the dataset to reproduce the same training and validation data. We use Sentence-BERT as a baseline for comparison with other Transformer models. Similarly, for the ML-models, we apply a Word2Vec approach to convert the medical data into embeddings, as suggested by Almazaydeh et al. (2023). We did not train a Word2Vec model from scratch, but used the pre-trained German Word2Vec[6] model from Yamada et al. (2020). In the next step, fixed parameters such as batch size, learning rate, and maximum padding size calculated over both datasets are set on the Transformer models, and hyperparameter optimization is performed on the ML-models via grid search using the validation data. The final step involves evaluating all ML-models on the IntRec and BRONCO150 test datasets.

5 Evaluation

As discussed in Section 4, the IntRec training and validation data are initially utilized to train and optimize the four proposed Transformer models. This process aims to identify the optimal parameters, enabling the selection of the most suitable model for the subsequent steps. To ensure optimal computational efficiency, we first performed document analysis on both datasets to determine the maximum token length and padding size. The German medical model MedBERT of Bressem et al. (2024) was used for this determination. As shown in Table 4, a maximum padding size of 143 tokens is sufficient to cover all sequences in both the normalized IntRec $(1,418)$ and BRONCO150 $(6,863)$ datasets, each consisting of labelled sentences. We also found that the BRONCO150 dataset contains documents of a greater length than the IntRec dataset. This discrepancy can be attributed to the divergent nature of the text: while the IntRec dataset is primarily composed of doctor questions directed at patients, the BRONCO150 dataset consists of discharge summaries that require a more comprehensive level of understanding.

The hyperparameters were set uniformly for all Transformer models with a number of epochs of 20, a batch size of 20, a learning rate of 2×10^{-5},

[2] https://huggingface.co/sentence-Transformers/paraphrase-multilingual-MiniLM-L12-v2

[3] https://huggingface.co/dbmdz/bert-base-german-cased

[4] https://huggingface.co/pritamdeka/BioBert-PubMed200kRCT

[5] https://huggingface.co/GerMedBERT/medbert-512

[6] https://huggingface.co/Word2vec/wikipedia2vec_dewiki_20180420_300d

Metric	IntRec 1418	BRONCO150 6863
Maximum Token Count	143	142
Average Length	14.90	18.32
Median Length	13.0	14.0
Standard Deviation	8.53	14.94

Table 4: Statistical properties of token lengths for both datasets.

and a maximum padding size of 143. For optimization, the AdamW optimizer is employed, as it offers more robust convergence compared to the traditional Adam optimizer due to its enhanced regularization through weight decay (Baevski et al., 2020). To minimize overfitting, a linear scheduler uniformly reduces the learning rate during training. Early stopping is implemented to terminate training if the validation accuracy (*val_acc*) fails to improve over three consecutive epochs. This approach prevents overfitting and reduces unnecessary computation. The training loss is calculated using *BCEWithLogitsLoss* from Pytorch (Paszke et al., 2019). Our analysis indicates that mean pooling is the most effective method for extracting embeddings. Consequently, it is consistently applied across all ML-models (see Appendix Table 9).

Hyperparameter optimization of ML-models is performed using Word2Vec embeddings with grid search and triple cross-validation, evaluated based on weighted F1-Score. The CatBoost model undergoes a separate optimization process, since *GridSearchCV* (Pedregosa et al., 2011b) is incompatible with the *Pool* format of CatBoost. Instead, the model is trained on a training dataset (*train_pool*) and evaluated on a validation dataset (*val_pool*). The best parameter configuration is determined based on the highest F1-Score. In addition to the application of hyperparameter optimization using Word2Vec embeddings, extensive hyperparameter exploration was simultaneously performed on the full set of ML-models, incorporating every available variant of BERT embeddings. The best parameters for each model are listed in the Appendix in the Table 8. These parameters are consistently applied to all ML and Transformer models without explicit mention in the Tables, as the optimal parameters are always used.

After determining the best hyperparameters, both the Transformer-based BERT models and all variations of the ML-models with BERT and Word2Vec embeddings were trained and validated on the cleaned IntRec test data. To measure the performance of the models, we use well-known metrics such as accuracy, precision, recall and F1-Score for both classes (medical and general). The individual results on the validation data are shown in Table 5.

Classifier	Acc.	Macro F1	Weighted F1	Gen. F1	Med. F1
Sentence-BERT	0.84	0.83	0.84	0.87	0.80
BERT$_{\text{ger}}$	**0.84**	0.84	0.84	0.87	**0.81**
BioBERT	0.77	0.76	0.77	0.80	0.72
MedBERT	**0.85**	**0.84**	**0.85**	**0.88**	**0.81**

Table 5: Performance of classification models on the IntRec validation data.

The metric Medical F1-Score (Med. F1) indicates how well the model correctly classifies medical statements, in contrast to the metric General F1-Score (Gen. F1), which represents the F1-Score over documents labeled as general. The Macro Avg F1-Score (Macro F1) calculates the average F1-Score across all classes, regardless of their size. In contrast, the Weighted Avg F1-Score (Weighted F1) additionally weights the size of each class and adjusts the F1-Score accordingly. The results show that MedBERT delivers the best overall performance, achieving an accuracy of 0.85 and a high F1-Score in both the macro and weighted average. The MedBERT model achieves a macro F1-Score of 0.84 and a weighted F1-Score of 0.85, indicating its ability to effectively perform both balanced and weighted classifications. In comparison, the English BioBERT shows the weakest performance, especially in the medical context, with an F1-Score of only 0.72. This model only achieves an accuracy of 0.77, indicating its limited ability to correctly classify medical statements in this specific dataset. Interestingly, both Sentence-BERT and BERT$_{\text{ger}}$ achieve similar performance, with an accuracy of 0.84 and a consistent Weighted and Gen. F1-Score of 0.84 and 0.87, respectively. Both models show strong and balanced classification performance, but they perform slightly worse than the MedBERT model. For the evaluation of the ML-models on the IntRec validation data with the respective text representations, the Weighted F1-Score is used as evaluation metric (see Table 7).

Classifier	Word2Vec	Sentence-BERT	BERT$_{\text{ger}}$	BioBERT	MedBERT
CatBoost	0.6839	**0.8372**	0.7621	0.6678	0.7813
RandomForest	0.6611	**0.8121**	0.7086	0.6535	0.7310
XGBoost	0.6551	**0.8059**	0.7519	0.6720	0.7671
SVM	0.6946	**0.8330**	0.7616	0.6668	0.7854
LightGBM	0.6654	**0.8107**	0.7599	0.6551	0.7567

Table 7: Weighted F1-Scores of ML-models with varying text representations on the IntRec validation data.

Model	**Word Rep.**	**Acc. IntRec**	**F1-IntRec**	**Acc. BRONCO**	**F1-BRONCO**
CatBoost	Word2Vec	$0.6620 \pm 3.33\,e^{-16}$	$0.6218 \pm 1.11\,e^{-16}$	$0.5993 \pm 1.11\,e^{-16}$	$0.4931 \pm 5.55\,e^{-17}$
	Sent.-BERT	$0.7746 \pm 1.11\,e^{-16}$	$0.7614 \pm 1.11\,e^{-16}$	$0.5974 \pm 0.00\,e^{-16}$	$0.4787 \pm 5.55\,e^{-17}$
	BERT_{ger}	$0.7183 \pm 1.11\,e^{-16}$	$0.7006 \pm 1.11\,e^{-16}$	$0.6016 \pm 0.00\,e^{-16}$	$0.4526 \pm 5.55\,e^{-17}$
	BioBERT	$0.6162 \pm 1.11\,e^{-16}$	$0.6000 \pm 2.22\,e^{-16}$	$\mathbf{0.6283 \pm 2.22\,e^{-16}}$	$\mathbf{0.5542 \pm 1.11\,e^{-16}}$
	MedBERT	$0.6514 \pm 1.11\,e^{-16}$	$0.6189 \pm 1.11\,e^{-16}$	$0.6012 \pm 1.11\,e^{-16}$	$0.4521 \pm 0.00\,e^{-16}$
RF	Word2Vec	$0.6479 \pm 1.11\,e^{-16}$	$0.6103 \pm 1.11\,e^{-16}$	$0.5951 \pm 0.00\,e^{-16}$	$0.4808 \pm 1.66\,e^{-16}$
	Sent.-BERT	$0.7394 \pm 2.22\,e^{-16}$	$0.7145 \pm 1.11\,e^{-16}$	$0.5920 \pm 1.11\,e^{-16}$	$0.4584 \pm 1.66\,e^{-16}$
	BERT_{ger}	$0.6866 \pm 1.11\,e^{-16}$	$0.6540 \pm 1.11\,e^{-16}$	$0.6013 \pm 0.00\,e^{-16}$	$0.4519 \pm 0.00\,e^{-17}$
	BioBERT	$0.6268 \pm 1.11\,e^{-16}$	$0.6049 \pm 1.11\,e^{-16}$	$0.6209 \pm 2.22\,e^{-16}$	$0.5226 \pm 2.22\,e^{-16}$
	MedBERT	$0.6549 \pm 0.00\,e^{-16}$	$0.6073 \pm 1.11\,e^{-16}$	$0.6015 \pm 0.00\,e^{-16}$	$0.4520 \pm 1.66\,e^{-16}$
XGBoost	Word2Vec	$0.6268 \pm 1.11\,e^{-16}$	$0.6091 \pm 2.22\,e^{-16}$	$0.6088 \pm 1.11\,e^{-16}$	$0.5087 \pm 2.22\,e^{-16}$
	Sent.-BERT	$0.7183 \pm 1.11\,e^{-16}$	$0.6994 \pm 1.11\,e^{-16}$	$0.5997 \pm 1.11\,e^{-16}$	$0.4953 \pm 0.00\,e^{-17}$
	BERT_{ger}	$0.6937 \pm 2.22\,e^{-16}$	$0.6711 \pm 3.33\,e^{-16}$	$0.6031 \pm 2.22\,e^{-16}$	$0.4587 \pm 5.55\,e^{-17}$
	BioBERT	$0.6232 \pm 1.11\,e^{-16}$	$0.6130 \pm 0.00\,e^{-16}$	$0.6200 \pm 0.00\,e^{-16}$	$0.5399 \pm 0.00\,e^{-16}$
	MedBERT	$0.6585 \pm 2.22\,e^{-16}$	$0.6301 \pm 1.11\,e^{-16}$	$0.6013 \pm 0.00\,e^{-16}$	$0.4548 \pm 5.55\,e^{-17}$
SVM	Word2Vec	$0.6549 \pm 0.00\,e^{-16}$	$0.6469 \pm 1.11\,e^{-16}$	$0.5659 \pm 0.00\,e^{-16}$	$0.5043 \pm 1.11\,e^{-16}$
	Sent.-BERT	$\mathbf{0.7958 \pm 1.11\,e^{-16}}$	$\mathbf{0.7881 \pm 2.22\,e^{-16}}$	$0.5885 \pm 0.00\,e^{-16}$	$0.4806 \pm 5.55\,e^{-17}$
	BERT_{ger}	$0.7465 \pm 1.11\,e^{-16}$	$0.7370 \pm 2.22\,e^{-16}$	$0.5990 \pm 0.00\,e^{-16}$	$0.4518 \pm 0.00\,e^{-16}$
	BioBERT	$0.6338 \pm 1.11\,e^{-16}$	$0.6093 \pm 0.00\,e^{-16}$	$0.5911 \pm 0.00\,e^{-16}$	$0.5322 \pm 1.11\,e^{-16}$
	MedBERT	$0.6761 \pm 2.22\,e^{-16}$	$0.6395 \pm 2.22\,e^{-16}$	$0.6013 \pm 0.00\,e^{-16}$	$0.4516 \pm 1.11\,e^{-16}$
LightGBM	Word2Vec	$0.6514 \pm 1.11\,e^{-16}$	$0.6288 \pm 1.11\,e^{-16}$	$0.5957 \pm 1.11\,e^{-16}$	$0.5080 \pm 0.00\,e^{-16}$
	Sent.-BERT	$0.7852 \pm 0.00\,e^{-16}$	$0.7746 \pm 0.00\,e^{-16}$	$0.5955 \pm 1.11\,e^{-16}$	$0.4717 \pm 0.00\,e^{-16}$
	BERT_{ger}	$0.7148 \pm 0.00\,e^{-16}$	$0.6951 \pm 1.11\,e^{-16}$	$0.6016 \pm 0.00\,e^{-16}$	$0.4562 \pm 5.55\,e^{-17}$
	BioBERT	$0.6479 \pm 1.11\,e^{-16}$	$0.6358 \pm 1.11\,e^{-16}$	$0.6159 \pm 0.00\,e^{-16}$	$0.5380 \pm 2.22\,e^{-16}$
	MedBERT	$0.6831 \pm 1.11\,e^{-16}$	$0.6560 \pm 0.00\,e^{-16}$	$0.6010 \pm 2.22\,e^{-16}$	$0.4526 \pm 1.66\,e^{-16}$
Sent.-BERT	-	$0.7676 \pm 1.11\,e^{-16}$	$0.7671 \pm 1.11\,e^{-16}$	$0.5280 \pm 0.00\,e^{-16}$	$0.5191 \pm 1.11\,e^{-16}$
BERT_{ger}	-	$0.7711 \pm 2.22\,e^{-16}$	$0.7655 \pm 2.22\,e^{-16}$	$0.5790 \pm 2.22\,e^{-16}$	$0.4538 \pm 1.66\,e^{-16}$
BioBERT	-	$0.7218 \pm 2.22\,e^{-16}$	$0.7101 \pm 1.11\,e^{-16}$	$0.6018 \pm 2.22\,e^{-16}$	$0.4796 \pm 5.55\,e^{-17}$
MedBERT	-	$0.7782 \pm 2.22\,e^{-16}$	$0.7752 \pm 0.00\,e^{-16}$	$0.6048 \pm 1.11\,e^{-16}$	$0.4938 \pm 0.00\,e^{-17}$

Table 6: Performance of various classification models on IntRec and BRONCO150 test data, based on accuracy and weighted F1-Score. The results include the mean and standard deviation from 100 evaluations.

The results show that, in contrast to the direct comparison with the Transformer models, all ML-models achieve the best results with multilingual Sentence-BERT embeddings, reaching an average Weighted F1-Score of 0.8198 with a low standard deviation of 0.0114. This indicates a consistent performance of the ML-models with this embedding. In comparison, the BioBERT and Word2Vec embeddings have an average performance that is 19.12% and 18.02% worse, respectively. These differences in model performance indicate that the multilingual Sentence BERT embeddings are best suited for the given classification task. The stable results show that this representation not only delivers high F1-Scores, but also exhibits low variance between models, further demonstrating its robustness. However, the overall results are worse than those of the Transformer variants.

To evaluate the robustness of the Transformer and ML-models, a data-driven analysis was performed during inference. Both the test data of the IntRec dataset and the normalized and labeled $6,863$ large BRONCO150 dataset were randomly shuffled 100 times with different but fixed seeds for the iteration index. Table 6 presents the results obtained, showing the mean and standard deviation for all metrics. Since the standard deviations for all models are in the range of 10^{-16}, they are presented with the factor e^{-16}. The results underline how crucial both the choice of the classification model and the underlying embedding representation are. Although MedBERT showed the best performance on the validation data, the MedBERT embeddings overall do not perform optimally on the IntRec test data. Notably, pure Transformer models do not outperform on average an SVM working in combination with Sentence-BERT embeddings. In particular, this combination achieves the best results with an accuracy of 0.7958 and a weighted F1-Score of 0.7881. The superiority of the Sentence-BERT embeddings over alternative representations such as Word2Vec, BERT_{ger}, BioBERT or MedBERT highlights the importance of a powerful embedding base, especially in the analysis of medical datasets. Furthermore, the extremely low standard deviations confirm the high robustness and reproducibility of the results, a factor further favored by the weighted F1-Scores, which take into account the class frequencies. Overall, the analysis shows that for optimal classification performance in the medical domain, not only the model complexity, but also the targeted selection of embeddings is of central importance.

Given that the BRONCO150 dataset consists entirely of domain-specific medical statements, and that no prior model training has included such data, its evaluation provides a potential means of exploring the transfer learning ability of different approaches when confronted with novel and, to some extent, partially different sentence structures. Table 6 shows that all models achieve robust results, with accuracy values mostly above 59% and weighted F1-Scores delivering consistent results. It is worth noting that the CatBoost model combined with BioBERT embeddings achieves the best results with an accuracy of 0.6283 and a weighted F1 Score of 0.5542. These results suggest that BioBERT embeddings, which are already pre-trained on medical texts, offer a significant advantage in the classification of purely medical sentences. The observed differences in performance can mainly be explained by the different characteristics of the datasets. While the IntRec dataset used for training mainly contains doctor-patient interviews with comparatively simple medical terminology, the content of the BRONCO150 dataset is based on discharge summaries, which document the course of treatment and the main medical findings and therapy decisions in detail. This high degree of precision and the distinct linguistic style complicate the direct transfer of the classification capabilities acquired during training, thereby accounting for the divergent results.

6 Discussion

Our study investigated the classification of medical statements in German doctor-patient dialogues by integrating Transformer-based models with traditional ML-models that leverage BERT-based embeddings. The evaluation provided key insights into model performance and domain adaptability, while highlighting the trade-offs between general-purpose and domain-specific methods. Regarding RQ1, our findings reveal that domain-specific models such as MedBERT.de even though explicitly optimized for medical texts do not exhibit a significant advantage over general-purpose Transformer models in dialogue-based medical contexts. Sentence-BERT, a non-domain-specific model, achieved an F1-Score of 0.84, which is nearly equivalent to that of MedBERT.de (F1 = 0.85). This suggests that high-quality sentence embeddings extracted from general Transformers can compensate for the lack of domain-specific pretraining in certain scenarios. In contrast, the comparatively weaker performance of BioBERT shows challenges related to linguistic and data-specific adaptation, particularly in cross-lingual settings. Our evaluation indicates that hybrid approaches such as combining an SVM classifier with Sentence-BERT embeddings yield strong performance on the test set, achieving the highest accuracy (0.80) and weighted F1-Score (0.79). This finding emphasizes the importance of careful selection of embedding strategies and model architectures for the effective classification of medical statements. To understand the performance differences observed in RQ2, it is important to note that, while both datasets contain German medical language, they differ in context and linguistic formality: IntRec features short, spoken questions, whereas BRONCO150 consists of structured discharge summaries. In the context of RQ2, the evaluation on the BRONCO150 dataset, which consists of structured medical texts, shows that models trained on conversational data struggle to generalize to more formal medical documents. While Sentence-BERT based models excel in doctor-patient dialogues, domain-specific embeddings like BioBERT deliver better performance for structured medical statements. This divergence shows the need to tailor embedding strategies to the specific nature of the text being analyzed. In conclusion, our research confirms that Transformer-based models, when optimally integrated with advanced embedding strategies, are capable of delivering accurate and robust classification of medical statements. The RQ1 is answered, showing the feasibility of employing hybrid approaches in doctor-patient interviews. This work not only sets a solid foundation for the evolution of more sophisticated models in the field but also highlights the critical importance of careful embedding selection and parameter tuning in navigating the challenges inherent in specialized medical language. Regarding RQ2, the complexity of the BRONCO150 dataset poses a significant challenge. None of the models achieved a good F1-Score on this data. Although accuracy remained higher than the F1-Score, this suggests that the models are more effective at classifying “OTHER” statements while struggling with “MEDICAL” ones.

7 Conclusion and Future Work

This study identifies several opportunities for future research. A practical evaluation of the proposed methods in real-world medical settings is essential to assess their effectiveness in automated text extraction within EHR systems. In this context, the application of knowledge distillation techniques should be explored to adapt models for resource-constrained environments, such as mobile devices and smartwatches, enabling real-time processing. In addition, future work should systematically investigate the extent to which automatically generated examples (e.g., via GPT-4o or other Large Language Models (LLMs)) can reduce the need for manual labeling. In particular, it is crucial to assess the quality of the resulting pseudo-labels and to explore how a hybrid approach (synthetic + manual) can yield robust models in resource-constrained environments. Furthermore, extending the approach to multi-turn dialogues and incorporating clinician feedback could enhance classification accuracy and system robustness. To better capture the context of IntRec’s short and isolated sentences, we plan to reframe the task as a QA problem by concatenating each QA instance into a single input and predicting its original label. Future work should also focus on optimizing embedding selection strategies, leveraging data augmentation techniques, and investigating transfer learning approaches to mitigate the performance gap between conversational and structured medical texts. Additionally, evaluating these models in real-world deployment scenarios, such as automated documentation systems, will provide valuable insights into their practical applicability. By addressing these challenges and refining current methodologies, future research can significantly improve the efficiency and domain relevance of automated medical text processing.

References

Laiali Almazaydeh, Mohammad Abuhelaleh, Arar Al Tawil, and Khaled Elleithy. 2023. Clinical text classification with word representation features and machine learning algorithms. *International Journal of Online and Biomedical Engineering (iJOE)*, 19(04):65–76.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint*.

Farnod Bahrololloomi. 2025. Bronco150 mapping: Medically relevant vs. non-medically relevant statements.

Bavarian State Library. 2025. bert-base-german-cased (revision 43cce13).

Florian Borchert, Christina Lohr, Luise Modersohn, Thomas Langer, Markus Follmann, Jan Philipp Sachs, Udo Hahn, and Matthieu-P. Schapranow. 2020. Ggponc: A corpus of german medical text with rich metadata based on clinical practice guidelines. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*. Association for Computational Linguistics.

Florian Borchert, Christina Lohr, Luise Modersohn, Jonas Witt, Thomas Langer, Markus Follmann, Matthias Gietzelt, Bert Arnrich, Udo Hahn, and Matthieu-P. Schapranow. 2022. GGPONC 2.0 - the German clinical guideline corpus for oncology: Curation workflow, annotation policy, baseline NER taggers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3650–3660, Marseille, France. European Language Resources Association.

Keno K. Bressem, Jens-Michalis Papaioannou, Paul Grundmann, Florian Borchert, Lisa C. Adams, Leonhard Liu, Felix Busch, Lina Xu, Jan P. Loyen, Stefan M. Niehues, Moritz Augustin, Lennart Grosser, Marcus R. Makowski, Hugo J.W.L. Aerts, and Alexander Löser. 2024. medbert.de: A comprehensive german bert model for the medical domain. *Expert Systems with Applications*, 237:121598.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German‘s next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, pages 785–794. ACM.

Amin Dada, Tim Leon Ufer, Moon Kim, Max Hasin, Nicola Spieker, Michael Forsting, Felix Nensa, Jan Egger, and Jens Kleesiek. 2023. Information extraction from weakly structured radiological reports with natural language queries. *European Radiology*, 34(1):330–337.

Pritam Deka, Anna Jurek-Loughrey, and Deepak P. 2022. *Evidence Extraction to Validate Medical Claims in Fake News Detection*, pages 3–15. Springer Nature Switzerland.

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

Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. Catboost: gradient boosting with categorical features support. *arXiv preprint*.

Ahmad Idrissi-Yaghir, Amin Dada, Henning Schäfer, Kamyar Arzideh, Giulia Baldini, Jan Trienes, Max Hasin, Jeanette Bewersdorff, Cynthia S. Schmidt, Marie Bauer, Kaleb E. Smith, Jiang Bian, Yonghui Wu, Jörg Schlötterer, Torsten Zesch, Peter A. Horn, Christin Seifert, Felix Nensa, Jens Kleesiek, and Christoph M. Friedrich. 2024. Comprehensive study on german language models for clinical and biomedical text understanding. *arXiv preprint*.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3149–3157, Red Hook, NY, USA. Curran Associates Inc.

Cyrille YetuYetu Kesiku, Andrea Chaves-Villota, and Begonya Garcia-Zapirain. 2022. Natural language processing techniques for text classification of biomedical documents: A systematic review. *Information*, 13(10):499.

Madeleine Kittner, Mario Lamping, Damian T Rieke, Julian Götze, Bariya Bajwa, Ivan Jelas, Gina Rüter, Hanjo Hautow, Mario Sänger, Maryam Habibi, Marit Zettwitz, Till de Bortoli, Leonie Ostermann, Jurica Ševa, Johannes Starlinger, Oliver Kohlbacher, Nisar P Malek, Ulrich Keilholz, and Ulf Leser. 2021. Annotation and initial evaluation of a large annotated german oncological corpus. *JAMIA Open*, 4(2).

Manuel Lentzen, Sumit Madan, Vanessa Lage-Rupprecht, Lisa Kühnel, Juliane Fluck, Marc Jacobs, Mirja Mittermaier, Martin Witzenrath, Peter Brunecker, Martin Hofmann-Apitius, Joachim Weber, and Holger Fröhlich. 2022. Critical assessment of transformer-based ai models for german clinical notes. *JAMIA Open*, 5(4).

Dominique Makowski and Viliam Simko. 2018. neuropsychology/psycho.r: 0.2.8.

Aurelie Mascio, Zeljko Kraljevic, Daniel Bean, Richard Dobson, Robert Stewart, Rebecca Bendayan, and Angus Roberts. 2020. Comparative analysis of text classification approaches in electronic health records. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*. Association for Computational Linguistics.

Luise Modersohn, Stefan Schulz, Christina Lohr, and Udo Hahn. 2022. *GRASCCO — The First Publicly Shareable, Multiply-Alienated German Clinical Text Corpus*. IOS Press.

MTSamples. 2025. Transcribed medical transcription sample reports and examples. Accessed: 31 January 2025.

Mariana L. Neves, Daniel Butzke, Antje Dörendahl, Nora Leich, Benedikt Hummel, Gilbert Schönfelder, and Barbara Grune. 2019. Overview of the CLEF ehealth 2019 multilingual information extraction. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019*, volume 2380 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011a. Scikit-learn: Machine learning in python. Accessed: 2025-02-18.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011b. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint*.

Robin Rojowiec, Benjamin Roth, and Maximilian Fink. 2020. Intent recognition in doctor-patient interviews. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 702–709, Marseille, France. European Language Resources Association.

Niloofer Shanavas, Hui Wang, Zhiwei Lin, and Glenn Hawe. 2020. Ontology-based enriched concept graphs for medical document classification. *Information Sciences*, 525:172–181.

Hanna Suominen, Liadh Kelly, Lorraine Goeuriot, and Martin Krallinger. 2020. *CLEF eHealth Evaluation Lab 2020*, pages 587–594. Springer International Publishing.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. 2020. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 23–30. Association for Computational Linguistics.

Bo Zhou, Dingling Su, and Zehui Qu. 2021. Medical text classification system based on deep learning. In *2021 International Conference on Intelligent Computing, Automation and Applications (ICAA)*, pages 388–392. IEEE.

A Appendix

Classifier	**Parameter**	**Word2Vec**	**Sentence-BERT**	**BERT$_{ger}$**	**BioBERT**	**MedBERT**
SVM	Kernel Type	poly	poly	poly	poly	rbf
	Kernel Degree	4	3	4	2	2
	Cost	10	1	1	10	1
	Gamma	scale	scale	scale	scale	0.01
	Coef0	0	0.5	0.5	0.5	0
RF	Bootstrap	False	False	False	False	False
	Max Depth	20	20	20	10	20
	Max Features	sqrt	sqrt	sqrt	sqrt	sqrt
	Min Samples Leaf	1	2	1	4	2
	Min Samples Split	5	5	2	10	2
	n Estimators	1500	500	500	500	500
LightGBM	Num Leaves	31	31	31	31	31
	n Estimators	1000	2000	2000	1000	1000
	Learning Rate	0.01	0.1	0.1	0.01	0.01
CatBoost	Depth	6	6	6	8	6
	Iterations	1000	1000	3000	1000	1000
	Learning Rate	0.01	0.1	0.01	0.01	0.01
XGBoost	Max Depth	8	8	8	6	8
	n Estimators	2000	1000	1000	2000	1000
	Learning Rate	0.01	0.01	0.01	0.01	0.01

Table 8: Optimized hyperparameters of ML-models based on grid search for different embeddings.

Classifier with Sentence-BERT	**Mean Pooling**	**Max Pooling**	**CLS Token**
CatBoost	**0.8372**	0.5242	0.6858
RandomForest	**0.8121**	0.6375	0.7671
XGBoost	**0.8059**	0.5130	0.7105
SVM	**0.8330**	0.6768	0.7196
LightGBM	**0.8107**	0.4799	0.7205

Table 9: Weighted F1-Scores for ML-models using different extraction strategies on the IntRec validation dataset.