Roman Christof
5057566
Joint Degree:
Sociology/Philosophy, Bachelor of Arts
Computer Science, Bachelor of Science
13. Semester of Study
s5562603@stud.uni-frankfurt.de

Bachelor Thesis

HospLetExtractor

A Pipeline for Automated Analysis of German Hospital Letters

Roman Christof

Submission Date: 28.03.2023

Text Technology Lab
Goethe-Universität Frankfurt am Main
Prof. Dr. Alexander Mehler

Bitte dieses Formular zusammen mit der Abschlussarbeit abgeben!

Erklärung zur Abschlussarbeit

gemäß § 35, Abs. 16 der Ordnung für den Bachelorstudiengang Informatik
vom 17. Juni 2019:

Hiermit erkläre ich

____

(Nachname, Vorname)

Die vorliegende Arbeit habe ich selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel verfasst.

Ich bestätige außerdem, dass die vorliegende Arbeit nicht, auch nicht auszugsweise, für eine andere Prüfung oder Studienleistung verwendet wurde.

Zudem versichere ich, dass alle eingereichten schriftlichen gebundenen Versionen meiner vorliegenden Bachelorarbeit mit der digital eingereichten elektronischen Version meiner Bachelorarbeit übereinstimmen.

Frankfurt am Main, den

____

Unterschrift der/des Studierenden

Abstract

This bachelor thesis developed a pipeline for automatic processing of scanned hospital letters: HospLetExtractor. Hospital letters can contain valuable information about potential adverse drug reactions and useful case information relevant to pharmacovigilance. To make this data accessible, this thesis presents a pipeline consisting of image pre-processing, optical character recognition and post-processing. Pre-processing deskews the images, removes lines and rectangles, reduces noise and applies super-resolution. For the post-processing a spell checking system was set up including a newly built word frequency dictionary for german medical terms based on a created corpus of german medical texts. Furthermore, classical and deep learning models for the classification of hospital letters were compared, in which the transformer-based models performed best. In order to train and test the models, a new gold standard was created. By making these medical documents accessible for automatic analysis, hopefully a contribution can be made to expand the scope of pharmacovigilance.

Acknowledgement

I would especially like to thank Prof. Dr. Alexander Mehler, whose lectures and seminars in the area of Natural Language Processing introduced me to this exciting field and provided me with help and inspiration while working on the topic. Furthermore, I would like to thank the Paul-Ehrlich-Institut and Dr. Dirk Mentzer, who made this cooperation possible, especially Farnaz Zeidi, Manuela Messelhäußer and Dr. Liam Childs for your help and the very productive collaboration. I also want to thank all the colleagues who contributed to the correction of hospital letters for the gold standard.

Model and Data Availability Statement

The gold standard and the parts of the dictionary marked with ”closed” are not openly accessible and are owned by the Paul-Ehrlich-Institut, as are the developed models.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

This thesis was conducted in cooperation with the Paul-Ehrlich-Institut (PEI). The PEI is a German federal regulatory authority reporting to the Federal Ministry of Health. It is responsible for marketing authorization, batch release and continuous assessment of benefit and harm (pharmacovigilance) of biomedicines and vaccines in Germany[1]. The PEI receives reports of suspected adverse drug reactions according to legal requirements from marketing authorization holders and physicians, e.g. the Medicinal Products Act[2]. Consumers are also encouraged to report suspected adverse reactions to the PEI. In particularly relevant cases, the PEI also receives hospital or doctor letters. The manual processing of such reports is very time-consuming. However, they often contain valuable information about adverse drug reactions and case information. To automate the processing of such reports would significantly save time and could expand the scope of pharmacovigilance. However, this poses several challenges. Most reports are only available as scanned documents and cannot be processed immediately. In addition, they are not available collectively, so they must first be classified. Since very sensitive health data is involved, external software can't be used if it processes or stores the data outside of PEI to ensure full data protection. Another difficulty is that many models and datasets are only available in English. And if they are available in German, they are often not freely accessible due to the sensitive nature of health data. This limits the use of such systems and makes it necessary to implement the models independently.

1.2 Contributions

The thesis presented here contributes to the utilization of scanned hospital letters through a pipeline consisting of pre-processing, optical character recognition (OCR), post-processing and classification: HospLetExtractor (Hospital Letter Extractor). It comprises the following components:

1. A gold standard for training and testing the OCR process (including pre- and post-processing) as well as classification models.
2. An image pre-processing pipeline.
3. Post-processing via spell checking including a newly built word frequency dictionary for German medical terms based on a created corpus of German medical texts.

[1] `https://www.pei.de/SharedDocs/Downloads/EN/institute/paul-ehrlich-institut-profile-en.pdf?__blob=publicationFile&v=5`, accessed Mar. 20, 2023

[2] `https://www.gesetze-im-internet.de/englisch_amg/`, accessed Mar. 20, 2023

4. A comparison of classical and deep learning models for the classification of hospital letters.

These models are not only applicable to PEI data. They could also support the work of hospitals or other organizations on their way to digitizing their work processes.

1.3 Outline

The thesis starts by explaining the theoretical background of the important building blocks of this work. For this purpose, first some general remarks on optical character recognition are made, and then image pre-processing steps to improve the OCR result are discussed. Since these results are still error-prone, text post-processing steps are then introduced. In the following part on document classification, classification models are first distinguished, and then different types of text representations are presented. In the second chapter, an overview of related work is given, which also deals with the use of hospital reports or related document types. For this purpose, scientific works from the field of English NLP as well as German NLP are presented and related to this thesis. The fourth chapter presents the gold standard created in this context, which serves not only as a basis for optimizing and testing OCR and related components (pre- and post-processing), but also for the classification models. The procedure for the creation of the dataset is explained, as well as the relevant language-related peculiarities of the medical text. The following chapter describes the implementation of the pipeline. This includes image pre-processing, OCR application, and text post-processing. To improve the post-processing step, a medical corpus and a corresponding word frequency dictionary are presented. The last part of the pipeline deals with document classification. In the sixth chapter, the individual pipeline components are evaluated and an error analysis is performed. An outlook on possible extensions of the work presented here is given in the next chapter, followed by a summary and a final assessment of the results in the last chapter.

2 Theoretical Background

This chapter lays out the theoretical foundation of this thesis. For this purpose, optical character recognition is discussed first, followed by image pre-processing and text post-processing, and finally document classification.

2.1 Optical Character Recognition

Optical character recognition (OCR) is the process of converting handwritten and printed text into machine-readable text (Chaudhuri et al. 2016). This allows scanned or photographed documents to be made available for further text processing steps. OCR consists of several stages (N. Islam, Z. Islam, and Noor 2016). The first is character segmentation, where the characters are isolated from each other so that they can be processed by the recognition system. This can be done for example with projection profiles or component analysis. The next step is to extract features from the separated characters. Point distributions, series expansions, transformations, and structural analysis are used to determine essential properties of the characters (Chaudhuri et al. 2016). In the last step, character classification, the extracted features are assigned to classes for each character. There are several techniques: template matching, statistical techniques, structural techniques and artificial neural networks. Often these approaches are combined (Chaudhuri et al. 2016).

2.2 Image Pre-processing

The OCR result is only as good as the actual image material. Therefore, the first step, image acquisition, where paper documents are digitized and converted into an image file is already decisive. The quality of the scan or photo is critical to the subsequent OCR result. Low image resolution, the angle of capture or scanning, existing noise, shadows and curvature can all make it difficult for the OCR engine to recognize letters. If the original image quality is insufficient, image pre-processing steps are required to achieve a reasonable OCR result.

For many of the further processing steps, it is a prerequisite that the image is grayscale, so this is the first transformation.

Sometimes the images are slightly tilted to the left or right, which makes it difficult for the OCR engine to detect the letters correctly. These rotated images can be straightened using deskewing techniques (Chaudhuri et al. 2016).

Some images have a low resolution and therefore have to be rescaled. When the image is enlarged, the additional pixels must be interpolated. Since classical methods do not take the context of the pixels into account, the images are often blurred (Lat and Jawahar 2018). To avoid this, deep learning models such as Convolutional Neural Networks (CNNs) are also used in this area. For example Fast Super-Resolution Convolutional Neural Network (FS-RCNN) (Dong, Loy, and Tang 2016) or Efficient Sub-pixel Convolutional Neural Network

(ESPCN) (Shi et al. 2016), which extends the CNN approach with an efficient sub-pixel convolutional layer.

Methods such as erosion, dilation, median filtering, Gaussian blur, or Gaussian pyramid based on histogram analysis can be applied to filter noise from images (Badoiu, Ciobanu, and Craitoiu 2016).

Another technique often used for this purpose is binarization. In binarization, the image is converted into black and white pixels based on a threshold. One of the most commonly used methods is the Otsu binarization. It is automatically calculates a global threshold from a histogram and applies it to the whole image (Sezgin and Sankur 2004). The problem with such global thresholds is that parts of images with different lighting conditions are often blacked out. There are specialized binarization techniques for text documents such as Niblack thresholding (Niblack 1986) and Sauvola thresholding (Sauvola and Pietikäinen 2000), which are so called adaptive thresholds. They do not assume a global threshold, but determine a local threshold based on the pixel environment. As mentioned above, it can be applied to reduce noise and correct for different lighting conditions, but it also carries the risk of affecting topological features of letters (Chaudhuri et al. 2016).

2.3 Text Post-processing

Because OCR results often still contain errors, despite the image processing, text post-processing is necessary. This is an important step because subsequent tasks suffer from error-prone OCR results, as Nguyen et al. (2021) and van Strien. et al. (2020) have shown.

The task of post-processing can be stated as follows: For a given document d consisting of a sequence of tokens $d = t_1 t_2 ... t_n$, we get through processing the document with OCR a result of $d' = t'_1 t'_2 ... t'_m$, where it's not necessarily true that $n = m$ because of omissions and segmentation errors. If $t'_1 \neq t_1$ then t'_1 is an error. The task is to compare every token of the OCR output with the original document, detect errors and correct them, so $d = d'$ (Nguyen et al. 2021).

The following types of errors can be observed: segmentation errors (e.g. 'Palpi tationen' instead of 'Palpitationen'), misrecognized characters (e.g. 'Kermspintomografie' instead of 'Kernspintomografie'), punctuation errors, omissions (complete words are missing), order errors (lines and thus words are swapped), artifacts (additional characters) (modified list from Kumar 2016).

The errors can be further distinguished into two categories: non-word errors and real-word errors. For non-word errors it is the case, that $t'_1 \notin D$, where D is a given dictionary of words, whereas in the case of real-word errors, $t'_1 \in D$ (Nguyen et al. 2021).

There exist a variety of techniques for post-processing. This is only a brief, not necessarily complete, overview. Besides the manual approach (correction by humans), they can be broadly divided into two branches (Nguyen et al. 2021): (1) isolated-word approaches and (2) context-dependent approaches. For the isolated-word approach the first option is to combine the results of different OCR engines to take advantage of the fact that these models make different mistakes. Additionally a voting system has to be implemented to find the best result. For example, Long Short-Term Memory (LSTM) and Conditional Random Fields (CRF) can be used. Secondly, one could use lexical approaches, which are based on a dictionary or word frequency list. Typically the Damerau-Levenshtein edit distance is used to measure the distance between potentially erroneous words and dictionary entries in order to provide an appropriate correction. The Damerau-Levenshtein finds the minimum number of operations to transform a given string into a target string, where the operations are insertions, deletions, substitutions, transpositions of single characters (Boytsov 2011). The underlying dictionaries play a crucial role, since they decide which words can be corrected and how many false corrections will occur. Smith (2011) not only shows how insufficient dictionaries can lead to a deterioration of the result, but also questions the assumption that the probability of word occurrence is related to the correctness of a word. Nevertheless, the authors consider this to be a good approximation. In contrast to the isolated-word approaches context-dependent approaches can detect and correct the real-word errors mentioned above. An example of this branch are sequence-to-sequence models, which formulate the task of post-processing as one of machine translation and use for example text-to-text transformers (Nguyen et al. 2021).

2.4 Document Classification

The task of text classification is to assign a given document to a predefined category. For each d_i in a set of documents $D = \{d_1, d_2, ..., d_n\}$, where n is the total number of documents and $i \in \{1, ..., n\}$, a class c_j in $C = \{c_1, c_2, ..., c_m\}$ is assigned, where $j \in \{1, ..., m\}$.

In the here relevant case of binary classification $m = 2$, thus the document can belong to one of two classes.

2.4.1 Classification

This is only a small selection of possible models that can be used for classification. They include classifiers that are relevant to this thesis.

2.4.1.1 Naive Bayes

Naive Bayes is based on Bayes' theorem , which states the conditional probability of X given Y as follows (Anandarajan, Hill, and Nolan 2019):

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \tag{2.1}$$

The underlying assumption is, that given a class the different features are independent of each other. This contributes to simplification, but does not take into account that features can be interrelated.

The class C that most likely corresponds to a document D is then computed by:

$$\hat{C} = argmaxP(D|C)P(C) \tag{2.2}$$

It is the product of the conditional probability of a document given a class and the probability of that class.

2.4.1.2 k-Nearest Neighbors

k-Nearest Neighbors (kNN) is a method to compute the neighborhood relations of data based on a similarity metric. K is to be defined beforehand in order to then determine the k nearest neighbors. In the case of document classification for each document its similarity to all other documents is computed, which results in a neighborhood of this document. Similarity measures can be for example the Euclidean distance, the Manhattan distance, the Hamming distance or the Minkowski distance. The most probable class can then be determined by weighted or majority voting between the adjacent document classes (Guo et al. 2006).

2.4.1.3 Decision Trees and Random Forests

Decision trees split the classes through a descending structure of rules. Each node represents a test criterion by which the data to be classified is further partitioned. (Mitchell 1997) The aim is to achieve a higher purity, i.e. a homogeneity of the split data. Various methods for splitting exist that specify when a new node is introduced, each with a different definition of purity. Some important approaches are mentioned here. The Gini index ”measures the divergence between probability distributions”, ”entropy measures the homogeneity” and ”Chi-Square Test measures the likelihood of a split” (Anandarajan, Hill, and Nolan 2019 p. 142).

The concept of Random Forests is an extension of the decision tree method. A set of decision trees is built and then the documents are classified by voting or averaging on this collection.

2.4.1.4 Deep Neural Networks

In this approach a neural network is used for classification. It consists of three layers: an input layer, a hidden layer and an output layer. The input layer holds the text features which will be described in the next section. It is connected to the first hidden layer. The hidden layers consist of nodes whose activation depends on a previously defined activation function (e.g. sigmoid, ReLU). They are connected via weights that determine their relevance to the network and are optimized during training. All following hidden layers are connected to the previous and subsequent layers. The last hidden layer is then connected to the output layer, which in classification is often a softmax function that gives the probability value for the different classes. Figure 2.1 shows the different layers and how they are connected. The model is trained with a back-propagation algorithm (Kowsari et al. 2019).

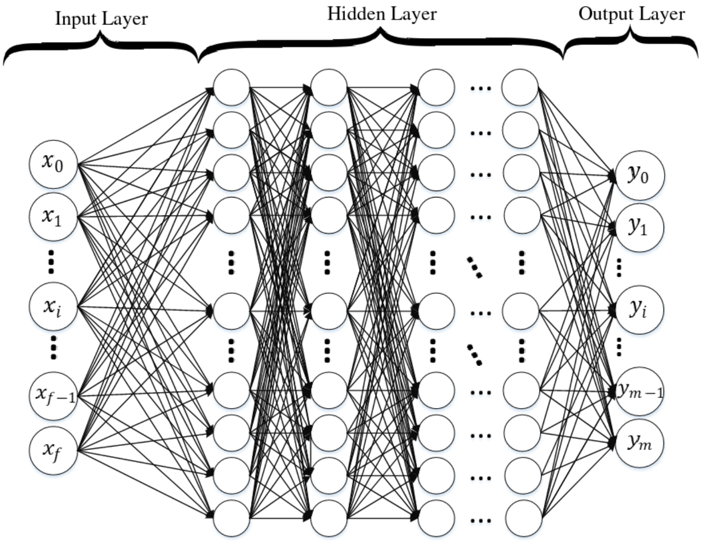


Figure 2.1: Fully connected deep neural network; taken from Kowsari et al. 2019, p 35

In practice, often only a linear layer is used, i.e. a neural network consisting of only one hidden layer.

2.4.2 Text Representations

The above mentioned classifiers need a text representation as an input. The following is a brief overview of different approaches to text representation that are relevant to this work or serve to better understand the methods used here.

2.4.2.1 Bag-of-words and Term Frequency-Inverse Document Frequency

Bag-of-words is a simple approach to transform a text into a vector representation. Its vectors have the length of the number of words based on the underlying vocabulary. The vector counts with each entry the frequency for every word. This type of encoding has the shortcomings that it does not take into account any relationship between the words in terms of their meaning and it produces very large vectors as the vocabulary size grows (Kowsari et al. 2019).

Term frequency-inverse document frequency (tfidf) is another way, which builds on the concept of bag-of-words, to obtain a representation of word and text by assessing the importance of specific words in a text. It is the product of the term frequency and the inverse document frequency:

$$tfidf_{i,j} = tf_{i,j} \cdot idf_i \tag{2.3}$$

where i is a term, j the document, $tf_{i,j}$ the frequency of i for j. The inverse document frequency is calculated by: $idf_i = \log(\frac{n}{1+df_i}) + 1$, where n equal the total number of documents in the corpus and df_i the document frequency of i, counting the number of documents in which i occurs. In this way, words that occur frequently in most of the texts are ranked lower, since they do not contribute to the separation of the texts. On the other hand, words that appear in only a few documents are rated higher. (Anandarajan, Hill, and Nolan 2019)

2.4.2.2 Word2vec

To overcome the shortcomings of simple word representations such as bag-of-words, neural network based approaches have been developed.

One prominent approach that uses a neural network architecture, is presented here: word2vec (Mikolov et al. 2013).

This method consists of two models. With the first model, continuous bag-of-words (CBOW), a neural network with one hidden layer is trained to predict a word based on the context. With the second model, continuous skip-gram, it learns the reverse, to predict the context based on a word.

The model trained in this way, learns a vector space in which the vectors of words with a similar meaning are close together.

2.4.2.3 Contextualized Word Representations

This method was originally developed by Melamud, Goldberger, and Dagan (2016) as context2vec. It uses the architecture of bidirectional long short-term memory (biLSTM). The word vectors are obtained by a forward and backward language model (Kowsari et al. 2019). This results in a left-to-right and right-to-left context word embedding, which is then combined with a multilayer perceptron (Melamud, Goldberger, and Dagan 2016). This idea has been extended by ELMo (Embeddings from Language Models) (Peters et al. 2018) with deep contextualized word representations that integrate not only syntax and semantics but also model polysemy. It also uses the bidirectional language models mentioned above and then maximizes the log likelihood of both directions.

2.4.2.4 Transformer-based Models

Transformer models have been triumphant in many fields in recent years, including natural language processing. One very successful architecture in the field of NLP is BERT, which stands for Bidirectional Encoder Representations from Transformer (Devlin et al. 2019). It can be used to learn word representations. The architecture of the model consists of a bidirectional Transformer encoder with multiple layers, which is build on the attention and self-attention mechanisms introduced by Vaswani et al. (2017). The attention mechanism allows the model to learn the relevance of each data point depending on its relevance to the task. This can be used to determine which words in a sentence are more important than others.

The model weights are trained using two different methods. The first is called masked language model, where some tokens of a text sequence are randomly masked and have to be predicted by the model. In the second task, next sentence prediction, the model has to to predict the next sentence based on the previous one. The pre-training is done on very large text corpora (about 3.3 billion words) (Devlin et al. 2019).

Another key advantage of the Transformer is that its computations can be highly parallelized due to a reduction of sequential computations (Vaswani et al. 2017). Computational dependencies are the bottleneck of traditional Recurrent Neural Networks (RNNs). This is what makes it possible to process such enormous amounts of training data.

The BERT representation can be fine-tuned for specialized tasks. The pre-trained parameters are used as a starting point and the weights of the top layers are retrained with additional data.

3 Related Work

This chapter gives a brief overview of existing work that presents methods for utilizing (scanned) medical documents.

In the field of English medical NLP, numerous approaches exist to make medical texts usable, mostly Electronic Health Records (EHRs). Exemplary, some particularly relevant ones are mentioned here.

The paper "Deep learning-based NLP data pipeline for EHR-scanned document information extraction" (Hsu et al. 2022) attempts to solve a similar problem to the one presented in this thesis, but for English medical reports. It is aimed at a more specific application, since it involves the classification of scanned reports in electronic health records with respect to a particular sleep disorder. Prior to classification, the documents are pre-processed and then OCRed into machine-processable text. In addition, classical machine learning models are evaluated and compared with deep learning models for classification. This thesis extends the pre-processing step with deskewing, noise reduction, super-resolution, includes a post-processing step with spell correction and trains models for German hospital letters.

Goodrum, Roberts, and Bernstam (2020) also utilize scanned documents from EHRs. For this purpose, models are developed to classify the documents (radiology reports, clinical correspondence, etc.). To prepare the documents for the classification models tesseract is used for OCR with a previous pre-processing that includes erosion and contrast enhancement. For post-processing a spell checker (SymSpell) based on an English word frequency dictionary is used. This thesis also evaluates this method for post-processing and builds its own word frequency dictionary for medical German language. In addition, simpler approaches such as Bag of Words combined with Naive Bayes, Logistic Regression and Random Forests are compared with deep learning models. The results show that a specialized BERT variant called ClinicalBERT performs best. However, this model is only available for the English language, so in this thesis different BERT variants are fine-tuned on German hospital letters.

In the area of German medical NLP, there are also several papers that try to utilize medical text data. Some particularly relevant ones will be mentioned here as well. Many of them already assume machine-processable text and do not deal with OCR, such as the paper "Critical assessment of transformer-based AI models for German clinical notes" (Lentzen et al. 2022). They built a new annotated dataset of clinical notes (not openly accessible) and used existing medical text corpora (BRONCO150 (Kittner et al. 2021), CLEF eHealth 2019 Task 1 (Kelly et al. 2019), GGPONC (Borchert et al. 2020), and JSynCC (Lohr, Buechel, and Hahn 2018)), only one of which is openly available (CLEF eHealth dataset). Due to the inaccessibility of the datasets, a gold standard is created in this thesis. In addition, they trained and fine-tuned several transformer-based models for document classification and named entity recognition. They conclude that even general-purpose language models perform well on clinical NLP tasks. Nevertheless, the fine-tuning of models can improve the results. Whereas model training from scratch has proven to be insufficient.

In contrast König et al. (2019) apply a rule- and knowledge-based approach with a medical ontology with the goal of detecting specific drug-disease interactions in discharge letters. They partially leverage scanned hospital letters, but proprietary software is used for OCR without any mentioned pre-processing. In contrast, in this thesis a publicly available OCR system tesseract is used.

The paper ”A Medical Information Extraction Workbench to Process German Clinical Text” (Roller et al. 2022) presents a new German medical corpus based on documents from a nephrology division (not open). It includes annotated clinical notes and discharge letters for NER and Relation Extraction. They compare different models for these tasks, extended with POS tagging and concept detection. Character based embeddings (like Flair) showed good results for clinical text. The authors assumes that the reason for this is the specialized language of clinical reports, which is characterized by many words of Greek and Latin origin.

4 Gold Standard for OCR and Classification

This chapter presents the gold standard for OCR and the pre- and post-processing component as well as classification and explains how it was created.

The starting point in terms of data was a collection of different types of documents, which in the case of hospital letters consist mainly of scanned letters. In addition to these letters, the relevant document types include non hospital letters, such as autopsy reports, laboratory reports, reporting forms (shown in figure 4.2)[3], letter and e-mail extracts. It was only possible to draw in part on documents that had already been classified. These included reporting forms and e-mail extracts, all other document types were not available in classified form, especially the hospital letters. Corrected OCR-processed data was not available for any of the documents. Thus, the first step was to establish a gold standard for the classification of hospital reports and their OCR processing.

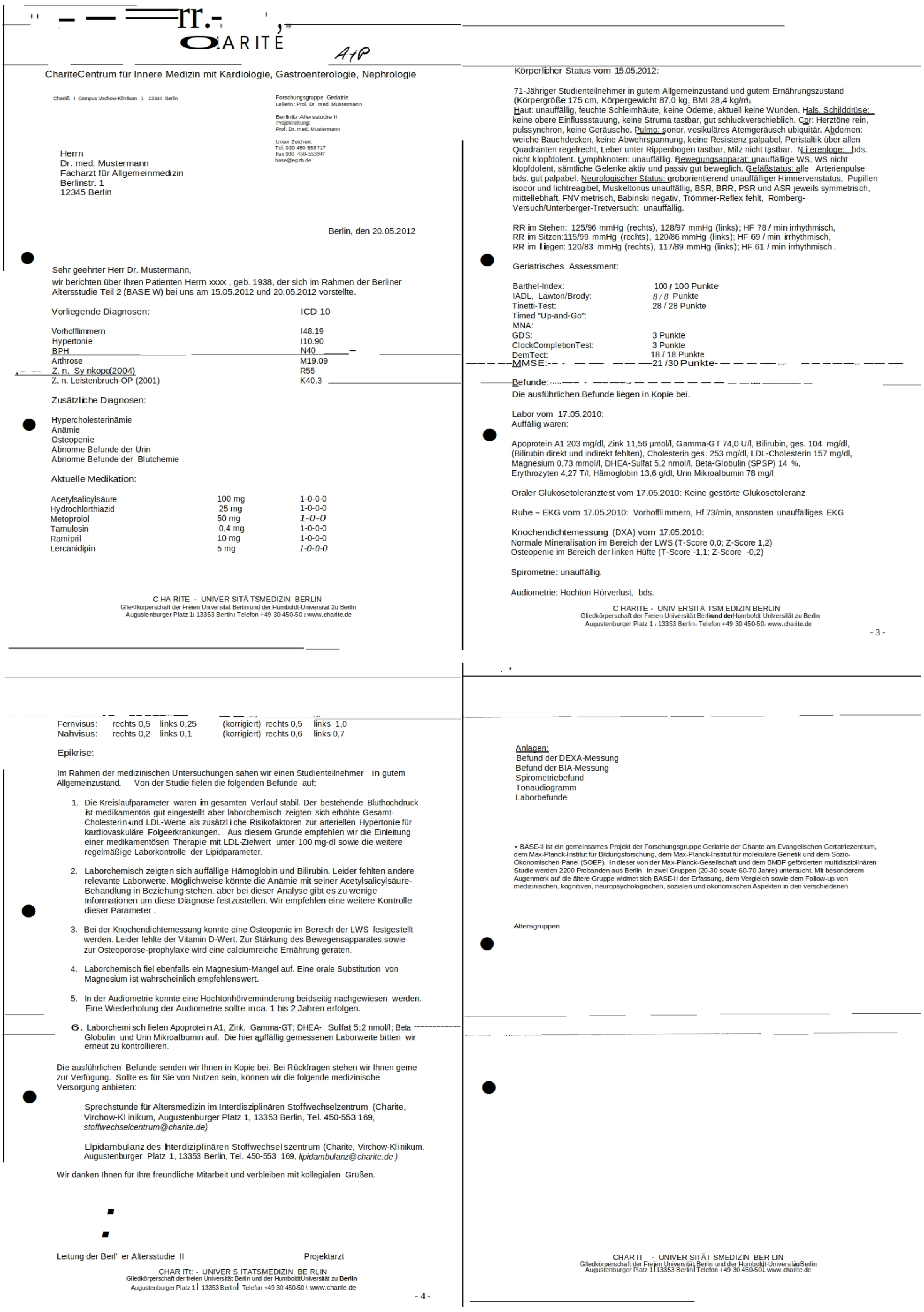
4.1 Data Collection

Primitive classifiers were trained to reduce the time required for manual classification. First a rule-based classifier based on text length and keywords (based on a manual review of the documents) and additionally a Naive Bayes model with term frequency-inverse document frequency (tf-idf) as the text representation were trained. In the absence of training data, a dataset presented in the context of the "2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records" (Joulin et al. 2017) was used. The dataset contains 505 English discharge summaries from the MIMIC-III (Medical Information Mart for Intensive Care-III) clinical care database[4]. Discharge summaries, as clinical reports of hospital stays and therapies, are textually similar to German hospital reports. The summaries were translated using the fairseq wmt19 transformer[5] (Ng et al. 2019). The German translation was previously used by Frei and Kramer (2022) to train a medical named entity recognition model called GERNERMED. Based on the indicators of the rule-based classifier and the Naive Bayes classifier, the results were manually validated. This method was used to aggregate 250 hospital letters with a total of 800 pages (most of the laboratory data was excluded due to complex tables). Tesseract (Kay 2007) was then used to convert the PDFs via OCR into machine-readable text. The next step was to correct the OCR errors to obtain a ground truth for optimizing the OCR, pre- and post-processing.

[3] `https://www.pei.de/DE/arzneimittelsicherheit/pharmakovigilanz/meldeformulare-online-meldung/meldeformulare-online-meldung-node.html`, accessed Mar. 20, 2023

[4] `https://portal.dbmi.hms.harvard.edu/projects/n2c2-2018-t2/,accessedMar.20,2023`

[5] `https://huggingface.co/facebook/wmt19-de-en`, accessed Mar. 20, 2023



CHARITÉ

A+P

ChariteCentrum für Innere Medizin mit Kardiologie, Gastroenterologie, Nephrologie

Charitß I Campus Virchow-Kllnlkum 1 13344 Berlin

Forschungsgruppe Geriatrie
Lelierin: Prol. Dr. med. Mustermann

Berlln&r Allersstudie II
Projektleltung:
Prof. Dr. med. Mustermann

Unser Zeichen:
Tel. 030 450-553717
Fax 030 450-553947
base@egzb.de

Herrn
Dr. med. Mustermann
Facharzt für Allgemeinmedizin
Berlinstr. 1
12345 Berlin

Berlin, den 20.05.2012

Sehr geehrter Herr Dr. Mustermann,

wir berichten über Ihren Patienten Herrn xxxx , geb. 1938, der sich im Rahmen der Berliner Altersstudie Teil 2 (BASE W) bei uns am 15.05.2012 und 20.05.2012 vorstellte.

Vorliegende Diagnosen:	ICD 10
Vorhofflimmern	I48.19
Hypertonie	I10.90
BPH	N40
Arthrose	M19.09
Z. n. Sy nkope(2004)	R55
Z. n. Leistenbruch-OP (2001)	K40.3

Zusätzliche Diagnosen:

Hypercholesterinämie
Anämie
Osteopenie
Abnorme Befunde der Urin
Abnorme Befunde der Blutchemie

Aktuelle Medikation:

Acetylsalicylsäure	100 mg	1-0-0-0
Hydrochlorthiazid	25 mg	1-0-0-0
Metoprolol	50 mg	*1-0-0*
Tamulosin	0,4 mg	1-0-0-0
Ramipril	10 mg	1-0-0-0
Lercanidipin	5 mg	*1-0-0-0*

C HA RITE - UNIVER SITÄ TSMEDIZIN BERLIN
Glle<Ikörperschaft der Freien Universität Bertin und der Humboldt-Universität 2u Bertln
Auguslenburger Platz 1ı 13353 Bertinı Telefon +49 30 450-50 ı www.charite.de

Körperlicher Status vom 15.05.2012:

71-Jähriger Studienteilnehmer in gutem Allgemeinzustand und gutem Ernährungszustand (Körpergröße 175 cm, Körpergewicht 87,0 kg, BMI 28,4 kg/m²). Haut: unauffällig, feuchte Schleimhäute, keine Ödeme, aktuell keine Wunden. Hals, Schilddrüse: keine obere Einflussstauung, keine Struma tastbar, gut schluckverschieblich. Cor: Herztöne rein, pulssynchron, keine Geräusche. Pulmo: sonor. vesikuläres Atemgeräusch ubiquitär. Abdomen: weiche Bauchdecken, keine Abwehrspannung, keine Resistenz palpabel, Peristaltik über allen Quadranten regelrecht, Leber unter Rippenbogen tastbar, Milz nicht tastbar. N i erenloge: bds. nicht klopfdolent. Lymphknoten: unauffällig. Bewegungsapparat: unauffällige WS, WS nicht klopfdolent, sämtliche Gelenke aktiv und passiv gut beweglich. Gefäßstatus: alle Arterienpulse bds. gut palpabel. Neurologischer Status: groborientierend unauffälliger Himnervenstatus, Pupillen isocor und lichtreagibel, Muskeltonus unauffällig, BSR, BRR, PSR und ASR jeweils symmetrisch, mittellebhaft. FNV metrisch, Babinski negativ, Trömmer-Reflex fehlt, Romberg-Versuch/Unterberger-Tretversuch: unauffällig.

RR im Stehen: 125/96 mmHg (rechts), 128/97 mmHg (links); HF 78 / min irrhythmisch,
RR im Sitzen:115/99 mmHg (rechts), 120/86 mmHg (links); HF 69 / min irrhythmisch,
RR im liegen: 120/83 mmHg (rechts), 117/89 mmHg (links); HF 61 / min irrhythmisch .

Geriatrisches Assessment:

Barthel-Index:	100 / 100 Punkte
IADL, Lawton/Brody:	*8 / 8* Punkte
Tinetti-Test:	28 / 28 Punkte
Timed "Up-and-Go":	
MNA:	
GDS:	3 Punkte
ClockCompletionTest:	3 Punkte
DemTect:	18 / 18 Punkte
MMSE:	21 /30 Punkte

Befunde:
Die ausführlichen Befunde liegen in Kopie bei.

Labor vom 17.05.2010:
Auffällig waren:

Apoprotein A1 203 mg/dl, Zink 11,56 µmol/l, Gamma-GT 74,0 U/l, Bilirubin, ges. 104 mg/dl, (Bilirubin direkt und indirekt fehlten), Cholesterin ges. 253 mg/dl, LDL-Cholesterin 157 mg/dl, Magnesium 0,73 mmol/l, DHEA-Sulfat 5,2 nmol/l, Beta-Globulin (SPSP) 14 %, Erythrozyten 4,27 T/l, Hämoglobin 13,6 g/dl, Urin Mikroalbumin 78 mg/l

Oraler Glukosetoleranztest vom 17.05.2010: Keine gestörte Glukosetoleranz

Ruhe – EKG vom 17.05.2010: Vorhoffli mmern, Hf 73/min, ansonsten unauffälliges EKG

Knochendichtemessung (DXA) vom 17.05.2010:
Normale Mineralisation im Bereich der LWS (T-Score 0,0; Z-Score 1,2)
Osteopenie im Bereich der linken Hüfte (T-Score -1,1; Z-Score -0,2)

Spirometrie: unauffällig.

Audiometrie: Hochton Hörverlust, bds.

C HARITE - UNIV ERSITÄ TSM EDIZIN BERLIN
Gliedkörperschaft der Freien Universität Berliund derHumbo!dt Unlversilät zu Berlin
Augustenburger Platz 1 ı 13353 Berlinı Telefon +49 30 450-50ı www.charite.de

- 3 -

Fernvisus:	rechts 0,5	links 0,25	(korrigiert)	rechts 0,5	links 1,0
Nahvisus:	rechts 0,2	links 0,1	(korrigiert)	rechts 0,6	links 0,7

Epikrise:

Im Rahmen der medizinischen Untersuchungen sahen wir einen Studienteilnehmer in gutem Allgemeinzustand. Von der Studie fielen die folgenden Befunde auf:

1. Die Kreislaufparameter waren im gesamten Verlauf stabil. Der bestehende Bluthochdruck ist medikamentös gut eingestellt aber laborchemisch zeigten sich erhöhte Gesamt·Cholesterin-und LDL-Werte als zusätzl i che Risikofaktoren zur arteriellen Hypertonie für kardiovaskuläre Folgeerkrankungen. Aus diesem Grunde empfehlen wir die Einleitung einer medikamentösen Therapie mit LDL-Zielwert unter 100 mg-dl sowie die weitere regelmäßige Laborkontrolle der Lipidparameter.
2. Laborchemisch zeigten sich auffällige Hämoglobin und Bilirubin. Leider fehlten andere relevante Laborwerte. Möglichweise könnte die Anämie mit seiner Acetylsalicylsäure-Behandlung in Beziehung stehen. aber bei dieser Analyse gibt es zu wenige Informationen um diese Diagnose festzustellen. Wir empfehlen eine weitere Kontrolle dieser Parameter .
3. Bei der Knochendichtemessung konnte eine Osteopenie im Bereich der LWS festgestellt werden. Leider fehlte der Vitamin D-Wert. Zur Stärkung des Bewegensapparates sowie zur Osteoporose-prophylaxe wird eine calciumreiche Ernährung geraten.
4. Laborchemisch fiel ebenfalls ein Magnesium-Mangel auf. Eine orale Substitution von Magnesium ist wahrscheinlich empfehlenswert.
5. In der Audiometrie konnte eine Hochtonhörverminderung beidseitig nachgewiesen werden. Eine Wiederholung der Audiometrie sollte inca. 1 bis 2 Jahren erfolgen.
6. Laborchemi sch fielen Apoprotei n A1, Zink, Gamma-GT; DHEA- Sulfat 5;2 nmol/l; Beta Globulin und Urin Mikroalbumin auf. Die hier auffällig gemessenen Laborwerte bitten wir erneut zu kontrollieren.

Die ausführlichen Befunde senden wir Ihnen in Kopie bei. Bei Rückfragen stehen wir Ihnen geme zur Verfügung. Sollte es für Sie von Nutzen sein, können wir die folgende medizinische Versorgung anbieten:

Sprechstunde für Altersmedizin im Interdisziplinären Stoffwechselzentrum (Charite, Virchow-Kl inikum, Augustenburger Platz 1, 13353 Berlin, Tel. 450-553 169, *stoffwechselcentrum@charite.de)*

Llpidambul anz des Interdiziplinären Stoffwechsel szentrum (Charite, Virchow-Kli nikum. Augustenburger Platz 1, 13353 Berlin, Tel. 450-553 169, *lipidambulanz@charite.de)*

Wir danken Ihnen für Ihre freundliche Mitarbeit und verbleiben mit kollegialen Grüßen.

Leitung der Berl' er Altersstudie II — Projektarzt

CHAR ITt: - UNIVER S ITATSMEDIZIN BE RLIN
Gliedkörperschaft der freien Universität Berlin und der HumboldtUniversität zu **Berlin**
Augustenburger Platz 1 | 13353 Berlin | Telefon +49 30 450-50 \ www.chanle.de

- 4 -

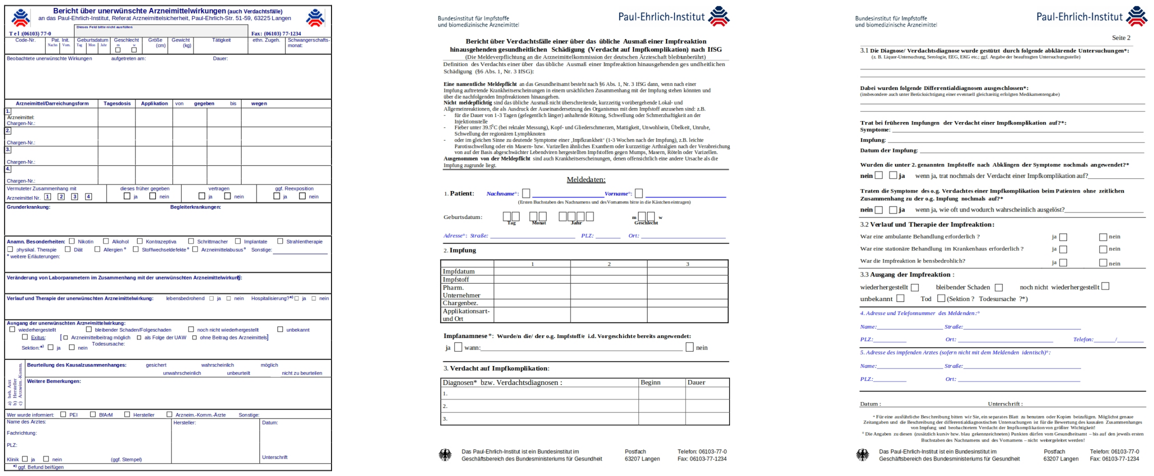
Anlagen:
Befund der DEXA-Messung
Befund der BIA-Messung
Spirometriebefund
Tonaudiogramm
Laborbefunde

• BASE-II ist ein gemeinsames Projekt der Forschungsgruppe Geriatrie der Chante am Evangelischen Gertatriezentrum, dem Max-Planck-Institut für Bildungsforschung, dem Max-Planck-Institut für molekulare Genetik und dem Sozio-Ökonomischen Panel (SOEP). In dieser von der Max-Planck-Gesellschaft und dem BMBF geförderten multidisziplinären Studie werden 2200 Probanden aus Berlin in zwei Gruppen (20-30 sowie 60-70 Jahre) untersucht. Mit besonderem Augenmerk auf die ältere Gruppe widmet sich BASE-II der Erfassung, dem Vergleich sowie dem Follow-up von medizinischen, kognitiven, neuropsychologischen, sozialen und ökonomischen Aspekten in den verschiedenen

Altersgruppen .

CHAR IT - UNIVER SITÄT SMEDIZIN BER LIN
Glledkörperschaft der Frejen Universität Berlin und der Humboldt-Universilät zu Berlin
Augustenburger Platz 1 | 13353 Berlin | Telefon +49 30 450-50 | www.charite.de

Figure 4.1: Example of an artificial hospital letter (König et al. 2019, supplementary information)



Bericht über unerwünschte Arzneimittelwirkungen (auch Verdachtsfälle)
an das Paul-Ehrlich-Institut, Referat Arzneimittelsicherheit, Paul-Ehrlich-Str. 51-59, 63225 Langen

T e l :(06103) 77-0 Dieses Feld bitte nicht ausfüllen Fax: (06103) 77-1234

Code-Nr.	Pat. Init. Nachn Vorn.	Geburtsdatum Tag Mon Jahr	Geschlecht m □ w □	Größe (cm)	Gewicht (kg)	Tätigkeit	ethn. Zugeh.	Schwangerschafts-monat:

Beobachtete unerwünschte Wirkungen aufgetreten am: Dauer:

Arzneimittel/Darreichungsform	Tagesdosis	Applikation	von gegeben	bis	wegen
1. Arzneimittel: Chargen-Nr.:					
2. Chargen-Nr.:					
3. Chargen-Nr.:					
4. Chargen-Nr.:					

Vermuteter Zusammenhang mit Arzneimittel Nr. 1 2 3 4 | dieses früher gegeben □ ja □ nein | vertragen □ ja □ nein | ggf. Reexposition □ ja □ nein

Grunderkrankung: **Begleiterkrankungen:**

Anamn. Besonderheiten: □ Nikotin □ Alkohol □ Kontrazeptiva □ Schrittmacher □ Implantate □ Strahlentherapie
□ physikal. Therapie □ Diät □ Allergien ° □ Stoffwechseldefekte ° □ Arzneimittelabusus ° Sonstige: ______
° weitere Erläuterungen:

Veränderung von Laborparametern im Zusammenhang mit der unerwünschten Arzneimittelwirkung*):

Verlauf und Therapie der unerwünschten Arzneimittelwirkung: lebensbedrohend □ ja □ nein Hospitalisierung?*) □ ja □ nein

Ausgang der unerwünschten Arzneimittelwirkung:
□ wiederhergestellt □ bleibender Schaden/Folgeschaden □ noch nicht wiederhergestellt □ unbekannt
□ Exitus: [□ Arzneimittelbeitrag möglich □ als Folge der UAW □ ohne Beitrag des Arzneimittels]
Sektion:*) □ ja □ nein Todesursache:

a) beh. Arzt b) Hersteller c) Arzneim.-Komm.

Beurteilung des Kausalzusammenhanges: gesichert wahrscheinlich möglich unwahrscheinlich unbeurteilt nicht zu beurteilen

Weitere Bemerkungen:

Wer wurde informiert: □ PEI □ BfArM □ Hersteller □ Arzneim.-Komm.-Ärzte Sonstige:

Name des Arztes:
Fachrichtung:
PLZ:
Klinik □ ja □ nein (ggf. Stempel)

Hersteller:

Datum:
Unterschrift

*) ggf. Befund beifügen

Bundesinstitut für Impfstoffe
und biomedizinische Arzneimittel

Paul-Ehrlich-Institut

Bericht über Verdachtsfälle einer über das übliche Ausmaß einer Impfreaktion hinausgehenden gesundheitlichen Schädigung (Verdacht auf Impfkomplikation) nach IfSG
(Die Meldeverpflichtung an die Arzneimittelkommission der deutschen Ärzteschaft bleibtunberührt)

Definition des Verdachts einer über das übliche Ausmaß einer Impfreaktion hinausgehenden ges undheitlichen Schädigung (§6 Abs. 1, Nr. 3 IfSG):

Eine namentliche Meldepflicht an das Gesundheitsamt besteht nach §6 Abs. 1, Nr. 3 IfSG dann, wenn nach einer Impfung auftretende Krankheitserscheinungen in einem ursächlichen Zusammenhang mit der Impfung stehen könnten und über die nachfolgenden Impfreaktionen hinausgehen.
Nicht meldepflichtig sind das übliche Ausmaß nicht überschreitende, kurzzeitig vorübergehende Lokal- und Allgemeinreaktionen, die als Ausdruck der Auseinandersetzung des Organismus mit dem Impfstoff anzusehen sind: z.B.

- für die Dauer von 1-3 Tagen (gelegentlich länger) anhaltende Rötung, Schwellung oder Schmerzhaftigkeit an der Injektionsstelle
- Fieber unter 39.5^{0}C (bei rektaler Messung), Kopf- und Gliederschmerzen, Mattigkeit, Unwohlsein, Übelkeit, Unruhe, Schwellung der regionären Lymphknoten
- oder im gleichen Sinne zu deutende Symptome einer ‚Impfkrankheit' (1-3 Wochen nach der Impfung), z.B. leichte Parotisschwellung oder ein Masern- bzw. Varizellen ähnliches Exanthem oder kurzzeitige Arthralgien nach der Verabreichung von auf der Basis abgeschwächter Lebendviren hergestellten Impfstoffen gegen Mumps, Masern, Röteln oder Varizellen.

Ausgenommen von der Meldepflicht sind auch Krankheitserscheinungen, denen offensichtlich eine andere Ursache als die Impfung zugrunde liegt.

Meldedaten:

1. **Patient:** *Nachname°:* □ ______ *Vorname°:* □ ______
(Ersten Buchstaben des Nachnamens und des Vornamens bitte in die Kästchen eintragen)

Geburtsdatum: □□ Tag □□ Monat □□□□ Jahr m □□ w Geschlecht

Adresse°: Straße: ______ *PLZ:* ______ *Ort:* ______

2. **Impfung**

	1	2	3
Impfdatum			
Impfstoff			
Pharm. Unternehmer			
Chargenbez.			
Applikationsart- und Ort			

Impfanamnese *: Wurde/n die/ der o.g. Impfstoff/e i.d. Vorgeschichte bereits angewendet:
ja □ wann: ______ □ nein

3. **Verdacht auf Impfkomplikation:**

Diagnosen* bzw. Verdachtsdiagnosen :	Beginn	Dauer
1.		
2.		
3.		

Das Paul-Ehrlich-Institut ist ein Bundesinstitut im
Geschäftsbereich des Bundesministeriums für Gesundheit
Postfach
63207 Langen
Telefon: 06103-77-0
Fax: 06103-77-1234

Bundesinstitut für Impfstoffe
und biomedizinische Arzneimittel

Paul-Ehrlich-Institut

Seite 2

3.1 **Die Diagnose/ Verdachtsdiagnose wurde gestützt durch folgende abklärende Untersuchungen*:**
(z. B. Liquor-Untersuchung, Serologie, EEG, EKG etc.; ggf. Angabe der beauftragten Untersuchungsstelle)

Dabei wurden folgende Differentialdiagnosen ausgeschlossen*:
(insbesondere auch unter Berücksichtigung einer eventuell gleichzeitig erfolgten Medikamentengabe)

Trat bei früheren Impfungen der Verdacht einer Impfkomplikation auf?*:
Symptome: ______
Impfung: ______
Datum der Impfung: ______

Wurden die unter 2. genannten Impfstoffe nach Abklingen der Symptome nochmals angewendet?*
nein □ □ **ja** wenn ja, trat nochmals der Verdacht einer Impfkomplikation auf?______

Traten die Symptome des o.g. Verdachtes einer Impfkomplikation beim Patienten ohne zeitlichen Zusammenhang zu der o.g. Impfung nochmals auf?*
nein □ □ **ja** wenn ja, wie oft und wodurch wahrscheinlich ausgelöst?______

3.2 **Verlauf und Therapie der Impfreaktion:**
War eine ambulante Behandlung erforderlich ? ja □ □ nein
War eine stationäre Behandlung im Krankenhaus erforderlich ? ja □ □ nein
War die Impfreaktion le bensbedrohlich? ja □ □ nein

3.3 **Ausgang der Impfreaktion :**
wiederhergestellt □ bleibender Schaden □ noch nicht wiederhergestellt □
unbekannt □ Tod □ (Sektion ? Todesursache ?*)

4. Adresse und Telefonnummer des Meldenden:°
Name: ______ *Straße:* ______
PLZ: ______ *Ort:* ______ *Telefon:* ______/______

5. Adresse des impfenden Arztes (sofern nicht mit dem Meldenden identisch)°:
Name: ______ *Straße:* ______
PLZ: ______ *Ort:* ______

Datum : Unterschrift :

* Für eine ausführliche Beschreibung bitten wir Sie, ein separates Blatt zu benutzen oder Kopien beizufügen. Möglichst genaue Zeitangaben und die Beschreibung der differentialdiagnostischen Untersuchungen ist für die Bewertung des kausalen Zusammenhanges von Impfung und beobachtetem Verdacht der Impfkomplikation von größter Wichtigkeit!
° Die Angaben zu diesen (zusätzlich kursiv bzw. blau gekennzeichneten) Punkten dürfen vom Gesundheitsamt – bis auf den jeweils ersten Buchstaben des Nachnamens und des Vornamens – nicht weitergeleitet werden!

Das Paul-Ehrlich-Institut ist ein Bundesinstitut im
Geschäftsbereich des Bundesministeriums für Gesundheit
Postfach
63207 Langen
Telefon: 06103-77-0
Fax: 06103-77-1234

Figure 4.2: Examples for report forms (they include handwritten or typed text, which can't be shown due to privacy reasons).

4.2 Guidlines and Procedure for Correction

In order to make the data processing transparent and to have a common basis for several correctors, guidelines for correcting were developed. They include the following rules:

- Wrongly recognized words should be corrected and artefacts removed
- Wrongly ordered lines should be corrected
- The correction should be done from left to right, in case of contiguous text blocks and several columns, one should start with the higher standing text block
- Tables should be corrected row by row, multi-line entries should stay together
- Page number and meta information should be included
- Hand written text should be included
- Upside down text from meta information should be removed
- Stamps should not be included
- Additional spaces and new lines do not need to be corrected
- Spelling errors in the original text should not be corrected

4.3 Peculiarities of the Data

There is no standardized layout for hospital letters. However, there are characteristics that many hospital letters have in common. The reports are divided into different sections with corresponding headings (diagnosis, preliminary diagnoses, admission findings, anamnesis, epicrisis, medication, therapy, laboratory data, etc.). The reports do not always contain all sections and the order also varies. Some of the reports contain tables of medications and laboratory data, but mostly they consist of continuous text. For privacy reasons, no real hospital letter can be shown. Instead, an artificially created hospital letter is shown in figure 4.1. Besides their format, hospital letters are characterized by their special language. Medical and biomedical terms of Latin and Greek origin are used, mixed with abbreviations and medical slang. The sentences are often choppy, contain many enumerations and sequences of observations and judgments. One example:

> Nierenloge: bds. nicht klopfdolent. Lymphknoten: unauffällig. Bewegungsapparat: unauffällige WS, Pupillen isocor und lichtreagibel ... Thrombozytenaggregationshemmendes Medikament verschrieben.

The quality of the scanned documents varies greatly. Some are completely clean and with high resolution, while others are blurry, have noise and are skewed.

4.4 The Dataset

The gold standard dataset includes 1,600 classified pages (800 hospital letters, 800 non hospital letters) and 210 hospital letter pages with corrected OCR. The latter were obtained from the previously classified 800 hospital letter pages (an overview is given in table 4.1). The individual pages of the hospital reports have an average of 238 words and 1654 characters.

The data is stored in json format (shown in 1, 2)

```
[{
    "file": "File_x_y_z.pdf"
    "ocr" : "Sehr geehrter Herr Dr. Mustermann, wir berichten
                 Ihnen über den o.g. Patienten ...",
    "hospitalLetter" : 1
}, ...]
```

Listing 1: Classification dataset (contains partly OCR errors)

```
[{
    "file": "File_x_y_z.pdf"
    "correctedOCR" : "Sehr geehrter Herr Dr. Mustermann, wir berichten
                Ihnen über den o.g. Patienten ..."
}, ...]
```

Listing 2: OCR dataset

Table 4.1: Gold standard dataset for training and testing

Type	**Task**	**Number (Pages)**
Classified documents: hospital letters and non hospital letters	Classification	1600 (800 each class)
Corrected OCRed hospital letters	OCR, pre- and post-processing	210

5 Pipeline Implementation

This chapter describes and explains the individual components of the pipeline. The pipeline is shown in the figure 5.1. It starts with image pre-processing, which prepares the images for the OCR engine. These are then converted into machine-processable text using Tesseract. The result is corrected in the next step of post-processing using a dictionary-based approach. This text is then classified as either a hospital letter or a non-hospital letter. Several classification models are presented. The pipeline is implemented in Python.

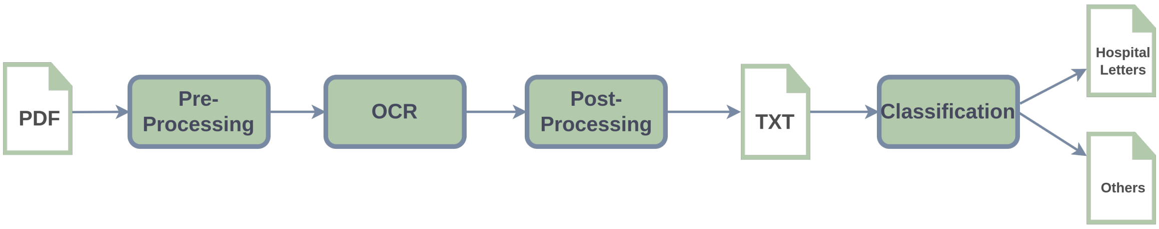


Figure 5.1: Overview of the whole pipeline

5.1 Image Pre-Processing

All the steps of image pre-processing are illustrated in the figure 5.3. It includes gray scaling and deskewing, line and rectangle removal, noise reduction and super resolution.

It was mainly implemented using the Python package of the Open Source Computer Vision Library (OpenCV)[6].

First, the image is gray scaled because the OCR engine has been trained on images processed in this way and the following processes require this step.

Since OCR only works well when the images are aligned straight, the second step is to correct any existing skew. For this purpose, the skew angle is determined by calculating the contours present in the image. To improve the recognition of the contours, dilation is first applied, which enlarges existing morphological structures. The angle of the largest contour is determined. This angle is then used to rotate the entire image.

In the third step, the lines and black rectangles are removed from the image. This is done because, it has bene observed that lines of tables and those that sometimes arise as by-products of the scanning process often cause artifacts when applying OCR. The same applies for black rectangles that are inserted in course of the anonymization of hospital reports.

To detect the lines and rectangles, the image is first inverted and then the morphological operations erosion and dilation are applied based on rectangle shape as the structuring element. The image is then binarized with a binary and OTSU threshold. Finally, all detected pixels belonging to a line or rectangle are set to white.

[6] `https://pypi.org/project/opencv-python/`, accessed Mar. 20, 2023

Since some documents contain noise, this is removed in a third step. However, this is only done if such noise has been detected, since the overall image quality, especially the clarity of morphological structures can be negatively affected. The potential noise is computed based on the histogram of the image. The threshold was determined by analyzing images with and without noise. If noise is detected, dilation and erosion is first performed with an identical kernel. In this way, the important morphological structures are preserved and noise related pixels disappear. After applying a bilateral filter, the image is then processed using Sauvola binarization.

In the final step super resolution is applied. Because the application of super resolution degrades the OCR result for images with good resolution and clear letters, it is only employed below a certain threshold. As all documents in the underlying dataset have already been resized to a more or less uniform size, the blur is used as the threshold. This is done by calculating the Laplacian of the image, as this is an indicator of well-defined edges. The threshold value is determined by piror examination of blurred and unblurred images. If the value falls below the threshold, super resolution is performed by ESPCN (Efficient sub-pixel Convolutional Neural Network). Figure 5.2 shows the images after each pre-processing step.

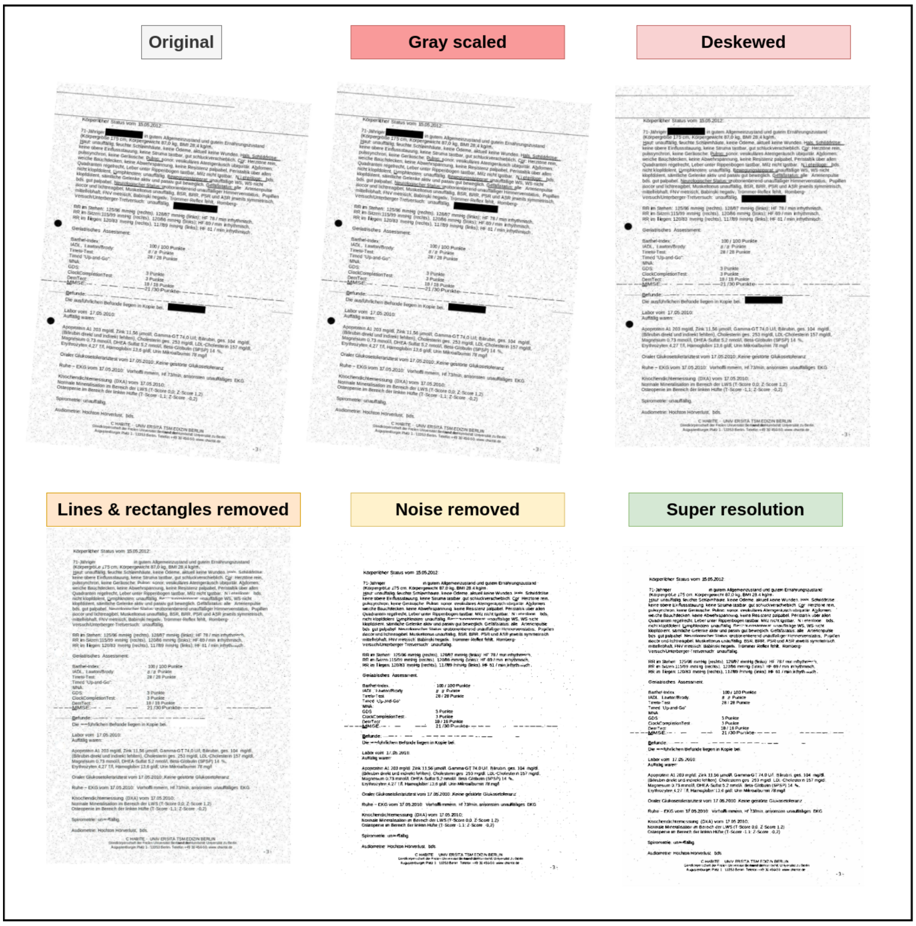


Figure 5.2: Example of applied pre-processing steps (based on artificially degraded hospital letter showed in figure 4.1).

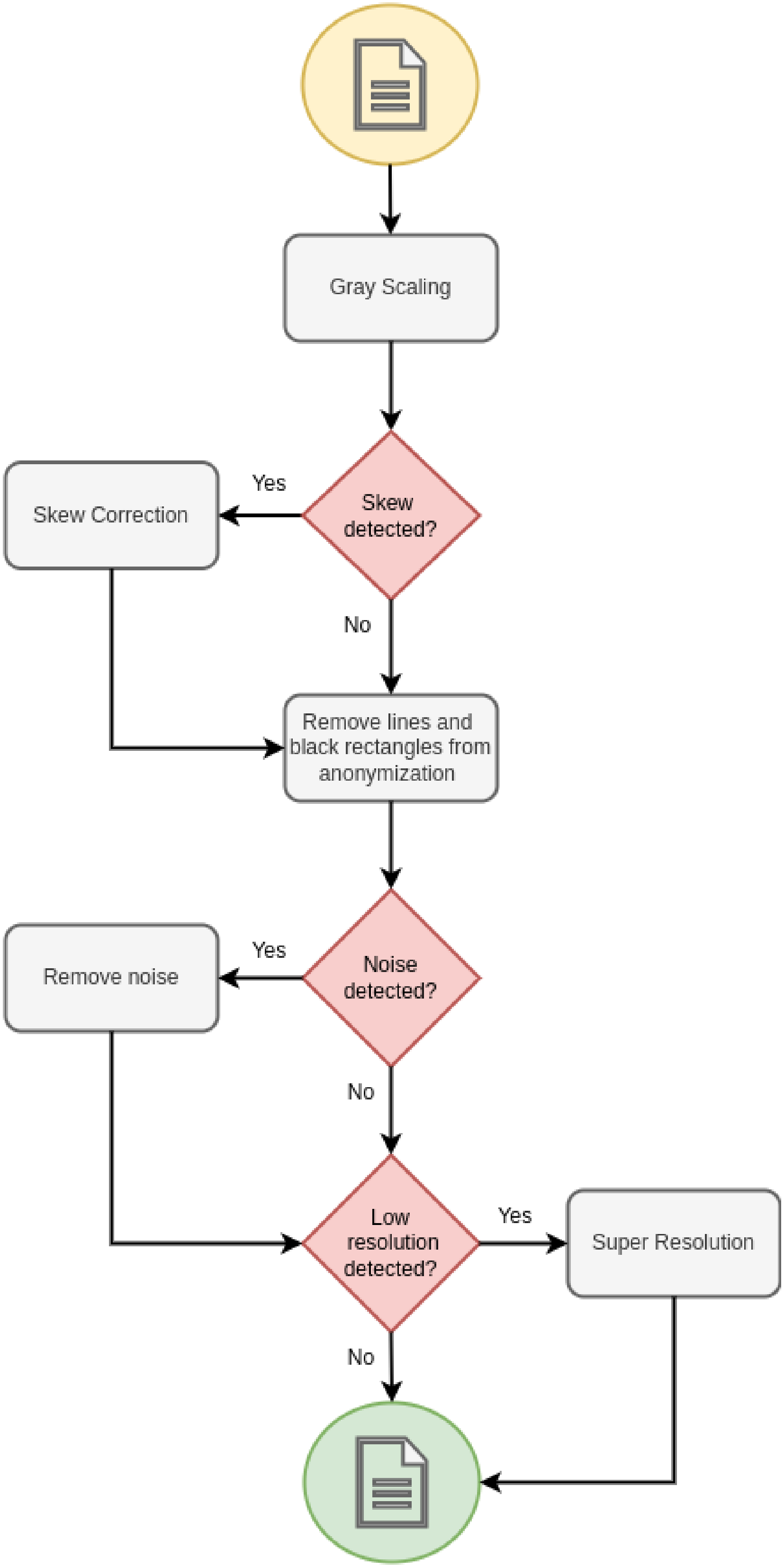


Figure 5.3: Pre-processing pipeline

5.2 Optical Character Recognition

Tesseract (Kay 2007) in version 5.2.0 was used for OCR. It consists of a long short-term memory (LSTM) neural network. The LSTM architecture is based on OCROpus[7] and consists of a forward LSTM and a backward LSTM, which are processed by a standard 1x1 convolution and passed to a softmax nonlinearity[8]. Tesseract provides several modes of page segmentation. The fully automatic page segmentation showed in own experiments that it sometimes skips text parts. Therefore the sparse text option was used, which finds as much text as possible. Tesseract was implemented using the Python wrapper pytesseract[9].

5.3 Text Post-Processing

5.3.1 Spell Correction

A dictionary approach is used for post-processing. For this, the spelling correction algorithm from SymSpell (Garbe 2012) was tested in different configurations and with different dictionaries. It calculates the Damerau–Levenshtein distance as the editing distance between the input word and the dictionary words. SymSpell uses symmetric deletion, which reduces the operations transpose, replace and insert to deletions. This leads to a significant reduction in complexity regarding the generation of edit candidates (Garbe 2012). The candidates are then ranked according to their frequency given in the dictionary. Several correction strategies are compared: Lookup, which only corrects single words; compound aware multi-word spelling correction, which does not work on single words but also considers previous words and thus can detect and correct splitting and concatenation errors, and word segmentation, which inserts spaces for wrongly concatenated words. It must be noted, that since a dictionary approach was used, no real-world error can be detected. The lookup method can't detect segmentation errors either.

For the pipeline presented here, a Python implementation [10] of SymSpell is used.

All numbers, dates and punctuation were excluded from the correction. Based on an error analysis, frequently recurring errors have already been corrected beforehand by substitution (e.g. 'ug' to' μg', '0.g.' to 'o.g.', which stands for 'oben genannte/n').

Since the dictionary on which the correction is based does not contain all word inflections, there is a risk of wrong corrections. To mitigate this risk, a primitive suffix checker was integrated. The token correction was skipped, if the word was equal to an entry in the dictionary modulo an inflectional suffix -, -e. -es, -r, -s, -n (based on Strohmaier et al. 2003).

An abstract version of the whole process is shown in the algorithm 1.

[7] `https://github.com/ocropus/ocropy`, accessed Mar. 20, 2023

[8] `https://tesseract-ocr.github.io/docs/das_tutorial2016/6ModernizationEfforts.pdf`, accessed Mar. 20, 2023

[9] `https://github.com/madmaze/pytesseract`, accessed Mar. 20, 2023

[10] `https://github.com/mammothb/symspellpy`, accessed Mar. 20, 2023

```
Input: String $S_{ocr}$, OCR result
Output: String $S_{post}$, Result of spell corection
S = substituteCommonErrors($S_{ocr}$) ;
$S_{post}$ = S ;
tokens = nltkTokenizer.tokenize(S);
for token in tokens do
    correction = symspell.lookup(maxEditDistance=2, exclude=[numbers, dates]);
    if token == correction mod inflection suffix then
        corrected = token ;
    else if token mod inflection suffix == correction then
        corrected = token ;
    else
        corrected= correction ;
    end
    $S_{post}$ = replace($S_{post}$, token, corrected) ;
end
return $S_{post}$;
```

Algorithm 1: Dictionary-based Post-Processing

5.3.2 Medical Corpus for Correction

The dictionary is an integral part of this spell correction approach and the results depend strongly on the size and quality of the dictionary. Hospital letters are characterized by a particular medical language that must be reflected in the dictionary. To achieve this a German medical corpus was composed. It includes data from google books/hunspell[11] and medicine related Wikipedia articles[12], as well as PubMed abstracts[13]. PubMed is an English-language text-based meta-database of references to medical articles related to the entire field of biomedicine from the United States National Library of Medicine. In addition, the corpus contains non-technical summaries (NTS) of animal experiments[14] and ICD-10 terms in the German modification[15]. The ICD- 10 is the International Statistical Classification Of Diseases And Related Health Problems and represents the official classification system for diagnoses in Germany. The corpus also includes MedDRA terms[16]. MedDRA (Medicinal Dictionary for Regulatory Activities) is an internationally agreed medical dictionary for regulatory purposes which provides standardised medicinal terminology. Furthermore, the corpus also contains internal data from the PEI: data of drugs and reactions and hospital letter data.
From the gold standard, 70 percent of the corrected hospital reports were integrated into the

[11] https://github.com/wolfgarbe/SymSpell/tree/master/SymSpell.FrequencyDictionary, accessed Mar. 20, 2023

[12] https://de.wikipedia.org/wiki/Portal:Medizin, accessed Mar. 20, 2023

[13] https://pubmed.ncbi.nlm.nih.gov/, accessed Mar. 20, 2023

[14] https://www.openagrar.de/receive/openagrar_mods_00046540, accessed Mar. 20, 2023

[15] https://www.bfarm.de/DE/Kodiersysteme/Klassifikationen/ICD/ICD-10-GM/_node.html, accessed Mar. 20, 2023

[16] https://www.meddra.org/, accessed Mar. 20, 2023

dictionary. Spelling errors in the original were corrected. An overview is given in table 5.1. Not all sources are freely accessible. This is the case for MedDRA and the PEI data.

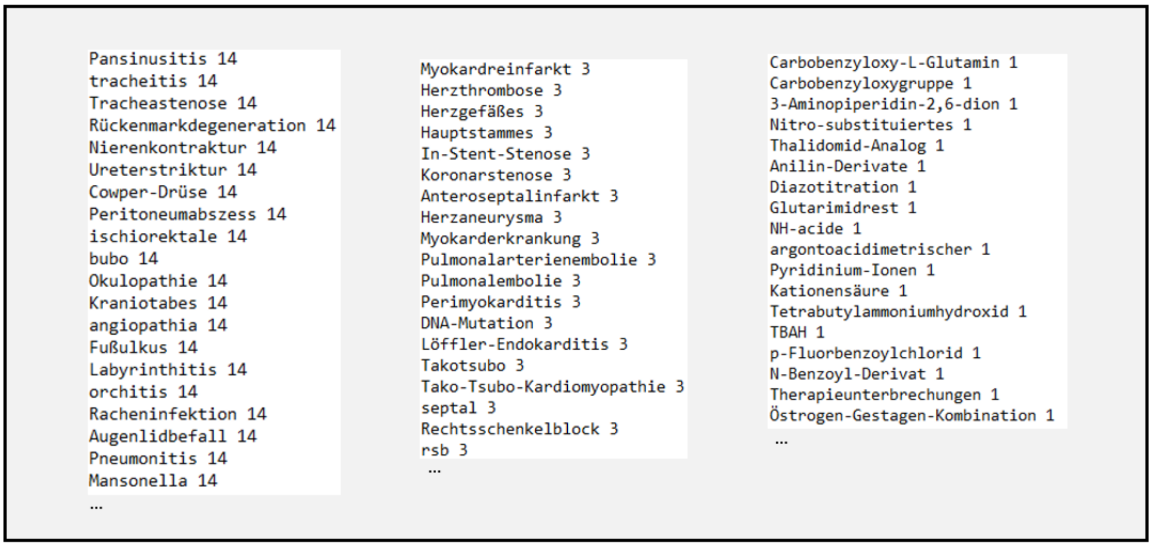
The PubMed data were accessed through the PubMed API called Entrez[17]. To access the API and load the data the biopython library[18] was used.

To access the Wikipedia API, the Python package Wikipedia-API[19] was employed. This was used to extract all articles of the Portal:Medizin. Afterwards, articles that were not relevant were manually removed (e.g. articles about medical personalities).

The other resources were downloaded from the links mentioned above.

The texts were then tokenized with nltk tokenizer[20], numbers, dates and punctuation were removed. For each corpus the frequency dictionary was computed and added together, except for the frequency dictionary of the hospital letters. For this dictionary, the frequencies were scaled by the highest frequency value of the combined frequencies of the other dictionaries. This is done to reflect the special importance of the vocabulary extracted from the hospital letters.

The generated dictionary contains nearly 700,000 terms in total. An exemplary extract from the dictionary is given in figure 5.4. The open dictionary contains 677,593 and the closed one 53,733 unique terms. 27,700 terms are exclusive to the closed dictionary. Figure 5.5 shows the intersection of the different dictionary sources. The open and closed data have a Jaccard Index of 0.0369. The Jaccard Index measures the similarity of two sets. 0.0 indicates that the sets are completely different and 1.0 indicates that the sets are identical (Liu 2006, p. 154).



```
Pansinusitis 14
tracheitis 14
Tracheastenose 14
Rückenmarkdegeneration 14
Nierenkontraktur 14
Ureterstriktur 14
Cowper-Drüse 14
Peritoneumabszess 14
ischiorektale 14
bubo 14
Okulopathie 14
Kraniotabes 14
angiopathia 14
Fußulkus 14
Labyrinthitis 14
orchitis 14
Racheninfektion 14
Augenlidbefall 14
Pneumonitis 14
Mansonella 14
...
```

```
Myokardreinfarkt 3
Herzthrombose 3
Herzgefäßes 3
Hauptstammes 3
In-Stent-Stenose 3
Koronarstenose 3
Anteroseptalinfarkt 3
Herzaneurysma 3
Myokarderkrankung 3
Pulmonalarterienembolie 3
Pulmonalembolie 3
Perimyokarditis 3
DNA-Mutation 3
Löffler-Endokarditis 3
Takotsubo 3
Tako-Tsubo-Kardiomyopathie 3
septal 3
Rechtsschenkelblock 3
rsb 3
...
```

```
Carbobenzyloxy-L-Glutamin 1
Carbobenzyloxygruppe 1
3-Aminopiperidin-2,6-dion 1
Nitro-substituiertes 1
Thalidomid-Analog 1
Anilin-Derivate 1
Diazotitration 1
Glutarimidrest 1
NH-acide 1
argontoacidimetrischer 1
Pyridinium-Ionen 1
Kationensäure 1
Tetrabutylammoniumhydroxid 1
TBAH 1
p-Fluorbenzoylchlorid 1
N-Benzoyl-Derivat 1
Therapieunterbrechungen 1
Östrogen-Gestagen-Kombination 1
...
```

Figure 5.4: Extract of the word frequency dictionary.

[17] https://www.ncbi.nlm.nih.gov/books/NBK25500/, accessed Mar. 20, 2023

[18] https://biopython.org/docs/1.75/api/Bio.Entrez.html, accessed Mar. 20, 2023

[19] https://pypi.org/project/Wikipedia-API/, accessed Mar. 20, 2023

[20] https://www.nltk.org/api/nltk.tokenize.html, accessed Mar. 20, 2023

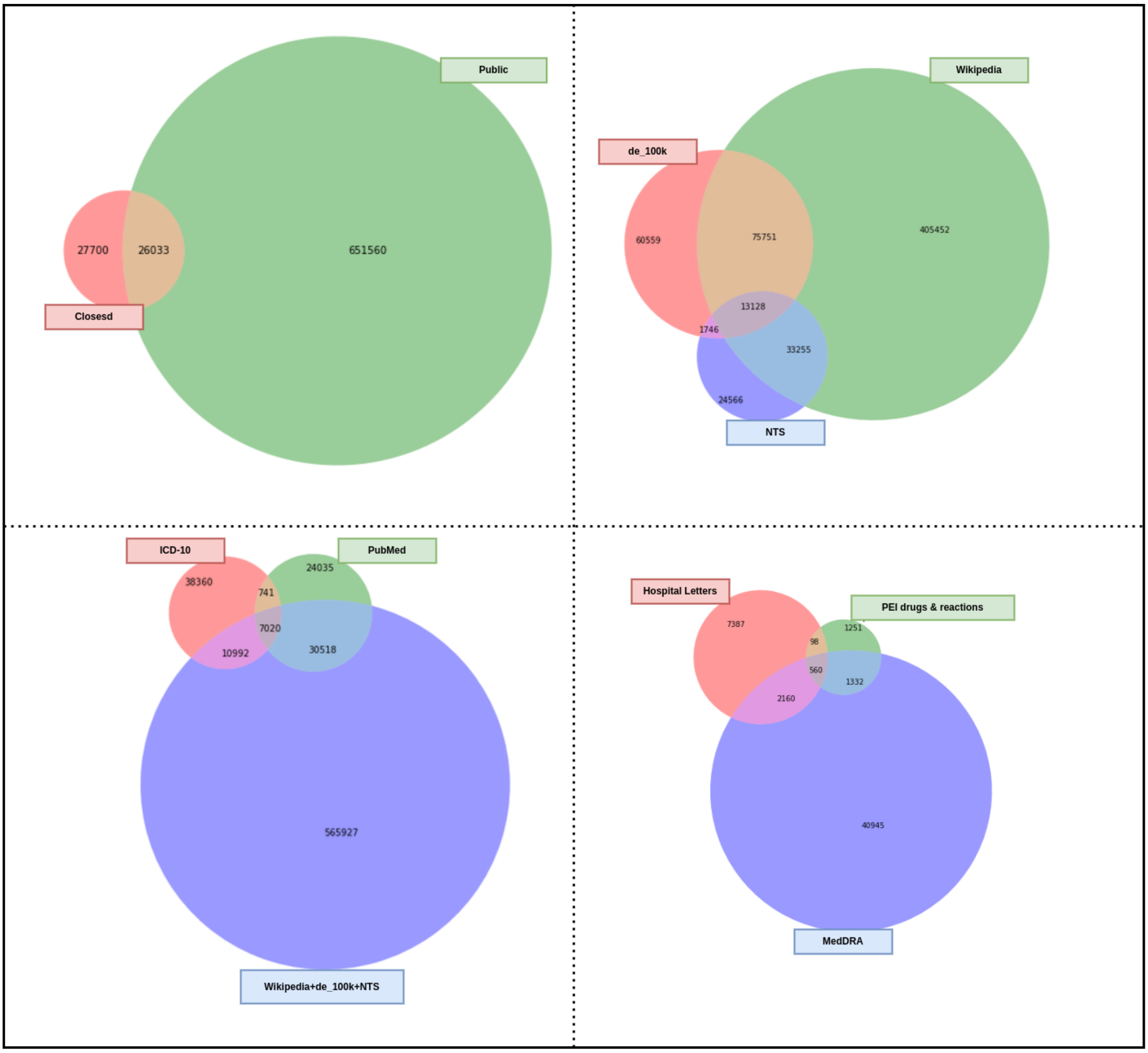


Figure 5.5: Overlaps of the different dictionary sources

.

Table 5.1: German medical corpus

Source	Number of terms	Description	Accessibility
de100k	100,000	Google books ngrams and hunspell	open
Wikipedia	476,405	Entries from Portal:Medizin and relevant categories	open
PubMed	62,316	Abstracts of scientific papers in german (3,240)	open
ICD-10	57,114	International Statistical Classification of Diseases, German Modification	open
NTS	123,882	Non-technical Summaries (NTS) of Animal Experiments Indexed with ICD-10 Codesv (8,386)	open
Sum	**677,593**	Unique terms	
MedDRA	35,368	Medical Dictionary for Regulatory Activities; terms without special german characters	closed
PEI drugs and reactions	3242	From PEI database	closed
PEI hospital letters	10,206	Corrected OCRed hospital letters from PEI database	closed
Sum	**44,042**	Unique terms	
Total Sum	**697,153**	Unique terms	

5.4 Document Classification

In this study three traditional methods (Naive Bayes, k-nearest neighbors, random forest, each with term frequency-inverse document frequency (tfidf) as text representation) were compared with fasttext, flair embeddings and three BERT-based models.

All classification models are trained and applied on a single page of hospital letters, not on entire letters. The reason for this is that many PDF documents are nested with different document types. The trained classifiers allow to resolve these entanglements and assign the pages to the appropriate class.

The data set of 1600 hospital and non-hospital letters was split into a training set (60 percent), development set (20 percent), and test set (20 percent).

The classical models were implemented using the machine learning library scikit-learn 1.2.1[21]. For every model a scikit-learn pipeline was built and the hyperparameters were tuned using grid search. TF-IDF was implemented via TfidfVectorizer.

[21] https://scikit-learn.org/, accessed Mar. 20, 2023

FastText[22] is a library for text representation and classification (Joulin et al. 2017), developed by the Facebook AI Research (FAIR) lab. It uses an embedding, which is partially solving the shortcomings of word2vec (Mikolov et al. 2013). While word2vec embeddings are limited to words and morphologies seen in the training data, fasttext embeddings use character n-grams. Through the modified skipgram model with bag of character n-grams, the word representations are not limited to the training data words and also include inflected forms (Bojanowski et al. 2017). This leads to fewer out-of-dictionary events. It is not only relevant for specialized language such as that used in hospital reports but it also suits well for noisy text. FastText uses a linear classifier for text classification. The softmax function calculates a probability value for each class. To reduce the computational effort, which arises from computing the probability value for each class, a hierarchical softmax is applied. A Huffman coding tree is built with the labels as leaves and the nodes represents the probability from the root this node. This allows a quick search for the appropriate label (does not have much effect in the case of binary classification) (Joulin et al. 2017). The parameters of the classification model with the character n-grams were fine-tuned using the built-in autotune functionality of fasttext[23].

In addition to FastText a classifier and embeddings from flair were trained. Flair[24] (Akbik, Bergmann, et al. 2019) is a NLP framework developed by Humboldt University of Berlin et al. Flair embeddings (Akbik, Blythe, and Vollgraf 2018) are word embeddings, which are based on character-level. Similar to FastText, here the character-level representation leads to an improved processing of texts which are error-prone due to OCR and spelling errors. They also take into account the surrounding words, which results in contextualized embeddings. The language model architecture consists of LSTMs, one that propagates the sequence information forward and one backward. Flair offers the functionality of stacked embeddings. With these it is possible to combine different embeddings through concatenation. The document representation is achieved via document pool embeddings. In this process a pooling operation (mean pooling) is applied to every word embedding to get the representation of the whole document. For this pipeline the pre-trained embeddings 'de-forward' and 'de-backword' (corpus based on web, wikipedia and subtitle data)[25] were stacked as the input for the document pool embeddings. They were further fine-tuned on the hospital letter dataset with a learning rate of 5.0e-5, a mini batch size of 8 and 20 epochs. The document pool embeddings were fine tuned with nonlinear transformation before the mean pooling step, because the baseline embeddings are not task specific. For the text classification part a linear layer predicts the class label based on the word representation.

For the transformer-based text representation the following models were tested and fine-tuned: GBERT, German-MedBERT and GottBERT.

GBERT[26] (Chan, Schweter, and Möller 2020) is a BERT-based language model trained on

[22] https://github.com/facebookresearch/fastText, accessed Mar. 20, 2023

[23] https://fasttext.cc/docs/en/autotune.html, accessed Mar. 20, 2023

[24] https://github.com/flairNLP/flair, accessed Mar. 20, 2023

[25] https://github.com/flairNLP/flair/blob/master/resources/docs/embeddings/FLAIR_EMBEDDINGS.md, accessed Mar. 20, 2023

[26] https://huggingface.co/deepset/gbert-large, accessed Mar. 20, 2023

the OSCAR dataset [27] dataset, which uses web crawled data from the Common Crawl[28]. It also includes data from OPUS[29] and Wikipedia.

German-MedBERT (Shrestha 2021)[30] is a fine-tuned version of GBERT (Chan, Schweter, and Möller 2020) on the NST-ICD dataset[31]. This dataset consists of non-technical summaries of animal experiment descriptions with International Statistical Classification of Diseases codes.

GottBERT (Scheible et al. 2020)[32] uses in contrast to German-MedBERT the RoBERTa architecture, which excludes the next sentence prediction for training and includes a dynamic masking process to change the masked tokens during epochs of training (Zhuang et al. 2021). GottBERT is also trained on a part of the OSCAR dataset.

The BERT-based text representations were fine-tuned on the hospital letter dataset with a learning rate of 5.0e-5, a mini batch size of 8 and 20 epochs. The flair framework was also used for this purpose. The three BERT variants were accessed through the transformer library HuggingFace[33] and implemented with the flair transformer document embeddings. The text classifier was trained analogously to the above described process for the flair embeddings.

Since the BERT models can only handle 512 tokens, in some cases tokens beyond that were truncated. This Paper (Sun et al. 2019) showed in experiments, that this trivial approach performs better than advanced hierarchical methods, which divide the text in fractions and combine the representations.

[27] `https://oscar-project.org/`, accessed Mar. 20, 2023

[28] `https://commoncrawl.org/the-data/`, accessed Mar. 20, 2023

[29] `https://opus.nlpl.eu/`, accessed Mar. 20, 2023

[30] `https://huggingface.co/smanjil/German-MedBERT`, accessed Mar. 20, 2023

[31] `https://www.openagrar.de/receive/openagrar_mods_00046540`, accessed Mar. 20, 2023

[32] `https://huggingface.co/LennartKeller/longformer-gottbert-base-8192-aw512`, accessed Mar. 20, 2023

[33] `https://huggingface.co/`, accessed Mar. 20, 2023

6 Evaluation

This chapter presents and evaluates the results of each pipeline step. In addition, a detailed error analysis of the individual components of pre- and post-processing is carried out. The character error rate (CER) and the word error rate (WER) were used to evaluate the OCR, pre- and post-processing (Carrasco 2014). The CER is derived from the Levenshtein distance and is defined as follows:

$$CER = \frac{S + D + I}{N} \tag{6.1}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions and N is the number of characters in the ground truth text. Where N can be calculated by $S + D + C$, where C is the number of correct characters. CER subsequently provides information about the percentage of incorrectly predicted characters.

The word error rate (WER) follows the same principle, but at the word level:

$$WER = \frac{S_w + D_w + I_w}{N_w} \tag{6.2}$$

where N_w is the number of words in the text, S_w the number of substituted words, D_w the number of deleted words, I_w the number of inserted words.

CER and WER were implemented using dinglehopper[34], which was developed in the context of the QURATOR[35] project with the participation of Prussian Cultural Heritage Foundation and the Berlin State Library. It was not only used to compute the errors, but also to visualize them, as shown in figure 6.1.

It is to be noted that the CER and. WER doesn't provide any information about the quality of the error. The errors can have consequences of varying severity regarding the accuracy of the documents. To compensate for this, an additional detailed error analysis will be provided in the next section.

For the evaluation of the classification models precision, recall and f1-score are used as metrics. They are defined as follows (Nguyen et al. 2021, p. 124:18):

$$Precision = \frac{TP}{TP + FP} \tag{6.3}$$

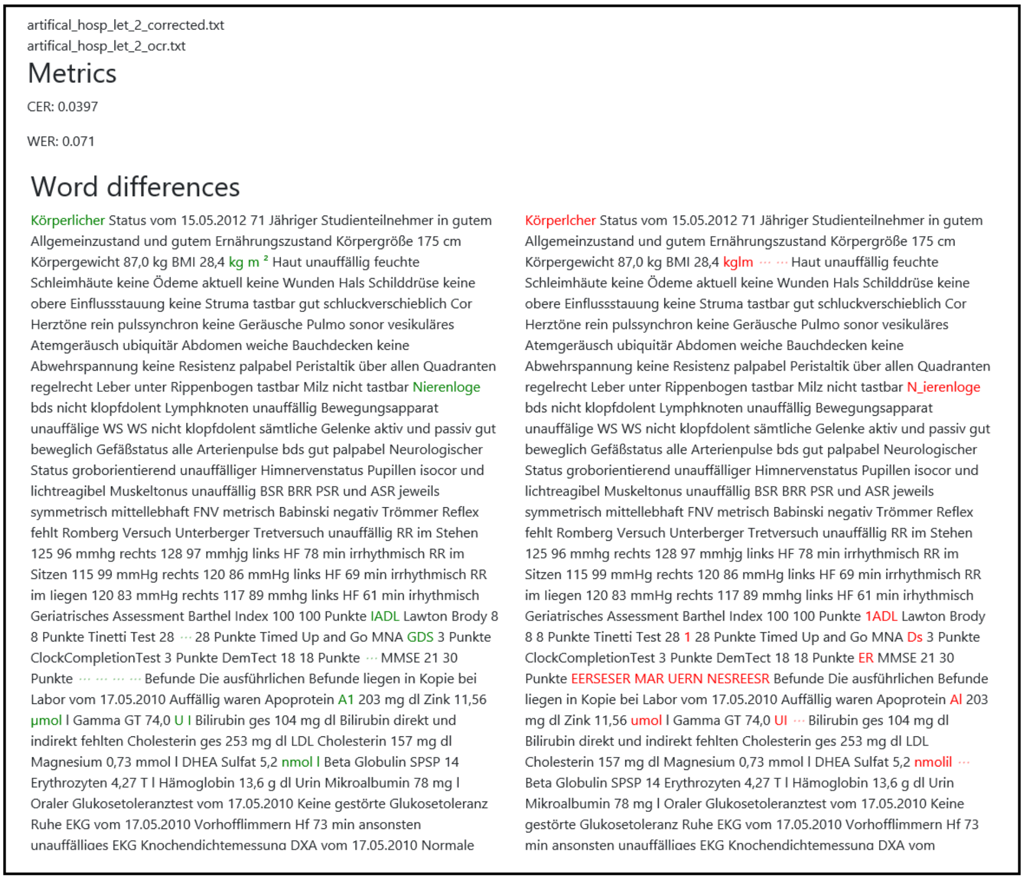
$$Recall = \frac{TP}{TP + FN} \tag{6.4}$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{6.5}$$

[34] `https://github.com/qurator-spk/dinglehopper`, accessed Mar. 20, 2023

[35] `https://ravius.sbb.berlin/`, accessed Mar. 20, 2023

where TP are the true positives, FP the false positives and FN the false negatives.



artifical_hosp_let_2_corrected.txt
artifical_hosp_let_2_ocr.txt

Metrics

CER: 0.0397

WER: 0.071

Word differences

Körperlicher Status vom 15.05.2012 71 Jähriger Studienteilnehmer in gutem Allgemeinzustand und gutem Ernährungszustand Körpergröße 175 cm Körpergewicht 87,0 kg BMI 28,4 kg m ^{2} Haut unauffällig feuchte Schleimhäute keine Ödeme aktuell keine Wunden Hals Schilddrüse keine obere Einflussstauung keine Struma tastbar gut schluckverschieblich Cor Herztöne rein pulssynchron keine Geräusche Pulmo sonor vesikuläres Atemgeräusch ubiquitär Abdomen weiche Bauchdecken keine Abwehrspannung keine Resistenz palpabel Peristaltik über allen Quadranten regelrecht Leber unter Rippenbogen tastbar Milz nicht tastbar Nierenloge bds nicht klopfdolent Lymphknoten unauffällig Bewegungsapparat unauffälige WS WS nicht klopfdolent sämtliche Gelenke aktiv und passiv gut beweglich Gefäßstatus alle Arterienpulse bds gut palpabel Neurologischer Status groborientierend unauffälliger Himnervenstatus Pupillen isocor und lichtreagibel Muskeltonus unauffällig BSR BRR PSR und ASR jeweils symmetrisch mittellebhaft FNV metrisch Babinski negativ Trömmer Reflex fehlt Romberg Versuch Unterberger Tretversuch unauffällig RR im Stehen 125 96 mmhg rechts 128 97 mmhjg links HF 78 min irrhythmisch RR im Sitzen 115 99 mmHg rechts 120 86 mmHg links HF 69 min irrhythmisch RR im liegen 120 83 mmHg rechts 117 89 mmhg links HF 61 min irhythmisch Geriatrisches Assessment Barthel Index 100 100 Punkte IADL Lawton Brody 8 8 Punkte Tinetti Test 28 ··· 28 Punkte Timed Up and Go MNA GDS 3 Punkte ClockCompletionTest 3 Punkte DemTect 18 18 Punkte ··· MMSE 21 30 Punkte ··· ··· ··· ··· Befunde Die ausführlichen Befunde liegen in Kopie bei Labor vom 17.05.2010 Auffällig waren Apoprotein A1 203 mg dl Zink 11,56 µmol l Gamma GT 74,0 U l Bilirubin ges 104 mg dl Bilirubin direkt und indirekt fehlten Cholesterin ges 253 mg dl LDL Cholesterin 157 mg dl Magnesium 0,73 mmol l DHEA Sulfat 5,2 nmol l Beta Globulin SPSP 14 Erythrozyten 4,27 T l Hämoglobin 13,6 g dl Urin Mikroalbumin 78 mg l Oraler Glukosetoleranztest vom 17.05.2010 Keine gestörte Glukosetoleranz Ruhe EKG vom 17.05.2010 Vorhofflimmern Hf 73 min ansonsten unauffälliges EKG Knochendichtemessung DXA vom 17.05.2010 Normale

Körperlcher Status vom 15.05.2012 71 Jähriger Studienteilnehmer in gutem Allgemeinzustand und gutem Ernährungszustand Körpergröße 175 cm Körpergewicht 87,0 kg BMI 28,4 kglm ··· ··· Haut unauffällig feuchte Schleimhäute keine Ödeme aktuell keine Wunden Hals Schilddrüse keine obere Einflussstauung keine Struma tastbar gut schluckverschieblich Cor Herztöne rein pulssynchron keine Geräusche Pulmo sonor vesikuläres Atemgeräusch ubiquitär Abdomen weiche Bauchdecken keine Abwehrspannung keine Resistenz palpabel Peristaltik über allen Quadranten regelrecht Leber unter Rippenbogen tastbar Milz nicht tastbar N_ierenloge bds nicht klopfdolent Lymphknoten unauffällig Bewegungsapparat unauffälige WS WS nicht klopfdolent sämtliche Gelenke aktiv und passiv gut beweglich Gefäßstatus alle Arterienpulse bds gut palpabel Neurologischer Status groborientierend unauffälliger Himnervenstatus Pupillen isocor und lichtreagibel Muskeltonus unauffällig BSR BRR PSR und ASR jeweils symmetrisch mittellebhaft FNV metrisch Babinski negativ Trömmer Reflex fehlt Romberg Versuch Unterberger Tretversuch unauffällig RR im Stehen 125 96 mmhg rechts 128 97 mmhjg links HF 78 min irrhythmisch RR im Sitzen 115 99 mmHg rechts 120 86 mmHg links HF 69 min irrhythmisch RR im liegen 120 83 mmHg rechts 117 89 mmhg links HF 61 min irhythmisch Geriatrisches Assessment Barthel Index 100 100 Punkte 1ADL Lawton Brody 8 8 Punkte Tinetti Test 28 1 28 Punkte Timed Up and Go MNA Ds 3 Punkte ClockCompletionTest 3 Punkte DemTect 18 18 Punkte ER MMSE 21 30 Punkte EERSESER MAR UERN NESREESR Befunde Die ausführlichen Befunde liegen in Kopie bei Labor vom 17.05.2010 Auffällig waren Apoprotein Al 203 mg dl Zink 11,56 umol l Gamma GT 74,0 UI ··· Bilirubin ges 104 mg dl Bilirubin direkt und indirekt fehlten Cholesterin ges 253 mg dl LDL Cholesterin 157 mg dl Magnesium 0,73 mmol l DHEA Sulfat 5,2 nmolil ··· Beta Globulin SPSP 14 Erythrozyten 4,27 T l Hämoglobin 13,6 g dl Urin Mikroalbumin 78 mg l Oraler Glukosetoleranztest vom 17.05.2010 Keine gestörte Glukosetoleranz Ruhe EKG vom 17.05.2010 Vorhofflimmern Hf 73 min ansonsten unauffälliges EKG Knochendichtemessung DXA vom

Figure 6.1: Example for error visualization with dinglehopper for artificial hospital letter.

Image pre-processing and spell correction was tested on 30 percent of the corrected hospital letter pages of the gold standard (60 hospital letter pages). The document classification was tested on 20 percent of the gold standard, but the classification set (160 hospital letter pages, 160 non hospital letter pages).

6.1 Results

6.1.1 Image Pre-Processing and Optical Character Recognition

Figure 6.2 shows the CER and WER for plain OCR, each pre-processing step and a combination of all pre-processing steps. The plain OCR has a CER of 10.8 percent and a WER of 22.2 percent. When the full pipeline is applied, the results improve the CER by 4.4 percent and the WER by 6.2 percent. The skew correction has the the largest impact on the improved results (a 3.7 percent reduction in WER). It also shows that noise reduction is the second most important component. Although the line and rectangle reduction used alone dramatically worsens the result, it leads to an improvement in combination with the other pipeline steps.

This is probably because removing the lines and rectangles creates noise that is removed in the subsequent steps. Super Resolution alone neither improves or worsens the result. However, in combination with noise reduction and the sometimes resulting blur, it could correct it. Nearly 2 percent of the improvement in results is due to the combination of steps.

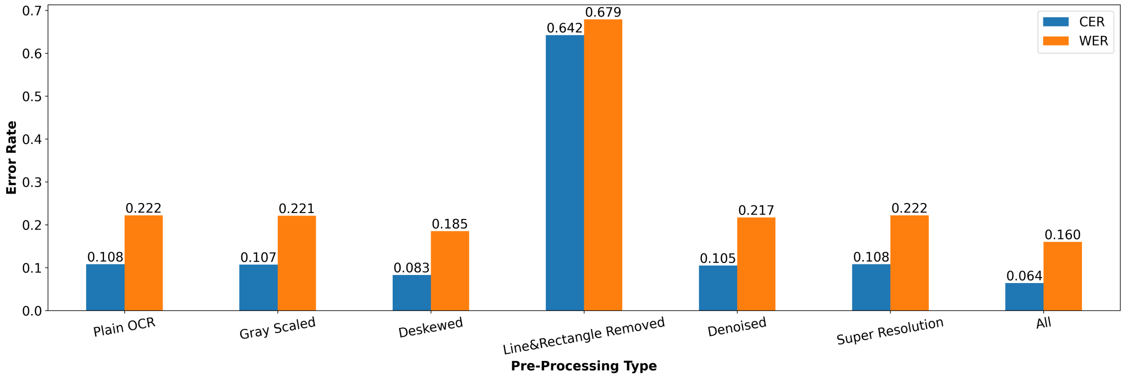


Figure 6.2: Average WER and CER for each pre-processing type.

6.1.2 Spell Correction

Figure 6.3 shows the CER and WER results of the spell correction based on different dictionaries. The first dictionary that was tested is the standard dictionary. It is the de100k dictionary (5.1) which includes only standard German terms. Since highly specialized language is used in the hospital reports, many terms are not in this dictionary. This explains why the result is significantly worse than without the dictionary. Many of the words are wrongly corrected. Introducing medical language to the dictionary leads to a considerable improvement. With the open dictionary, the WER drops to 15.7 percent. This could be further improved by adding the closed dictionary. In this case, the WER is 14.8, an improvement of 1.2 percent compared to plain pre-processing. However, the CER is increasing very slightly (by 0.6 percent). This is due to the fact that words are sometimes wrongly corrected here as well. This indicates that the dictionary needs to be expanded, since only a few processed hospital letters have been integrated into the dictionary so far.

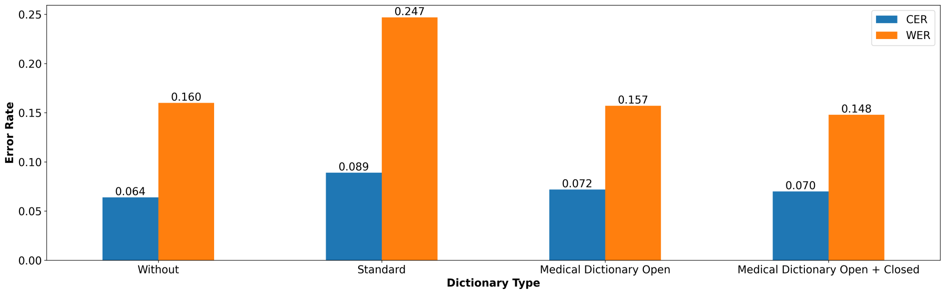


Figure 6.3: Average WER and CER for each dictionary.

In addition to the different dictionaries, different types of correction were also investigated. Figure 6.4 shows the results for these correction types. It can be seen that the simple lookup method performs significantly better than lookup compound and word segmentation.

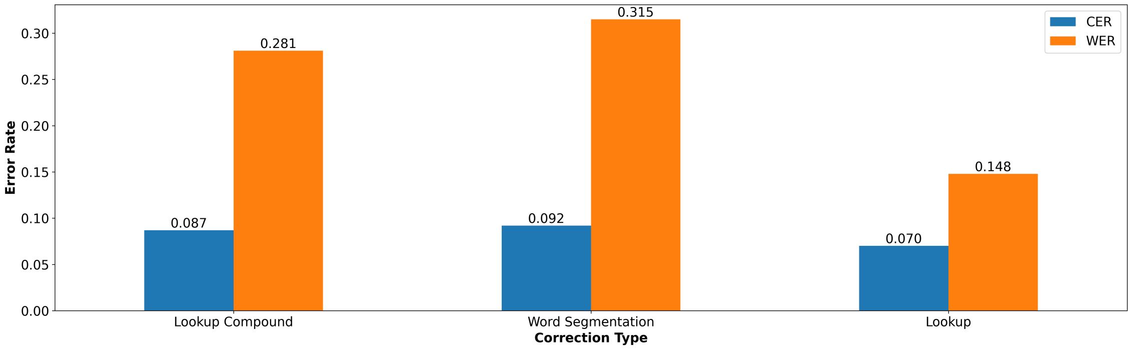


Figure 6.4: Average WER and CER for each correction type based on medical dictionary open+closed

6.1.3 Document Classification

For document classification, the models were trained on the one hand on plain OCR data and on the other hand with the pre- and post-processed data. The comparison of the results is show in table 6.1. For each of the models, the precision, recall, and f1 score were calculated.

The results show that especially the classical models like naive bayes and k-nearest neighbors perform better with pre- and post-processing. For these, the F1 score improves between 2 and 3 percent. Some of the BERT-based models also improve slightly, but not significantly. The F1 score of fasttext and flair decrease minimally.

GottBERT achieves the best result regarding the F1 score with 97.66 percent. This may be due to the fact that this BERT variant was trained on significantly more data than the other two BERT models.

Table 6.1: Comparison of classification results with plain OCR and pre- and post-processing

	Plain OCR			Pre- and Post-Processing		
	P	**R**	**F1**	**P**	**R**	**F1**
TF-IDF + Naive Bayes	89.37	94.70	91.96	91.41	98.67	94.90
TF-IDF + K-Nearest Neighbours	90.25	92.05	91.14	92.31	95.36	93.81
TF-IDF + Random Forest	91.55	93,37	92.45	92.45	97.35	94.84
Fasttext *(trained)*	94.93	99.33	97.09	93.59	1.0	96.69
Flair *(fine-tuned)*	94.26	99.34	96.10	94.67	96.60	95.62
German-MedBERT *(fine-tuned)*	96.12	98.67	97.38	95.40	98.64	96.99
GBERT base *(fine-tuned)*	94.93	99.33	97.08	96.00	99.29	97.61
GottBERT base *(fine-tuned)*	95.54	99.33	97.40	96.05	99.32	**97.66**

6.2 Error Analysis

6.2.1 Image Pre-Processing and Optical Character Recognition

The text resulting from OCR and previous image pre-processing contains various types of errors. A common error is the misrecognition of individual letters. These are often specific characters. So it occurs that instead of an ”a” a ”g” is recognized, e.g. ”Vorerkrankunaen” rather than ”Vorerkrankungen”, ”e” instead of ”c”, e.g ”oceipital” rather than ”occipital” or ”ii” instead of ”m”, e.g. ”Pflegeheiii” rather than ”Pflegeheim”, ”i” instead of ”l”, e.g. ”iumbalpunktiert” instead of ”lumbalpunktiert”. Sometimes also German umlauts are not recognized correctly, e.g. ”Ischämiezeichen” wrongly becomes ”Ischamiezeichen”, as well as some special characters (”u” rather than μ) The OCR also fails sometimes with dots and commas, e.g. with dates (”04.01 2021” rather than ”04.01.2021”). This incorrect recognitions are caused, among other things, by parts of letters being missing either due to erroneous printing or individual pre-processing steps (e.g. noise reduction). It can also be caused by different fonts in the same document (e.g. the actual text in one font and the meta information in another). Sometimes black vertical lines from top to bottom (scanning or printing artifacts) overlap with letters that are completely covered or cut in half (e.g. ”woraufhin” to ”wrozufhir”). It is difficult for the model to recognize the letters correctly when the font size is very small. This is the case, for example, when the address, information about the doctor and hospital are given in the upper right part of the document. These are often printed in a smaller font size. In addition to misrecognized characters, the OCR result occasionally also contains segmentation errors. Either connected letters are torn apart or two words are wrongly connected. Examples include the following: ”keine Provokationsfaktoren” to ”keineProvokationsfaktoren or ”abgrenzbareSeitenventrikelhörner” instead of ”abgrenzbare Seitenventrikelhörner”. In addition, in a few cases complete lines are swapped. This can be caused when the skew is not completely corrected. As the results of the pre-processing have shown, the skew correction has a significant influence on the OCR quality. Another influencing factor lies in documents with a more complex structure and layout.

6.2.2 Spell Correction

After applying the spell correction, some OCR errors remain, in some cases, new errors are added. The problem of errors not being corrected may be due to the fact that the set edit distance is too low. It was set to two, because otherwise the number of erroneous corrections would increase. In some cases, however, an edit distance of three would be necessary, e.g. ”unauffalkg” (correct: ”unauffällig”) and ”Hauntdicinosen” (correct: ”Hauptdiagnosen”). Furthermore, as mentioned above, real-world errors cannot be corrected by the dictionary approach (e.g. ”Keime Allergien” instead of ”Keine Allergien”). Besides, in some cases, the wrong corrections are made when the actual word is not in the dictionary. For example ”diatisch” to ”drastisch” (correct would be ”diätisch”). Actually correct words are then incorrectly replaced: e.g. ”nexthaler” becomes wrongly ”lethaler”, ”pleurallinie” becomes wrongly ”pleuralgie” and ”cardiolipin igm” wrongly changed to ”cariolipin igg”.

7 Future Work

This chapter presents how the pipeline could be improved and its functionality extended.

7.1 Pre-processing

To improve denoising and deblurring, more advanced techniques such as u-shaped transformers for image restoration (Wang et al. 2022) could be tested.

Complex tables, which are often included in the laboratory data, were excluded from this thesis. In order to make these usable, a layout detection system could be added to the pipeline. For example, ClinicalLayoutLM (Wei et al. 2022) or LayoutParser (Shen et al. 2021) would be potentially suitable for this. However, additional training data would have to be created for this.

7.2 OCR

One way to improve the OCR engine is to fine tune tesseract[36] on the hospital letters and the other medical documents. The gold standard created for testing OCR and pre- and post-processing would already provide an initial data set for this purpose. This could improve the overall accuracy of the OCR engine and in particular the recognition of special symbols (e.g. bullet points, tick boxes, arrows) that are currently poorly recognized.

In addition, it could be tested whether alternatives to Tesseract produce better results. There exist also transformer-based systems in this area. For example TrOCR, a transformer-based optical character recognition with pre-trained models (Li et al. 2021).

7.3 Post-processing

The existing dictionary-based approach to spell correction would still benefit from additional data on hospital letters, since the language used is very specialized and includes jargon and slang. Even with medical text corpora (e.g PubMed) this cannot be covered. For this it would be necessary to correct more hospital letters and to integrate the data into the frequency dictionary. It could also be tested whether bigrams in the frequency dictionary improve the spell correction result.

As an additional approach, it could be investigated whether sequence-to-sequence models for post-processing of the OCR results improve the outcome (e.g. Ramirez-Orta et al. 2022). In this context, BART (Lewis et al. 2019, FLAN-T5 (Chung et al. 2022) or the preciously published ByT5 Xue et al. 2022 could be tested. ByT5 uses UTF-8 bytes instead of tokens and would therefore be particularly suitable for the spell correction task.

[36] `https://tesseract-ocr.github.io/tessdoc/tess5/TrainingTesseract-5.html`, accessed Mar. 20, 2023

7.4 Document Classfication

The document classification could be extended to multi-classification. This could be used to classify not only hospital reports, but also autopsies, report forms, emails and other documents.

7.5 Named Entity Recogntion and Relation Extraction

The hospital reports could now be further analyzed using named entity recognition and relation extraction to detect adverse drug events and related case information. However, this requires not only the development of appropriate models, but also the acquisition of additional labeled data. In terms of models, the work of GERNERMED++ (Frei, Frei-Stuber, and Kramer 2022) could be followed up and extended.

8 Conclusion

This thesis presents a pipeline for the automatic processing of hospital letters. This includes image pre-processing and OCR as well as a spell correction approach. The results show that preprocessing can improve the OCR result, deskewing is particularly important. They also show that spell correction with ordinary dictionaries significantly worsens the result. To improve the spell correction result, a German medical corpus and a corresponding frequency list were created. This also leads to improved results in terms of WER, but needs to be enriched with hospital report data for actual application. The classification task shows that the classical machine learning models perform surprisingly well, but are surpassed by the Transformer models. The classical models benefit most from pre- and post-processing, but minimal improvements in the BERT-based models can also be observed. Other tasks such as NER and relation extraction could also take advantage of pre- and post-processing. There are still some improvements to be made for the actual application of this pipeline, but the foundation for further optimizations could be established.

Bibliography

Akbik, Alan, Tanja Bergmann, et al. (June 2019). "FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations).* Minneapolis, Minnesota: Association for Computational Linguistics, pp. 54–59. DOI: 10.18653/v1/N19-4010. URL: https://aclanthology.org/N19-4010.

Akbik, Alan, Duncan Blythe, and Roland Vollgraf (Aug. 2018). "Contextual String Embeddings for Sequence Labeling". In: *Proceedings of the 27th International Conference on Computational Linguistics.* Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 1638–1649. URL: https://aclanthology.org/C18-1139.

Anandarajan, M., C. Hill, and T. Nolan (2019). *Practical Text Analytics: Maximizing the Value of Text Data.* Advances in analytics and data science. Springer International Publishing. ISBN: 9783319956640. URL: https://books.google.de/books?id=RNhbzwEACAAJ.

Badoiu, Vlad, Andrei-Constantin Ciobanu, and Sergiu Craitoiu (May 2016). "OCR quality improvement using image preprocessing". English. In: *Journal of Information Systems & Operations Management* 10. Report, pp. V1+. ISSN: 18434711. URL: https://link.gale.com/apps/doc/A483829355/AONE?u=anon~69fd4429&sid=googleScholar&xid=d6bb88c8.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017). "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146. ISSN: 2307-387X.

Borchert, Florian et al. (Nov. 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.* Online: Association for Computational Linguistics, pp. 38–48. DOI: 10.18653/v1/2020.louhi-1.5. URL: https://aclanthology.org/2020.louhi-1.5.

Boytsov, Leonid (May 2011). "Indexing Methods for Approximate Dictionary Searching: Comparative Analysis". In: *ACM J. Exp. Algorithmics* 16. ISSN: 1084-6654. DOI: 10.1145/1963190.1963191. URL: https://doi.org/10.1145/1963190.1963191.

Carrasco, Rafael C. (2014). "An Open-Source OCR Evaluation Tool". In: *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage.* DATeCH '14. Madrid, Spain: Association for Computing Machinery, pp. 179–184. ISBN: 9781450325882. DOI: 10.1145/2595188.2595221. URL: https://doi.org/10.1145/2595188.2595221.

Chan, Branden, Stefan Schweter, and Timo Möller (Dec. 2020). "German's Next Language Model". In: *Proceedings of the 28th International Conference on Computational Linguistics.* Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 6788–6796. DOI: 10.18653/v1/2020.coling-main.598. URL: https://aclanthology.org/2020.coling-main.598.

Chaudhuri, A., K. Mandaviya, P. Badelia, and S.K. Ghosh (2016). *Optical Character Recognition Systems for Different Languages with Soft Computing.* Studies in Fuzziness and Soft

Computing. Springer International Publishing. ISBN: 9783319502526. URL: https://books.google.de/books?id=-ZXJDQAAQBAJ.

Chung, Hyung Won et al. (2022). *Scaling Instruction-Finetuned Language Models.* DOI: 10.48550/ARXIV.2210.11416. URL: https://arxiv.org/abs/2210.11416.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

Dong, Chao, Chen Change Loy, and Xiaoou Tang (2016). "Accelerating the Super-Resolution Convolutional Neural Network". In: *Computer Vision – ECCV 2016.* Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling. Cham: Springer International Publishing, pp. 391–407. ISBN: 978-3-319-46475-6.

Frei, Johann, Ludwig Frei-Stuber, and Frank Kramer (2022). *GERNERMED++: Transfer Learning in German Medical NLP.* DOI: 10.48550/ARXIV.2206.14504. URL: https://arxiv.org/abs/2206.14504.

Frei, Johann and Frank Kramer (2022). "GERNERMED: An open German medical NER model". In: *Software Impacts* 11, p. 100212. ISSN: 2665-9638. DOI: https://doi.org/10.1016/j.simpa.2021.100212. URL: https://www.sciencedirect.com/science/article/pii/S2665963821000944.

Garbe, Wolf (June 2012). *SymSpell.* URL: hhttps://github.com/wolfgarbe/SymSpell.

Goodrum, Heath, Kirk Roberts, and Elmer V. Bernstam (2020). "Automatic classification of scanned electronic health record documents". In: *International Journal of Medical Informatics* 144, p. 104302. ISSN: 1386-5056. DOI: https://doi.org/10.1016/j.ijmedinf.2020.104302. URL: https://www.sciencedirect.com/science/article/pii/S1386505620309977.

Guo, Gongde, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer (Mar. 2006). "Using kNN model for automatic text categorization". In: *Soft Computing* 10.5, pp. 423–430. ISSN: 1433-7479. DOI: 10.1007/s00500-005-0503-y. URL: https://doi.org/10.1007/s00500-005-0503-y.

Hsu, Enshuo, Ioannis Malagaris, Yong-Fang Kuo, Rizwana Sultana, and Kirk Roberts (June 2022). "Deep learning-based NLP data pipeline for EHR-scanned document information extraction". In: *JAMIA Open* 5.2. ooac045. ISSN: 2574-2531. DOI: 10.1093/jamiaopen/ooac045. eprint: https://academic.oup.com/jamiaopen/article-pdf/5/2/ooac045/44023378/ooac045.pdf. URL: https://doi.org/10.1093/jamiaopen/ooac045.

Islam, Noman, Zeeshan Islam, and Nazia Noor (Dec. 2016). "A Survey on Optical Character Recognition System". In: *ITB Journal of Information and Communication Technology.*

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov (Apr. 2017). "Bag of Tricks for Efficient Text Classification". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers.* Association for Computational Linguistics, pp. 427–431.

Kay, Anthony (July 2007). "Tesseract: An Open-Source Optical Character Recognition Engine". In: *Linux J.* 2007.159, p. 2. ISSN: 1075-3583.

Kelly, Liadh et al. (2019). "Overview of the CLEF EHealth Evaluation Lab 2019". In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings*. Lugano, Switzerland: Springer-Verlag, pp. 322–339. ISBN: 978-3-030-28576-0. DOI: 10.1007/978-3-030-28577-7_26. URL: https://doi.org/10.1007/978-3-030-28577-7_26.

Kittner, Madeleine et al. (Apr. 2021). "Annotation and initial evaluation of a large annotated German oncological corpus". In: *JAMIA Open* 4.2. ooab025. ISSN: 2574-2531. DOI: 10.1093/jamiaopen/ooab025. eprint: https://academic.oup.com/jamiaopen/article-pdf/4/2/ooab025/38830128/ooab025.pdf. URL: https://doi.org/10.1093/jamiaopen/ooab025.

König, Maximilian, André Sander, Ilja Demuth, Daniel Diekmann, and Elisabeth Steinhagen-Thiessen (Nov. 2019). "Knowledge-based best of breed approach for automated detection of clinical events based on German free text digital hospital discharge letters". In: *PLOS ONE* 14.11, pp. 1–14. DOI: 10.1371/journal.pone.0224916. URL: https://doi.org/10.1371/journal.pone.0224916.

Kowsari, Kamran et al. (2019). "Text Classification Algorithms: A Survey". In: *Information* 10.4. ISSN: 2078-2489. DOI: 10.3390/info10040150. URL: https://www.mdpi.com/2078-2489/10/4/150.

Kumar, Atul (2016). "A Survey on Various OCR Errors". In: *International Journal of Computer Applications* 143, pp. 8–10.

Lat, Ankit and C. V. Jawahar (2018). "Enhancing OCR Accuracy with Super Resolution". In: *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 3162–3167. DOI: 10.1109/ICPR.2018.8545609.

Lentzen, Manuel et al. (Nov. 2022). "Critical assessment of transformer-based AI models for German clinical notes". In: *JAMIA Open* 5.4. ooac087. ISSN: 2574-2531. DOI: 10.1093/jamiaopen/ooac087. eprint: https://academic.oup.com/jamiaopen/article-pdf/5/4/ooac087/47042094/ooac087_supplementary_data.pdf. URL: https://doi.org/10.1093/jamiaopen/ooac087.

Lewis, Mike et al. (2019). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. DOI: 10.48550/ARXIV.1910.13461. URL: https://arxiv.org/abs/1910.13461.

Li, Minghao et al. (2021). *TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models*. DOI: 10.48550/ARXIV.2109.10282. URL: https://arxiv.org/abs/2109.10282.

Liu, Bing (2006). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Berlin, Heidelberg: Springer-Verlag. ISBN: 3540378812.

Lohr, Christina, Sven Buechel, and Udo Hahn (May 2018). "Sharing Copies of Synthetic Clinical Corpora without Physical Distribution — A Case Study to Get Around IPRs and Privacy Constraints Featuring the German JSYNCC Corpus". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). URL: https://aclanthology.org/L18-1201.

Melamud, Oren, Jacob Goldberger, and Ido Dagan (Aug. 2016). "context2vec: Learning Generic Context Embedding with Bidirectional LSTM". In: *Proceedings of the 20th SIGNLL Con-*

ference on Computational Natural Language Learning. Berlin, Germany: Association for Computational Linguistics, pp. 51–61. DOI: 10.18653/v1/K16-1006. URL: https://aclanthology.org/K16-1006.

Mikolov, Tomas, Kai Chen, Gregory S. Corrado, and Jeffrey Dean (2013). "Efficient Estimation of Word Representations in Vector Space". In: *International Conference on Learning Representations*.

Mitchell, T.M. (1997). *Machine Learning*. McGraw-Hill International Editions. McGraw-Hill. ISBN: 9780071154673. URL: https://books.google.de/books?id=EoYBngEACAAJ.

Ng, Nathan et al. (Aug. 2019). "Facebook FAIR's WMT19 News Translation Task Submission". In: *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics, pp. 314–319. DOI: 10.18653/v1/W19-5333. URL: https://aclanthology.org/W19-5333.

Nguyen, Thi Tuyet Hai, Adam Jatowt, Mickael Coustaty, and Antoine Doucet (July 2021). "Survey of Post-OCR Processing Approaches". In: *ACM Comput. Surv.* 54.6. ISSN: 0360-0300. DOI: 10.1145/3453476. URL: https://doi.org/10.1145/3453476.

Niblack, W. (1986). *An Introduction to Digital Image Processing*. Delaware Symposia on Language Studies5. Prentice-Hall International. ISBN: 9780134806747. URL: https://books.google.de/books?id=XOxRAAAAMAAJ.

Peters, Matthew E. et al. (June 2018). "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: https://aclanthology.org/N18-1202.

Ramirez-Orta, Juan Antonio, Eduardo Xamena, Ana Maguitman, Evangelos Milios, and Axel J. Soto (June 2022). "Post-OCR Document Correction with Large Ensembles of Character Sequence-to-Sequence Models". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.10, pp. 11192–11199. DOI: 10.1609/aaai.v36i10.21369. URL: https://ojs.aaai.org/index.php/AAAI/article/view/21369.

Roller, Roland et al. (2022). *A Medical Information Extraction Workbench to Process German Clinical Text*. DOI: 10.48550/ARXIV.2207.03885. URL: https://arxiv.org/abs/2207.03885.

Sauvola, J. and M. Pietikäinen (2000). "Adaptive document image binarization". In: *Pattern Recognition* 33.2, pp. 225–236. ISSN: 0031-3203. DOI: https://doi.org/10.1016/S0031-3203(99)00055-2. URL: https://www.sciencedirect.com/science/article/pii/S0031320399000552.

Scheible, Raphael, Fabian Thomczyk, Patric Tippmann, Victor Jaravine, and Martin Boeker (2020). *GottBERT: a pure German Language Model*. DOI: 10.48550/ARXIV.2012.02110. URL: https://arxiv.org/abs/2012.02110.

Sezgin, Mehmet and Bülent Sankur (2004). "Survey over image thresholding techniques and quantitative performance evaluation". In: *J. Electronic Imaging* 13, pp. 146–168.

Shen, Zejiang et al. (2021). "LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis". In: *arXiv preprint arXiv:2103.15348*.

Shi, Wenzhe et al. (2016). “Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1874–1883. DOI: 10.1109/CVPR.2016.207.

Shrestha, Manjil (2021). “Development of a Language Model for Medical Domain”. masterthesis. Hochschule Rhein-Waal, p. 141.

Smith, Ray (2011). “Limits on the Application of Frequency-Based Language Models to OCR”. In: *Proceedings of the 2011 International Conference on Document Analysis and Recognition.* ICDAR ’11. USA: IEEE Computer Society, pp. 538–542. ISBN: 9780769545202. DOI: 10.1109/ICDAR.2011.114. URL: https://doi.org/10.1109/ICDAR.2011.114.

Strohmaier, C.M., C. Ringlstetter, K.U. Schulz, and S. Mihov (2003). “Lexical postcorrection of OCR-results:the web as a dynamic secondary dictionary?” In: *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.* Pp. 1133–1137. DOI: 10.1109/ICDAR.2003.1227833.

Sun, Chi, Xipeng Qiu, Yige Xu, and Xuanjing Huang (2019). *How to Fine-Tune BERT for Text Classification?* DOI: 10.48550/ARXIV.1905.05583. URL: https://arxiv.org/abs/1905.05583.

van Strien., Daniel et al. (2020). “Assessing the Impact of OCR Quality on Downstream NLP Tasks”. In: *Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 1: ARTIDIGH,* INSTICC. SciTePress, pp. 484–496. ISBN: 978-989-758-395-7. DOI: 10.5220/0009169004840496.

Vaswani, Ashish et al. (2017). *Attention Is All You Need.* DOI: 10.48550/ARXIV.1706.03762. URL: https://arxiv.org/abs/1706.03762.

Wang, Zhendong et al. (2022). “Uformer: A General U-Shaped Transformer for Image Restoration”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17662–17672. DOI: 10.1109/CVPR52688.2022.01716.

Wei, Qiang et al. (2022). “ClinicalLayoutLM: A Pre-trained Multi-modal Model for Understanding Scanned Document in Electronic Health Records”. In: *2022 IEEE International Conference on Big Data (Big Data)*, pp. 2821–2827. DOI: 10.1109/BigData55660.2022.10020569.

Xue, Linting et al. (2022). “ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models”. In: *Transactions of the Association for Computational Linguistics* 10, pp. 291–306. DOI: 10.1162/tacl_a_00461. URL: https://aclanthology.org/2022.tacl-1.17.

Zhuang, Liu, Lin Wayne, Shi Ya, and Zhao Jun (Aug. 2021). “A Robustly Optimized BERT Pre-training Approach with Post-training”. In: *Proceedings of the 20th Chinese National Conference on Computational Linguistics.* Huhhot, China: Chinese Information Processing Society of China, pp. 1218–1227. URL: https://aclanthology.org/2021.ccl-1.108.