Dissertation

Leveraging Machine Learning Methods for Processing Non-Lexical Content in Clinical Narratives

submitted by

Amila KUGIC
Ing.[in] Dipl.-Ing.[in] BSc BSc

for the Academic Degree of

Doctor of Medical Science
(Dr.scient.med.)

at the

Medical University of Graz

Institute for Medical Informatics, Statistics and Documentation

under the Supervision of

Univ.-Prof. Dr.med. Stefan SCHULZ

2025

Declaration

I hereby confirm that this thesis is the result of my own independent scholarly work. I also confirm that in all cases, where material from the work of others (in books, articles, essays, dissertations, and on the internet) is acknowledged, quotations and paraphrases are clearly indicated. No material other than that cited in the reference list has been used. I have read and understood the Medical University's regulations and procedures concerning plagiarism.

Furthermore, I hereby declare that if artificial intelligence (AI) tools were used for the generation and/or correction of certain text passages in the creation of this work, such employment was conducted in compliance with ethical principles, academic integrity, and the regulations of my university. Additionally, it was ensured that this usage was transparently disclosed and appropriately attributed.

Graz, 21st February 2025 Amila Kugic

Disclosures

The following list of publications corresponds to papers, which were authored during the course of my studies, either as first author or co-author. The first-authored publications form the basis of this dissertation.

Kugic A[1], Potjan LM[2], Hammer LM[2], Schulz S[2], Kreuzthaler M[2]. **Alcohol Status Standardization from Clinical Real World Data with Transformer Architectures.** 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI), pp. 233-238, DOI: 10.1109/ICHI54592.2022.00043.

[1] CBmed GmbH - Center for Biomarker Research in Medicine, Graz, Austria.

[2] Institute for Medical Informatics, Statistics and Documentation,
Medical University of Graz, Austria.

Kugic A[1], Pfeifer B[1], Schulz S[1], Kreuzthaler M[1]. **Data-Driven Identification of Clinical Real-World Expressions Linked to ICD**. Stud Health Technol Inform. 2023 May 18;302:827–8. DOI: 10.3233/SHTI230279.
[1] Institute for Medical Informatics, Statistics and Documentation,
Medical University of Graz, Austria.

Kugic A[1], Pfeifer B[1], Schulz S[1], Kreuzthaler M[1]. **Embedding-based terminology expansion via secondary use of large clinical real-world datasets**. Journal of Biomedical Informatics. 2023 Sep 29;104497. DOI: 10.1016/j.jbi.2023.104497.
[1] Institute for Medical Informatics, Statistics and Documentation,
Medical University of Graz, Austria.

Kugic A[1], Kreuzthaler M[1], Schulz S[1]. **Clinical Acronym Disambiguation via ChatGPT and BING**. Stud Health Technol Inform. 2023 Oct 20;309:78–82. DOI: 10.3233/SHTI230743.
[1] Institute for Medical Informatics, Statistics and Documentation,
Medical University of Graz, Austria.

Kugic A[1], Schulz S[1], Kreuzthaler M[1]. **Identification of Non-Lexical Content in Croatian Health Forum Entries**. In: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Istanbul, Turkiye: IEEE; 2023. p. 4328–35. DOI: 10.1109/BIBM58861.2023.10386069
[1] Institute for Medical Informatics, Statistics and Documentation,
Medical University of Graz, Austria.

Abdulnazar A[1,2], **Kugic A**[1], Schulz S[1], Stadlbauer V[3], Kreuzthaler M[1]. **O2 supplementation disambiguation in clinical narratives to support retrospective COVID-19 studies**. BMC Medical Informatics and Decision Making. 2024 Jan 31;24(1):29. DOI: 10.1186/s12911-024-02425-2.

[1] Institute for Medical Informatics, Statistics and Documentation,
Medical University of Graz, Austria.

[2] CBmed GmbH - Center for Biomarker Research in Medicine, Graz, Austria.

[3] Division of Gastroenterology and Hepatology, Department of Internal Medicine,
Medical University of Graz, Graz, Austria.

Kugic A[1], Schulz S[1], Kreuzthaler M[1]. **Disambiguation of acronyms in clinical narratives with large language models**. Journal of the American Medical Informatics Association. 2024 Sep 1;31(9):2040–6. DOI: 10.1093/jamia/ocae157.

[1] Institute for Medical Informatics, Statistics and Documentation,
Medical University of Graz, Austria.

Kugic A[1], Abdulnazar A[1,2], Knezovic A[1], Schulz S[1], Kreuzthaler M[1]. **Smoking Status Classification: A Comparative Analysis of Machine Learning Techniques with Clinical Real World Data**. In: Finkelstein J, Moskovitch R, Parimbelli E, editors. Artificial Intelligence in Medicine. Cham: Springer Nature Switzerland; 2024. p. 182–91. (Lecture Notes in Computer Science; vol. 14844). DOI: 10.1007/978-3-031-66538-7_19.

[1] Institute for Medical Informatics, Statistics and Documentation,
Medical University of Graz, Austria.

[2] CBmed GmbH - Center for Biomarker Research in Medicine, Graz, Austria.

Sharma V[1], Thalhammer A[2], **Kugic A**[3], Schulz S[3], Kreuzthaler M[3]. **Sequence-Model-Based Medication Extraction from Clinical Narratives in German**. In: Finkelstein J, Moskovitch R, Parimbelli E, editors. Artificial Intelligence in Medicine. Cham: Springer Nature Switzerland; 2024. p. 334–44. (Lecture Notes in Computer Science; vol. 14844). DOI: 10.1007/978-3-031-66538-7_33.

[1] Roche Diagnostics, California, USA

[2] F. Hoffmann-La Roche AG, Basel, Switzerland

[3] Institute for Medical Informatics, Statistics and Documentation,
Medical University of Graz, Austria.

Kugic A[1], Schulz S[1], Kreuzthaler M[1]. **Term Candidate Generation to Enrich Clinical Terminologies with Large Language Models**. In: Mantas J, Hasman A, Demiris G, Saranto K, Marschollek M, Arvanitis TN, et al., editors. Studies in Health Technology and Informatics. IOS Press; 2024. DOI: 10.3233/SHTI240509.

[1] Institute for Medical Informatics, Statistics and Documentation,
Medical University of Graz, Austria.

Kugic A[1], Kreuzthaler M[1], Schulz S[1]. **Annotation of Non-Lexical Entities in Croatian Health Forum Entries With Large Language Models**. In: Book of Abstracts of the XXI EURALEX International Congress. Cavtat, Croatia: Institut za hrvatski jezik; p. 147–51. Available from: https://euralex.jezik.hr/wp-content/uploads/2021/09/Euralex_boa_20.pdf.

[1] Institute for Medical Informatics, Statistics and Documentation,
Medical University of Graz, Austria.

Kugic A[1], Martin I[2], Modersohn L[2], Pallaoro P[2], Kreuzthaler M[1], Schulz S[1], Boeker M[2]. **Processing of Short-Form Content in Clinical Narratives: Systematic Scoping Review**. Journal of Medical Internet Research. 2024 Sep 26;26:e57852. DOI: 10.2196/57852.

[1] Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria.

[2] Institute for AI and Informatics in Medicine, School of Medicine and Health, Technical University of Munich, Munich, Germany.

Acknowledgements

I want to thank three groups of people, who contributed to the successful completion of this four-year-long journey.

First and foremost, I would like to express my heartfelt gratitude to my supervisors, Stefan Schulz, Markus Kreuzthaler, and Martin Boeker, for their mentorship, patience, and guidance. Especially in the beginning, understanding the field and creating direction towards possible research investigations was instrumentally helpful. Patiently explaining the intended goals and research steps that were required, majorly shaped the dissertation and research questions answered during the dissertation. Their collective encouragement and feedback at every stage of my studies have been invaluable.

Second, to all colleagues and co-authors, thank you for the laughter, support, and perseverance. The shared in-depth discussions helped me gain a deeper appreciation and understanding of the field of medical informatics.

Third, this work was partially funded and made possible through the generous support of the Institute for Medical Informatics, Statistics and Documentation at the Medical University of Graz, and CBmed GmbH - Center for Biomarker Research in Medicine. I am especially grateful to them for their financial assistance that enabled opportunities to attend international conferences and share my work with the wider academic community, and to publish in high-impact journals. As a doctoral student, I received funding from the following grants:

- Digital Biomarkers for Precision Medicine (DBM4PM) project within the K1 COMET Competence Center CBmed (https://www.cbmed.at/), funded by the Federal Ministry of Transport, Innovation and Technology (BMVIT); the Federal Ministry of Science, Research and Economy (BMWFW); Land Steiermark (Department 12, Business and Innovation); the Styrian Business Promotion Agency (SFG); and the Vienna Business Agency. The COMET program is executed by the FFG.

- European Union's Horizon Research and Innovation Programme under grant agreement No 101057062; AI-powered Data Curation & Publishing Virtual Assistant (AIDAVA), https://aidava.eu/)
- FFG Basisprogramm; Predicting Patient Outcomes in Emergency Departments with Causal Machine Learning (PREMEDICAL)
- Medical University of Graz through the Doctoral School Sustainable Health Research (SHR)

Graz, 21st February 2025 Amila Kugic

Contents

List of Abbreviations

A

AUROC area under the receiver operating characteristic.

B

BERT Bidirectional Encoder Representations from Transformers.
BiLSTM Bidirectional Long Short-Term Memory.
BIO beginning-inside-outside.

C

CASI Clinical Abbreviation Sense Inventory.
CNN Convolutional Neural Network.
CRF Conditional Random Field.
cTAKES clinical Text Analysis and Knowledge Extraction System.

E

EHR Electronic Health Record.
ELECTRA Efficiently Learning an Encoder that Classifies Token Replacements Accurately.
ETL Extract Transform Load.

G

GPT Generative Pretrained Transformer.

L

LLM Large Language Model.
LSTM Long Short-Term Memory.

M

Main-NLC Main Categorical Non-Lexical Content.
MAP Mean Average Precision.
MIMIC Medical Information Mart for Intensive Care.
ML Machine Learning.

N

NER Named Entity Recognition.
NLC Non-Lexical Content.
NLE Non-Lexical Entity.
NLP Natural Language Processing.

S

SDOH Social Determinants of Health.
SNOMED CT Standardized Nomenclature of Medicine Clinical Terms.
SOTA state-of-the-art.
Sub-NLC Sub Categorical Non-Lexical Content.
SVM Support Vector Machine.

U

UMLS Unified Medical Language System.

List of Figures

List of Tables

Zusammenfassung

Einleitung. Jegliche Art der textuellen Dokumentation in elektronischen Patientenakten, unabhängig von der Sprache, enthält spezialisierte Formen von Ausdrücken, die oft Elemente aus mehreren Sprachen umfassen. So enthält die deutsche Klinikspra-che auch zahlreiche Wärter und Wortstämme aus dem Lateinischen, Griechischen, und Englischen. Die Verwendung von Fachausdrücken, Kurzformen und anderen sprachlichen Merkmalen erfordern spezielle Verarbeitungspipelines um klinische Bedeutungen zu extrahieren und zu disambiguieren. Die größte Herausforderung besteht darin, dass all diese Merkmale häufig nicht-lexikalischer Natur und kontextabhängig sind.

Methoden. Für die Analyse wurden verschiedene Datensätze und Rahmenstrukturen für Untersuchungen im Bereich des maschinellen Lernens herangezogen, darunter Fachwörterbücher, klinische Terminologien, Lexika, Sprachmodelle und grundlegende Strukturen zur Verarbeitung natürlicher Sprache. Fachexperten validierten die Untersuchungen, um die Leistung der einzelnen, durch Computeralgorithmen bewältigten Aufgaben zu bewerten.

Ergebnisse. Die Studien zu nicht-lexikalisierten Sprachkomponenten integrierten sowohl aktuelle Techniken des maschinellen Lernens einschießlichlich großer Sprachmodelle (LLM). Für die Erkennung dieser Sprachkomponenten wurden Namenserkenner und LLM eingesetzt. Zur Erweiterung von Domänenlexika dienten Kookkurrenzanalysen, Vektordarstellungen und LLM, um neue Bedeutungen aus Bestandteilen klinischer Texte zu identifizieren. Die Disambiguierung von nicht-lexikalisierten Sprachkomponenten, insbesondere von Akronymen, erfolgte durch Text-Mining und LLM-basierte Ansätze. Eine Übersichtsarbeit unterzog bisherige Ansätze der automatischen Interpretation von Abkürzungen in klinischen Texten einer detaillierten Analyse. Außerdem wurden Informationen zu Risikofaktoren wie Raucher- und Alkoholstatus in die Klassifizierung klinischer Texte einbezogen.

Diskussion. Hochwertige, ausgewogene Datensätze und Sprachressourcen sind essenziell für leistungsstarke Ergebnisse im maschinellen Lernen. Mangelhafte oder unzureichende Terminologien mindern die Effizienz. Die Generierbarkeit und Wiederverwendbarkeit von Lösungen wird beeinträchtigt, da vergleichbare Aufgaben mit denselben Ressourcen in unterschiedlichen Sprachen oft nicht einheitlich umsetzbar sind.

Abstract

Introduction. Any kind of narrative in electronic health records, regardless of the language, contains specialized forms of expressions. These often include elements from multiple languages, particularly Latin, Greek and English. The use of jargon expressions, short forms, and other linguistic features require specialized processing pipelines to extract and disambiguate clinical mentions. The main challenge to overcome is that these features are often non-lexical and fully context-dependent.

Methods. This dissertation used various datasets and computational frameworks for machine learning investigations, such as clinical terminologies, domain dictionaries, short form sense inventories, language models and natural language processing frameworks. Each investigation underwent rigorous testing including annotation of results by domain experts to correctly evaluate the performance of each subtask.

Results. Many of the investigations focusing on non-lexical entities (NLEs) did not only cover state-of-the-art methods for processing NLEs, but also incorporated the application of large language models (LLMs) to the tasks at hand. First, for the identification of NLEs, named entity recognition and LLMs were used for the recognition of NLEs. Second, for the expansion of domain lexicons, co-occurrence analyses, embedding representations, and LLMs were applied to extract unknown term candidates from large clinical datasets for terminology expansion. Third, the disambiguation of NLEs, particularly of acronyms, used both a text mining approach, as well as LLM-based disambiguation. Fourth, short-form processing of clinical narratives were explored through a scoping systematic review. Finally, risk factor retrieval, *viz.* smoking status and alcohol status mentions in clinical texts, helped in analyzing the role of NLEs in clinical text classification.

Discussion. Quality of datasets and language resources need to be balanced and available to reach high-performance state-of-the-art results with machine learning methods. Falsely mapped, missing or low terminology resources decreased performances. Generizability and reusability were impacted, as realizations of the same tasks with the same resources in two different languages might not be applicable in the same manner.

Chapter 1

Introduction

Free-text data, such as in clinical narratives (clinical reports, summaries, etc.) and in database fields within electronic health record (EHRs) implementations, exist to support documentation in healthcare in a form that is suitable for human-to-human communication. Implementations for documentation vary across jurisdictions and institutions, as well as the services provided to them by the healthcare provider, from analogous methods, such as handwriting, to dictation, which is processed by human typists or medical speech-recognition software, and transformed into written text. In recent years, the rate of clinicians having to type the necessary information into the EHRs by themselves has increased [1]. Among clinicians, EHRs are often seen as a burden, even though the utility of EHRs bring added value such as data management, clinical decision support, and tracking administrative and operational costs. A study on the role of EHRs in emergency departments demonstrated that EHRs increased the workload and were not optimized to fit the clinicians' fast-paced workflow. Moy et al. [2] conducted a systematic review on clinical documentation burden and found multiple factors that contribute to this problem, often with the result that clinicians work extra hours completing documentation and reviewing tasks at home. In these circumstances, clinicians often document information quickly and concisely, using jargon and abbreviations, to save time and focus on interpersonal communication. As a result, detailed structured formats, such as tables are rarely used, leaving critical information locked in complex, unstructured texts intended primarily for other clinicians, disregarding secondary use scenarios of texts.

Globally, healthcare providers store vast amounts of data annually in various forms, e.g., imaging, text, laboratory results tables, etc. Language constantly evolves and mirrors the progress being made in medicine, and to process this large influx of data in an automated way to aid clinicians, natural language processing (NLP) systems can be implemented to make personalized medicine and real-time evaluations possible [3]. The rationales for using NLP in clinical data processing, include but are not limited to, several factors that aim to bridge the gap between clinical narratives and the need for standardized, structured documentation. These include enhancing the reusability of clinical data for both primary care and secondary research purposes, improving interoperability across institutions and linguistic groups, enabling efficient summarization of complex patient histories, and supporting predictive analytics and health surveillance [4].

The field of clinical NLP improved in recent years as new methods, such as transformer architectures [5], and better processing capabilities, made machine learning (ML) methods widely applicable in various settings [6]. A review on deep learning in clinical NLP by Wu et al. [7] showcased that almost 90% of relevant literature was on information extraction tasks. Recognizing linguistic patterns in clinical narratives and assigning labels are at the core of these tasks, which is challenging due to the writing style of clinicians, where multiple factors impede information extraction, such as redundancy, linguistic variations, jargon expressions, and short form content. These linguistic variations emphasize the need for approaches to address out-of-vocabulary content.

1.1 Definition of Non-Lexical Entities

Lexicality is connected to a resource, which differentiates entities[1] of language on their linguistic characteristics, whether these are found in the necessary domain or not. Non-Lexical Entities (NLEs) in context of this work are defined as "***meaningful text passages not found in a domain-specific lexicon or dictionary***".[2] In NLP, OOV (out-of-vocabulary) content is a common but very broad concept. This concept is refined to NLEs.

There are several types of NLEs in clinical narratives, particularly with reference to clinical jargon expressions.

- **Misspellings.** Misspellings occur when a word is written incorrectly because the writer does not know or is unsure of the correct spelling. These errors can stem from misunderstanding word patterns, rules, or simply from learning the wrong spelling, e.g., cognitive orthographic errors by writers using a second language or having dyslexia. Examples: "Diabetis", corr.: "Diabetes" (diabetes); "Hipertyreose", corr.: "Hyperthyreose" (hyperthyroidism).

- **Mistyping.** Mistypings happen when the writer knows the correct spelling but accidentally presses the wrong keys while typing, i.e., inexact strokes, neighboring keys are hit, not corrected for lack of proofreading. Examples: "Rippenserien-draktur" (corr.: "Rippenserienfraktur" (serial rib fracture)), "Kofpshcmerz" (corr.: "Kopfschmerz" (headache)).

- **Transcription errors, speech recognition errors.** Transcription errors arise when spoken words are inaccurately converted into written text due to mishearing, misinterpretation, or typographical mistakes by the transcriber. Speech recognition errors happen when automated systems incorrectly transcribe spoken language, often due to background noise, accents, or unclear pronunciation. Examples: "mexikanische Aortenklappe" (mexican aortic valve); corr.: "mechanische

[1] Text spans refer to things (entities) in the universe of discourse (domain). These are referred to by names or terms. Named entity recognition (NER) is the technique to identify text spans that mention a given entity by identifying its name. Note that in the NER contexts the "entity mentions" are often just called "entities". In this work, "text span" will be used instead to avoid confusion.

[2] Both words, lexicon and dictionary, are used mostly interchangeably. According to Merriam-Webster, a dictionary is defined as "*a reference source in print or electronic form containing words usually alphabetically arranged along with information about their forms, pronunciations, functions, etymologies, meanings, and syntactic and idiomatic uses*" [8].

Aortenklappe" (mechanical aortic valve, "positiv auf am vitamine" (positive for am vitamins) (corr.: "positiv auf Amphetamine" (positive for amphetamines)).

- **Optical Character Recognition (OCR) Errors.** This happens when text is inaccurately converted from images to digital format due to poor image quality, complex fonts, or misinterpretation by the OCR software. Example: "TYPL Diabetes" (corr.: "TYP I Diabetes" (type 1 diabetes)).

- **Spelling variants.** Spelling variants are alternative correct spellings of a word that differ based on regional, historical, or contextual factors. These are particularly common in German clinical language by variation of the letters "c", "z" and "k". Examples: "cerebral" (corr.: "zerebral"), "Magenulcus" (corr.: "Magenulkus" (stomach ulcer)).

- **Acronyms.** Acronyms are specialized forms of abbreviations that are mostly formed on the basis of initial letters of a series of words. In clinical practice, these are often not introduced as required for published texts, which leads to ambiguities. Acronyms may include abbreviations of proper names. Examples: acronym "MM" for "malignes Melanom" (malignant melanoma),"Morbus Menière" (disorder of the inner ear); "BHB", which corresponds to the long form "Krankenhaus der Barmherzigen Brüder" (abbreviated hospital name).

- **Non-acronym abbreviations or Dot-based abbreviations.** These are shortened words, created by truncating a word at a place where the meaning is clear from context, and the rest of the letters are omitted or replaced by a dot, and have a high variability. Dot-based short forms are more common in German, while abbreviations without dots are more common in English. Examples: "chron." for "chronisch" (chronic); "lymphozyteninf. Schleimhaut" ("lymphozyteninfiltrierte Schleimhaut" (lymphocyte infiltration of mucosa)); "Fx" for "fracture".

- **Single-Character Abbreviations.** This type of abbreviation are simplified representations of words, most often only consist of the initial starting letter of the word. Depending on the language, dots are either used or not used in documentation. Examples: "N" or "N." for "Nerv" (nerve) or "Neoplasie" (neoplasia); "V." for "Vena" (vein) or "Vulnus" (wound); "M." for "Musculus" (muscle) or "Morbus" (disease).

- **Conventionalized Abbreviations.** Conventionalized abbreviations are standardized and widely accepted shortened forms of words or phrases, such as "etc" for "et cetera", commonly used in written and spoken language. These appear in clinical narratives often without dot, typical in concise clinical jargon. Examples: "Leukos" for "Leukozyten" (leukocytes), "Gabla" for "Gallenblase" (gallbladder).

- **Ellipses**: Ellipses are intentional omissions of term components in texts. Example: "disseminierte Nekrosen im Groß- und Kleinhirn" (disseminated necrosis in the cerebrum and cerebellum).

- **Single-word compounds.** As the name suggests, these types of words are created by combining one or more words into one, to create a new compound word. Examples: "Coronaangst" (fear of COVID-19); "Opiatverschreibung" (opiate prescription); "Maushand" or "Mausarm" (compound word used as a reference for repetitive strain injury of the hand or arm); "verhaltensoriginell" (behaviorally abnormal).

- **Multi-word terms.** Phrases or words that together represent an entity. These are often newly added and might remain untranslated. Examples: "Long Covid"; "COPD cough" (chronic obstructive pulmonary disease cough).

- **Codes.** These are standardized alphanumeric identifiers used to document diagnoses, procedures, and medical services, and aggregated in medical lexicons and terminologies. Non-domain-related codes also occur in clinical narratives, such as zone improvement plan (ZIP) codes, bank account numbers, phone numbers, email addresses, uniform resource locators (URLs), etc. Examples: from clinical coding system, the International Classification of Diseases (ICD-10): "C90.0" (multiple myeloma)); "109989006" code from SNOMED CT for "multiple myeloma (disorder)"; 8036 (ZIP code for Graz, Austria).

- **Numeric expressions.** Numeric expressions in clinical narratives are quantitative data points, such as measurements, dosages, and lab values, and provide precise and essential information about a patient's health status and treatment. These particular NLEs should be distinguished from terms that include digits according to the lexicon. Examples: "Multigravida 4" (female patient pregnant four times); "Niereninsuffizienz Grad 4" (renal failure stage 4). Furthermore, specialized forms of numeric expressions, which are used to denote stages of diseases,

grades of conditions, or classifications of medical procedures, provide a standardized and clear way to convey critical clinical information. Due to using letters to convey that information, these expressions can be mistaken for acronyms. Example: "IV" for "intravenous" or "four".

- **Brand names.** Descriptions of brand names consist of proprietary names of medications or medical devices, specifically in clinical narratives listed in notes describing the treatment plans and prescriptions. These distinguish themselves from their generic counterparts, i.e., drug ingredients, and ensure specific identification and documentation. Examples: "Kalinor" (tablets for the normalization of potassium balance); "Neo Safe T CU 380" (specific version of an intrauterine device).

- **Proper names of persons, institutions, geographic locations.** These types of NLEs identify individuals, healthcare facilities, and places, to provide context and clarity to patient histories, treatment locations, and care coordination. Examples: "Mayo Clinic" (healthcare facility, treatment location); "Dr. John Doe" (treating physician).

- **Inflection forms.** Inflected forms of lexical entries either are not or only indirectly handled in lexicons, such as declension, and conjugations. Example: "diagnoses, diagnosed, diagnosing" are varied conjugations of the lexical entry "diagnose".

1.2 Processing of Non-Lexical Entities

NLEs have three cornerstones: identification, expansion, and disambiguation in NLP. With reference to NLEs, identification refers to the detection and delineation of NLEs in clinical narratives, expansion covers the evaluation and processing required to extrapolate possible senses and meanings from suitable lexical resources, and disambiguation deals with the final selection and linkage of the NLE to its corresponding entry in the lexical resource. The rationale is to systematically process NLEs to ensure accurate interpretation and integration into lexical resources.

There are two top-level approaches to process non-lexical content, which are (i) to *extend existing lexicons*, or (ii) to *map NLEs to existing lexicons*. The aim is the

enrichment of lexical resources, addressing gaps in dictionaries, and maintaining the relevance and utility of domain dictionaries.

The *extension of lexicons* can be accomplished in a variety of ways, and three of the most common ones are manual completion, production rules, or machine translation. First, using manual lexicon completion, domain experts annotate datasets for lexicon expansion. This comes at a high cost, as any updates to the lexical resources need to be done manually as well, and constant updates would be needed due to the high productivity of domain languages. Second, production rules cover all variations of existing lexical entries and possible new entries, such as compounds, spelling variants, inflections and derivations, and add these variations to the resources automatically. While this can be a high-throughput methodological approach for expansion, the tendency of expanding a lexical resource with any and all possible spelling variations tends towards combinatorial explosion and a need for further domain corpora to differentiate between main lexical entities and NLEs. Furthermore, the addition of lexical entries via rules and regular expressions, if covered by the rules, might increase false positive rates in automated text processing, i.e., instances where text spans are incorrectly identified or categorized due to the misapplication of the patterns defined in the rules. These effects need to be evaluated to see how such methodologies can help or hinder the processing of NLEs depending on the processing goal. Third, machine translation and subsequent expert validation can be applied, especially for low-resource languages, to expand lexicons in a given domain, with the help of high-resource languages, such as English.

Mapping NLEs to existing entries in lexical resources consist of two types: mapping rules and machine learning applications. In 2020, a systematic review by Kersloot et al. [9] evaluated the state of NLP in developing and assessing methods to map clinical texts onto ontology concepts. The review aimed to analyze the diverse NLP approaches for short forms. The variety of data sources and applications highlighted multiple evaluation methods for mapping text spans to ontological representations. The predominant objectives of the included studies covered information extraction, information enrichment, classification tasks. Generally, mapping rules encompass automatically mapping a language expression to an ontology or dictionary, which can lead to false mappings, particularly with short words. ML algorithms to identify NLEs, such as acronyms, non-standard linguistic variations, or to identify synonymous expressions, require a large quantity of annotated corpora to allow the creation of ML models with these func-

tionalities. In many ways, language models trained on annotated corpora have been scarce depending on the language and the domain being processed. In German, clinical language models were trained to become openly available for use in research. For instance, BioGottBERT [10], a biomedical German language model trained on medical and clinical datasets. Other languages with lower resources still remain without openly available language models in the given problem domain.

NLEs are always connected to a domain. If the term[3] itself is not found in a domain dictionary with the desired meaning, the text span is rendered to be non-lexical. An example for this might for instance be the German compound term "Heuschnupfen" (hay fever). In the international edition of SNOMED CT, synonymous terms "hayfever" or "hay fever" reference the lexical entry in SNOMED CT, with ID 21719001, where the fully specified name of the entry is "Allergic rhinitis caused by pollen (disorder)". Comparing this entry to the Austrian or German releases of SNOMED CT, the term "Heuschupfen" on its own is not represented in this manner, and would be considered non-lexical. The nearest entry, in the corresponding German or Austrian releases at the time of writing, for the term "Heuschnupfen", would be the compound term "Heuschnupfen-Konjunktivitis" (Hay fever conjunctivitis), which would be a false classification for the initial text span.

If one or multiple dictionaries together do not identify the term with that specific sense, then the term is classified as a NLE. As such classifications, on term as well as sense level, are far from trivial and need a lot of resources to be able to be done automatically, e.g., ML techniques, manual classification of terms by domain experts.

The representation of NLEs can pose hindrances to straight rule-based lookups as these entities often include inflected, derived, or contextually ambiguous forms that do not directly match their canonical or base forms. Application of lemmatizers would be able to resolve terms to their base forms, i.e., the verb form "prescribed" is transformed to the base infinite form "prescribe". By applying lemmatizers in the workflow, NLEs can be inherently linked to their base forms before determining their sense. However, the effectiveness of lemmatizers depends on the availability and suitability of language resources for the specific domain. Lemmatizers can facilitate automated text processing,

[3]The words "term" or "terms" refer to words or phrases associated with a specific meaning, and refer to a lexical entry in a domain dictionary. According to Chute [11] and Zeng et al. [12], term identification as part of vocabulary development consists of two steps, which are "*(1) the identification of candidate strings (ie, words or phrases) in a domain and (2) the determination of which of these should be included in a vocabulary as "valid" terms, also called "termhood determination."*"

but errors such as false positives or negatives often arise, particularly when processing text at token[4] level without accounting for the sense or context. Clinical narratives, with their significant textual variability, can hinder the performance of NLP functions and toolkits, especially when lemmatizers are trained on general, non-medical texts.

1.3 Related Work for Non-Lexical Entities

Throughout this work, the focus for processing NLEs with NLP methods centers on clinical narratives. A systematic review by Spasic and Nenadic [13] analyzed clinical narratives used in conjunction with ML applications. The analysis showed that clinical data is heterogeneous in nature, and that annotations, provenance and data availability are main contributing factors, which influence data processing, prior to the application of ML methods. Annotations or labels assign certain characteristics to text based data, so that in turn probabilities and ML models can be trained and used for different NLP tasks. Provenance, i.e., the source of the data, is varied as well, where sometimes only one type of document is used, e.g., discharge summaries or medication administrations. Data availability further influences how much data and what kind of clinical data is available for research purposes, and data privacy still plays a pivotal role in making real world data accessible. Additionally, the above mentioned review explained that the four most common topics covered in clinical NLP are classification tasks, named entity recognition, information extraction, and word sense disambiguation, in that order. The authors suggest focusing on applications, such as data augmentation or unsupervised learning, to approach NLP tasks in healthcare without annotated resources or to enhance clinical narrative processing through resource augmentation and transfer learning.

1.3.1 Identification of NLEs

The detection of NLEs is the first step to be able to automatically process NLEs. The easiest method to identify NLEs in NLP would be rules and regular expressions as explained in section 1.2. A rule-based approach is very focused on specific linguistic

[4]Tokens are a product of a process called tokenization. In NLP, to effectively process texts, it is necessary to break texts down into smaller components, which can be words, subwords, characters, or sentences, depending on the granularity required. For example, whitespace-tokenization would be separating an input text according to whitespaces.

traits or patterns in texts, which need to be updated and tested constantly, so that coverage and NLE-specific identification maintains a high accuracy. As the focus shifts towards ML-based approaches, larger amounts of labeled datasets are necessary for task-specific processing of clinical texts. State-of-the-art (SOTA) approaches for clinical named entity recognition (NER) utilize pre-trained language models, which are fine-tuned with domain-specific labeled data to accurately identify relevant entities. The necessity for data and current SOTA approaches still hold true with the current evolution in NLP towards large language models (LLMs). The promise of text-based prompts to contextually process and understand clinical data have been of interest for the research community since the introduction of LLMs.

In 2007, vocabulary development for consumer health information was investigated by Zeng et al. [14]. The aim was to automate term recognition, and consisted of three parts: candidate string extraction, manual review by human annotators, and application of methods to the curated dataset for testing. String matching for the extraction of candidate strings from the UMLS Metathesaurus [15] was done. Pre-processing included removal of symbols, stemming, normalization, and truncation. With the creation and analysis of n-grams, annotators reviewed and assessed the most frequent n-grams for eligibility to be included in the vocabulary. Two methods for automated classification of term recognition were implemented, a logistic regression model and a candidate collection formula. The logistic regression model seemed to be very effective for the identification of new terms, and achieved an AUROC (area under the receiver operating characteristic) of 0.95. In 2022, Faris et al. [16] implemented disease symptom identification for Arabic text-based medical consultations. The method used features from AraBERT [17] embeddings and fine-tuned on a Bidirectional Long Short-Term Memory (BiLSTM) [18] classifier. Based on manual evaluation of the identified symptom mentions, the method reached 0.706 in recall. In 2024, a study on disease symptom mentions and their identification by Sogandi [19] with supervised and unsupervised ML methods showed that a statistical method, such as statistical regression, outperformed other methods in accuracy with an accuracy of 0.84. Regarding F1-measure, the best performing method remained a combination of bootstrap aggregation and random forest for an F1-measure of 0.86.

Specifically on the topic of detecting short forms, Xu et al. [20] investigated abbreviation detection in clinical notes with four different methods in 2007. First, a baseline method, a lookup operation, compared all tokens with two word lists to find unknown

tokens, i.e., abbreviations. Secondly, a heuristic rule-based method with rules was combined with the lookup word list implementation from the baseline method. Third and fourth, both consisted of decision tree classifiers with different features, such as word formation, word frequency, while the fourth method also included outside knowledge sources. The results showed that the best performing method with a precision of 0.91 was the fourth method, i.e., the decision tree classifier. The worst performance was recorded with the baseline method by processing each token in discharge summaries and comparing them to known English word lists. Similarly in 2011, Kim et al. [21] demonstrated an abbreviation detection modeling approach. It consisted of document pre-processing of clinical narratives with a sentence splitter and a tokenizer, and joined with a specially trained clinical Text Analysis and Knowledge Extraction System (cTAKES) [22] model and the LRABR, list of abbreviations and acronyms, part of UMLS. This modeling approach achieved 0.66 for exact matches, and 0.75 for partial matches. From the error analysis, the most limiting factor for the detection of short forms was the list with which the model was trained, and therefore missing a lot of short forms based on their test set.

1.3.2 Expansion of Lexicons with NLEs

Clinical lexicons require continuous expansion to remain accurate and comprehensive, accommodating new terminology[5], advancements in research, and emerging healthcare practices. Two main approaches can be followed to automate the expansion of lexicons: manual, or data-driven solutions. Rules and regular expressions can be manually created to find new entries for a specific domain dictionary, or implementation of ML methods can help with the expansion of lexicons. Additionally, a data-driven solution would be able to aggregate semantic and lexical information found in clinical narratives, and allow to further create new resources, or to better connect and update existing resources. The lexical entries can be used as starting point to determine classifications or NLE types, and to ascertain their relationship or connection to other entries in the lexicon. Another possibility would be to prompt for further lexical entries with LLMs,

[5]A *terminology* is a standardized and structured set of terms, which "*are ideally unambiguous and self-explanatory. This often does not reflect clinical language use*". [4] "*Standardised medical terminology is a set of linguistic conventions and standardised vocabulary for the description of medical concepts and processes. It is used to ensure the accurate and consistent communication of information, and to facilitate the exchange of data between healthcare professionals, patients, and healthcare delivery organisations.*" [23–25] An example for a terminology in the clinical domain is SNOMED CT.

which would be one of the newest data-driven methods with regard to lexicon creation and expansion.

Relevant literature for the named two approaches for lexicon expansion should give a short overview of the topic. In 2020, Sarker [26] developed a system that automatically expands concepts in biomedical texts. Particularly due to three subtypes of NLEs, i.e., non-standard expressions, misspellings, and abbreviations, this system was developed to support lexicon creation processes. Built with threshold decay functions, dense vector and lexical similarities, the semantic similarity of terms was used to generate lexical variants of multi-word expressions. Methods, such as n-gram creation, Levensthein [27] ratios, semantic and lexical filters, and manually set thresholds, which were applied previously on different datasets, e.g., the use of this system for processing adverse drug reaction mentions by Sarker and Gonzalez [28] reached an F1-measure of 0.812 for text classification. Koroleva et al. [29] analyzed clinical trial outcomes with pre-trained language models, BERT [30], BioBERT [31], and SciBERT [32]. BioBERT achieved the best results and outperformed baseline measures, e.g., Levensthein [27] distance, statistical measures, etc., and reached an F1-measure of 0.89. With the addition of variants of outcomes, i.e., generating an extended dataset in which ambiguities such as abbreviations were automatically replaced with their long forms prior to training. The latter resulted in an improved F1-measure of 0.93. In 2023, Carpenter and Altman [33] searched for synonymous expressions for drug abuse through the application of the LLM GPT-3, filtering and cross-referencing generated terms to existing drug names and online resource, i.e., colloquial term candidates for drug names could be found. The combined method of LLM, filtering, and cross-referencing reached a precision of 0.77, and an F1-measure of 0.52.

1.3.3 Disambiguation of NLEs

Disambiguation of NLEs means using ML methods to assign a valid label, such as a lexical entry, to the NLE based on the context. The many possibilities to implement disambiguation capabilities vary from rule-based methods, to processing lexical resources for correctly expanded terms, to evaluations of deep learning methods via data driven methods, e.g., embeddings or LLM representations. To effectively investigate the disambiguation of NLEs, the datasets need to be extensively checked and curated prior to the implementation of a method. As one large focus of disambigua-

tion in related works in clinical NLP was on abbreviations, due to their ambiguity, the disambiguation of short forms was chosen as a special focus of this dissertation work.

Related works on short-form disambiguation give an overview of SOTA methods: In 2012, Wu et al. [34] tested three existent clinical NLP systems, MetaMap [35], MedLEE [36, 37], and cTAKES [22], to resolve abbreviations in 32 discharge summaries from the Vanderbilt Medical Center. MedLEE demonstrated the best performance across all evaluations compared to the other NLP systems, with an F1-measure of 0.60 for all abbreviations, 0.70 for clinically relevant abbreviations, and finally, 0.73 for a subset of ambiguous abbreviations. In 2015, Wu et al. [38] showed that training a Support Vector Machine (SVM) [39] model with additional 10-fold cross validation for each of the selected 25 abbreviations resulted in 0.89 in average accuracy, and outperformed the baseline majority sense classifier with 0.73 in average accuracy. In 2016, Mowery et al. [40] described the outcome of the ShARe/CLEF[6] eHealth Challenge 2013 Task 2 [41], giving a methods overview of participating teams for short form normalization and disambiguation. The performance of the five participating teams achieved accuracy ranges from 0.43 to 0.72. Most teams used cTAKES, Conditional Random Field (CRF) [42], custom lexicons, and online resources for the creation of the training data and the disambiguation system. Adams et al. [43] used the local context around acronyms and their metadata to draw word embeddings. On the Clinical Abbreviation Sense Inventory (CASI) [44], the embeddings approach performed reasonable well with an accuracy of 0.71. In 2024, Hosseini et al. [45] investigated abbreviation disambiguation with a BiLSTM model. Due to the imbalance with low-resourced clinical datasets, i.e., inconsistent text span counts for each abbreviation sense extracted from clinical narratives, the dataset was pre-processed to balance senses by curating additional examples with reverse substitution[7] from MeDAL [46], an abbreviation disambiguation dataset created from a medical source dataset. Word embeddings and the BiLSTM model were applied for disambiguation of abbreviations from the CASI dataset. On the re-balanced dataset, the BiLSTM approach reached an accuracy of 0.96, which is a slight improvement compared to the application on the imbalanced dataset of 0.94. Similarly in 2022, Agrawal et al. [47] researched the capabilities of LLMs to disambiguate acronyms from the CASI dataset. With GPT-3 edit in a zero-shot scenario, the method reached an accuracy of 0.86, and macro F1-measure of 0.69.

[6]Shared Annotated Resources/Conference and Labs of the Evaluation Forum

[7]Reverse substitution means aggregating examples, where the long form version of a short form is found, and then replacing the long form expansions with the corresponding short form.

1.4 Research Questions

Several research questions were stated as part of the published investigations, and the overarching themes of all research questions were summarized in the following way:

- Do specialized methods for NLEs improve clinical text processing?
- Do LLMs yield better results than traditional ML methods?
- Are there differences in performance when comparing languages?

The following chapters are organized as follows: The Materials and Methods section elaborates on the datasets, language models, and evaluation measures used across investigations. All investigations pertaining to answering the research questions are described separately in the main body of this dissertation. In five chapters, a systematic scoping review on short forms, the three cornerstones for processing NLEs, and text classification with NLEs, the utility of NLP for NLE processing is shown. Each investigation is further subdivided into Background and Significance, Materials and Methods, and Results and Discussion. Finally, the Discussion summarizes the scientific knowledge gain, answers the research questions, and gives a conclusion and outlook towards the impact of the results in context of existing research.

Chapter 2

Materials and Methods

Due to the variety of non-lexical entity (NLE) types and multitude of source materials available, from the local hospital information systems to available online resources, machine learning (ML) methodologies were applied in conjunction with the described datasets (see section 2.1), language models (see section 2.2), and evaluation measures (see section 2.3). An overview of the materials and methods used in this dissertation are provided here for overall understanding. General descriptions of natural language processing (NLP) methods for NLE processing can be found in the systematic scoping review on short forms by Kugic et al. [48] in the section "Overview of Methodologies". A more detailed description of each dataset and applied ML approach, is presented within the "Materials and Methods" sections in the respective chapters of this dissertation.

2.1 Datasets

2.1.1 Clinical Terminologies and Lexicons

Terminologies, such as UMLS and SNOMED CT, and lexicons can be used as lexical repositories of information for NLP tasks.

- **Unified Medical Language System (UMLS).** The UMLS is a hybrid, comprehensive biomedical terminology resource [15] developed by the U.S. National Library of Medicine to integrate diverse medical vocabularies. It includes a Metathesaurus of linked terms, a Semantic Network for categorizing concepts,

and an English Lexicon to support natural language processing in healthcare applications.

- **Standardized Nomenclature of Medicine Clinical Terms (SNOMED CT).** SNOMED CT is a large ontology-based clinical terminology system encompassing all fields of medicine and made available as an international standard, promoted by SNOMED International. It contains more than 360,000 concepts, and nearly a million English terms[1]. Although the main use case for SNOMED CT has been the standardization of structured data, its use for NLP is increasingly investigated [49].

- **Medical Dictionaries.** General medical dictionaries in German, Croatian, and English were used as manual and automated lookups for data annotations or deep learning implementations.

2.1.2 Clinical Narratives and Health Forums

- **Medical Information Mart for Intensive Care (MIMIC).** MIMIC is a collection of de-identified EHRs for over 40,000 patients in English comprising not only of clinical narratives, but furthermore include laboratory test results, measurements of various kinds, medical imaging, radiology reports, and other medical data. This resource exists in different editions, with MIMIC-IV [50] being the most up-to-date one, to date.

- **n2c2 Challenges.** Formerly known as i2b2 challenges (Informatics for Integrating Biology and the Bedside), and in subsequent years administered by the Harvard Medical School and George Mason University under the new name National NLP Clinical Challenges (n2c2). The goal was to provide de-identified annotated datasets to the research community with a medical informatics task as part of these challenges. The participating teams would then develop a method with the training set, and test the method on a hold-out set that the organizers provided. These datasets, starting from 2006 onward, are available at the homepage[2] to be used on premise for the implementation and testing of new tasks.

[1] https://www.snomed.org/what-is-snomed-ct
[2] https://n2c2.dbmi.hms.harvard.edu/data-sets

- **SemClinBr.** The Portuguese dataset "SemClinBr" is an annotated multi-institutional and multi-specialty corpus comprising 1,000 clinical notes, with over 65,000 entities and 11,000 relationships, designed to support biomedical research and clinical NLP tasks in Brazilian Portuguese [51].

- **German Clinical Narratives.** Two collections of de-identified clinical narratives in German were used for investigations. The first collection consisted of discharge summaries from dermatology, cardiology, and oncology departments. The second collection comprised approx. 1.9 million unique de-identified clinician-authored problem list entries for diagnosis coding at patient discharge.

- **Medical Health Forums.** Similar characteristics of NLEs cannot only be found in clinical narratives. Freely available online forum conversations share characteristics of clinical narratives. In particular, health forums offer similar context and linguistic characteristics in the medical domain, as medical professionals do not only author discharge summaries or other types of documents from a hospital provider or independent clinician, but also answer, in certain cases, patient-authored questions, online. These texts can be additional resources for clinical research, if clinician authored texts are not available.

2.1.3 Short Form Sense Inventories

Short forms are a broad categorical term referring to all subtypes of abbreviations, which include acronyms. From section 1.1, we had established that multiple versions of short forms exist, which differ slightly in representation, formation, and definition. To be able to automatically process these types of short forms with NLP, inventories of short forms are needed to identify, expand, and disambiguate them. Short form sense inventories are a type of lexicon, which contain possible long forms of short forms. These inventories were created in the beginning manually, and later through automated means, so that possible meanings of short forms are available for automated processing. The minimum requirement for short form sense inventories are short forms, which are mapped to all their possible long forms, as most abbreviations are ambiguous and not easily disambiguated. Some inventories add more information, metadata, if it is needed during processing, as well as examples, which contextually represent the short form.

- **Clinical Abbreviation Sense Inventory (CASI).** The anonymized CASI dataset [44, 52] was curated by the University of Minnesota consists of specialized abbreviations, mostly acronyms, with context and metadata information from clinical narratives, and their resolutions, to establish an English clinical narrative dataset with which NLP applications can be developed and evaluated. The dataset contains approximately 37,500 entries, which also allows for machine learning applications to be tested with this dataset.

- **Abbreviation Lexicons.** Due to the complexity of short forms, and common short forms occurring in clinical narratives, more generalized abbreviations can be a possible resource for the processing of short forms in clinical narratives. Multiple print and online resource lexicons are available. For this dissertation, reference short form corpora were used as manual lookup resource during the implementation of ML and NLP processing pipelines, and for the manual annotation of results in two languages, i.e., German [53, 54], and Croatian [55–60]. For English, the above described CASI dataset was used. For the annotation of Portuguese short forms, the domain expert was able to disambiguate short forms without additional lexicon resources.

2.2 Language Models

ML implementations often use language models to enable efficient processing of unstructured medical data for tasks, such as clinical text understanding, predictive analytics, and workflow automation. They offer benefits like scalability, adaptability, and improved accuracy, particularly with domain-specific models, such as BioBERT [31] and ClinicalBERT [61]. However, challenges like data privacy, domain specificity, interpretability, and bias must be addressed to fully harness their potential in healthcare [62, 63]. In this dissertation, similarly to Ling et al. [64] and Li [65], Large Language Models (LLMs) and pre-trained language models are seen as separate types of models. Pre-trained language models, such as BERT [30] models, are defined by Ling et al. [64] as models that "*process input text into vector representations without an explicit decoding phase to generate new text. Instead, they transform and embed text into a high-dimensional space*". Additionally, LLMs, such as Generative Pretrained Transformer (GPT) models, are described as "*autoregressive language models that generate the next word in a sequence based on previous words. They map a sequence of tokens*

to a vector representation and generate contextually relevant content autoregressively, calculating the probability of the next token based on the context" [64]. For diverse datasets, diverse language models are needed to fully exploit the ML capabilities when these are applied. The following constitutes a list of language models that were applied for a variety of tasks.

- **BERT.** The BERT (Bidirectional Encoder Representations from Transformers) [30] language model[3] is a multilingual transformer-based model pre-trained on a large corpus of text from 104 languages using a cased vocabulary, i.e., during training the casing was not eliminated, instead the distinction between uppercase and lowercase letters was kept.

- **SapBERT-XLMR.** The SapBERT [66] (Self-alignment Pretraining for BERT) Cross-lingual SapBERT-XLMR model is designed for biomedical term alignment across languages, leveraging a multilingual embedding space for cross-lingual knowledge representation. It combines self-alignment objectives with pre-trained language models, i.e., XLM-RoBERTa [67], to map semantically similar terms, even in different languages, into close vector representations for ML tasks, e.g., biomedical concept linking, multilingual search, etc. The XLM-Roberta model was applied, in combination with SapBERT, for its performance gains in cross-lingual transfer learning tasks.

- **BERTic.** The BERTic [68] language model is a transformer-based model specifically designed for Bosnian, Croatian, Serbian, and Montenegrin languages. This model was pre-trained on a large corpus of regional texts for classification and named entity recognition tasks. Being an ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately)-based model [69] means it uses a computationally efficient training method that involves a generator-discriminator setup, enabling the model to learn robust representations by detecting replaced tokens instead of simply predicting masked tokens.

- **GPT-3.5-turbo.** The GPT-3.5-turbo model [70] is a generative language model, classified as a LLM, by the company OpenAI built on transformer architectures, leveraging large-scale pre-training on diverse datasets, to synthesize texts based on contextual prompts.

[3]https://huggingface.co/google-bert/bert-base-multilingual-cased

- **GPT-4.** The GPT-4 model is an enhanced version of OpenAI's generative LLMs. It is built with transformer architectures used in GPT-3.5-turbo with significant improvements in scale, training techniques, and data diversity. While GPT-3.5-turbo was optimized for efficiency and real-time performance, GPT-4 is reported to have a larger model capacity, more refined fine-tuning processes, and a broader dataset, enabling it to understand and generate more nuanced and contextually complex responses with better reasoning and fewer biases.

- **Llama-2-7b-chat.** The Llama-2-7b-chat model [71] is an open source LLM developed by Meta, designed for conversational tasks and fine-tuned using reinforcement learning from human feedback to improve response quality and alignment with user intent. It is built on a transformer architecture with 7 billion parameters, allowing it to handle a range of language generation tasks with reasonable computational efficiency.

- **Llama-2-70b-chat.** The Llama-2-70b-chat model [71] is a more advanced, large-scale open source LLM by Meta, optimized for complex conversational AI applications with 70 billion parameters.

2.3 Evaluation Measures

Any NLP task, which was implemented and tested, needed to be evaluated, so that the effect of the newly implemented ML approach could be gauged. The evaluation needed to be in-line with the the aim, data and methods of the implemented approach, so that a comparative analysis could be performed. Based on this kind of information, the following five metrics were most often used for evaluation of NLP tasks: precision, recall, F1-measure, accuracy, and mean average precision (MAP).

2.3.1 Unranked Metrics

The following values need to be defined as these are required for calculating the unranked metrics. True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) represent the outcomes of the model's predictions compared to the actual labels. TP occur when a model correctly predicts a positive outcome, TN

when it correctly predicts a negative outcome, FP when it incorrectly predicts a negative outcome as positive, and FN when it incorrectly predicts a positive outcome, labeling it as negative instead.

The precision, recall, and F1 measures are commonly used metrics for evaluating the performance of classification models. Precision is a metric that measures the proportion of correctly identified positive instances among all predicted instances.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.1}$$

Recall is a similar metric to precision, which measures the proportion of correctly identified positive instances among all relevant instances.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.2}$$

When combining the metrics precision and recall, the F1-measure can be calculated.

$$\text{F1-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.3}$$

Accuracy measures the proportion of correct predictions made by a model out of the total number of predictions.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2.4}$$

2.3.2 Ranked Metrics

Mean Average Precision (MAP) is a rank-based evaluation metric commonly used in information retrieval to assess the quality of ranked lists produced for multiple queries.

For a given query q, assume there are R relevant documents, and the ranked positions of these relevant documents are denoted as $K_1, K_2, \ldots, K_R$. The Average Precision (AP) for the query is defined as:

$$\text{AP} = \frac{1}{R} \sum_{i=1}^{R} P(K_i), \tag{2.5}$$

where $P(K_i)$ is the precision at rank K_i, which is the fraction of relevant documents among the top K_i ranked documents. If a relevant document is not retrieved, its corresponding precision is considered to be zero.

The Mean Average Precision at K (MAP@K) across multiple queries is computed as the mean of the Average Precision (AP) values for all queries. Let Q denote the total number of queries.

$$\text{MAP@K} = \frac{1}{Q} \sum_{q=1}^{Q} \text{AP}_q. \tag{2.6}$$

Chapter 3

Systematic Scoping Review on Short Form Non-Lexical Entities

3.1 Background and Significance

The aim was to summarize and review the processing methods of all types of NLEs with NLP methods. Due to the vast amount of literature published about NLEs and their varied descriptions and multiple possible variations, a more nuanced and focused scope of the review on short forms was undertaken. The term "short forms" in context of this research is applied to summarize all types of abbreviations, such as acronyms, in one term. From the section on types of NLEs, it is noted that various types of short forms exist and are grouped together for better clarity. From existing research, it was not possible to find out which method works best for the three sub tasks, identification, expansion, and disambiguation, in analyzing and processing short forms with NLP. Additionally, the categorization of short form types, as well as analyzing processing methodologies, were other important aspects. The performed systematic scoping review [48] on this topic not only helped in understanding this research topic better but also structured the content and answered the above stated research questions.

3.2 Materials and Methods

To enable a systematic scoping review, a search strategy needed to be formulated. It should encompass the whole research landscape in context of this topic, be formulated

in a way that it would be manageable by the review team, while making sure that related research would not excluded. As part of the prerequisites to starting a review, the size of the review team needed to be determined for better task management and organization of this research project. The review team included four members handling most of the review work and three supervisors overseeing quality checks and providing decision support. The type of review, the systematic scoping review, in conjunction with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, determined the workflow [72]. First, five different literature databases, Web of Science, Embase, MEDLINE, EBMR (Evidence-Based Medicine Reviews), and ACL Anthology, were searched with the explained search strategy. In the next step, the de-duplication of papers was undertaken, i.e., the same paper can be indexed in different literature databases, which meant manually deleting non-relevant papers from the aggregated articles. Inclusion and exclusion criteria were discussed within the team, and consisted of the following criteria:

- **Language.** If the article was written in a language other than German or English, the article was excluded. The two languages were chosen as all team members were fluent in both.

- **Collections and Reviews.** Due to the search strategy not including filters for specific types of papers, such as review articles, and the searched literature databases being sometimes incomplete, the exclusion of collections and review articles was decided and manually performed.

- **Peer-Reviewed Articles.** Articles without peer review were excluded, such as preprints made available on electronic print archival sites, e.g., arXiv. If these papers were found in the literature databases, in archival form, another manual search was done via search engines for the published version of the same article, and if that could not be found, the article was excluded.

- **NLP.** Statistical analyses of the distribution and counts of short form content in texts were excluded. Processing clinical narratives with NLP methods was required for an inclusion of an article.

- **Short Forms.** If NLP methods were applied, but short forms were not processed as the main aim of the investigation, the article was excluded.

- **Evaluation and Validation.** Performance metrics needed to be reported and evaluated by the authors of the articles. Without an evaluation of the reported method, the functionality and applicability of the method would not be able to be gauged.

- **Originality.** The application of an already published method on a dataset would not be enough for the inclusion of the article, but would be an acceptable way to compare the original method to related works. Articles needed to offer a unique method for processing short forms, and should that not be the case, the investigation was excluded.

- **Human Clinical Narratives.** The processed datasets had to be clinical narratives from human medicine. Veterinarian or other types of narratives were excluded due to the aim of analyzing short form processing in clinical narratives. This distinction was already made clear through the search queries that listed all varieties of medical sub-specialties, such as "ophthalmolog*" for "ophthalmology".

3.3 Results and Discussion

From 6,579 articles, 131 full texts were screened for eligibility based on the reported criteria. Finally, all review team members agreed on 19 papers to be included in the final analysis and synthesis of results. One of the 19 articles was found via additional reference searches. The data synthesis evaluated the articles according to multiple factors: types of short forms, dataset restrictions, overview of processing methodologies for the identification, expansion, and disambiguation of short forms, languages under investigation, and NLP recommendations.

Mainly shallow approaches were utilized for the identification of short forms, i.e., rules, regular expressions and string similarity methods. The expansion of short forms was separated into lookup and non-lookup methods for finding a viable expansion candidate for a short form. Lookup methods referenced an additional sense inventory or database, such as PubMed. Non-lookup methods did not use predefined lists or dictionaries, but instead used text mining techniques, end-to-end encoder models, or even active learning approaches to find possible long form versions of short forms. Disambiguation of short forms, due to the context dependency, was generally focused on embedding representations and deep learning.

Most investigations processed English clinical narrative source texts, while four papers processed short forms in languages other than English, i.e., Russian, Serbian, Spanish and Polish. For English, short form lexicons and sense inventories exist, which can be used to create ML approaches, such as the Clinical Abbreviation Sense Inventory (CASI) [44, 52]. For low-resource languages, the creation of a clinical sense inventory often is the first step in processing short forms, which was also the case for the four mentioned non-English articles.

Types of short forms, i.e., restrictions on the type of short form processed in the articles, consisted of shortend words, ad-hoc abbreviations, acronyms of various lengths, dot-based abbreviations, common abbreviations, and physician-specific abbreviations. A majority of the articles selected specific short form types to process, although seven articles did not place any restrictions upon the source dataset. Extrapolating the type of short form for each article was difficult to ascertain, mostly due to a lack of definition of the short form type. Even though one article might state to investigate abbreviations, in actuality, based on external dataset descriptions, pre-processing and examples given in the text, the review concluded that the article only processed a portion of all available acronyms. These limitations in analyzing only a subset of a dataset occur quite often, which then impacts the informative value and comparability of the results. Connected to this, the results would be impacted, where the assumption is that it would artificially improve results, and would only be applicable in certain scenarios, e.g., for three-character length acronyms. Although the reason for the selection of specific short form types can be traced back to dataset restrictions, errors, or methodological decisions made at the beginning of the investigations.

Furthermore, to assess the quality and reproducibility of the investigations, NLP recommendations stated by Kersloot et al. [9] were used as a guide and the values were extracted from the included articles in this scoping systematic review, e.g., source code availability, linking of external datasets, performance metrics, error analysis, etc. The majority only partially adhered to the recommendations due to missing error analyses, confusion matrix, and external validation.

The limitations of the review was the selected review time frame of five years that might have excluded relevant publications, and might have lead to an under-representation of articles applying certain methods, such as large language models (LLMs).

Future research should aim to adhere to NLP recommendations, formulate the type of content analyzed in the short form processing article, and investigate short form

processing in languages other than English. Restrictions on the dataset set prior to processing short forms could introduce bias into the developed model that could be mitigated with selecting other processing methods for short forms and/or additional debiasing steps.

Chapter 4

Identification of Non-Lexical Entities

Two investigations focused on the identification of NLEs with NLP and ML methods. First, a SOTA approach to identify NLEs in Croatian health forum entries was implemented [73], specifically focusing on four subcategories of NLEs: short forms, lexical variations, brand names and proper names. Second, as continuation of the named investigation, the application of large language models for the identification of NLEs was evaluated [74].

4.1 Identifying NLEs with Named Entity Recognition

4.1.1 Background and Significance

Named entity recognition (NER) methods are one possibility to detect meaningful text spans in a data driven way. Fine-tuning a model further helps with increasing performance and utilizing labeled data for a specific purpose in NLP. High resource languages, for instance English, have online available resources to process and perform NER tagging for the three most common text span types, people, locations, organizations, etc. The tagging of these text span types by general purpose language models do not match the text spans of interest in clinical narratives, which would need to focus more on diverse entity types, such as medication or symptom recognition. Recognizing text spans for these entities would need specially trained models based on annotated corpora. Lower resource languages that have no accessible model for NER tagging have

to go through the whole process of annotating the datasets, as well as training a model to establish this functionality.

Heryawan et al. [75] implemented a long short-term memory (LSTM) model to find abbreviations in medical texts. Through embedding vectors and adding further text spans via a sample generator to mitigate problems due to an imbalanced dataset, an F1-measure of 0.49 could be reached. For the identification and normalization of abbreviations, Huang et al. [76] used word embeddings and sequence labeling for deep learning via a sequence-to-sequence model that achieved an F1-measure of 0.84. The best performing method to detect medical events, temporal expressions, and values was a BERT implementation by Kaplar et al. [77], which reached an F1-measure of 0.86, and therefore outperformed conditional random fields, LSTM, and ensemble methods.

4.1.2 Materials and Methods

Dataset. In total, 12,023 distinct entries from health forums were extracted covering the specialties (but not limited to) ophthalmology, gynecology, cardiology, and radiology. The content of health forums contain a variety of health related posts. Users question other forum users and experts among them on information around diseases, symptoms, medications, etc. or post their clinical narratives with further questions, which were either not adequately answered by their treating physician or establishing the need for a second opinion. These features of health related forum posts make them a near interchangeable resource in comparison to clinical texts from the hospital information system for medical informatics, especially when there are no available resources freely available for a low-resource language. For this investigation, Croatian health forum entries[1] were used as a resource.

Data Curation. The curation of the dataset was performed with a self-implemented website crawler following the following steps: (i) to find all possible web links to each health forum entry to aggregate data sources, (ii) forum entries need to be health related, which was predetermined based on the content of the website, (iii) filter out any entries that do not contain physician authored texts, made possible via accessible metadata attached to each online entry, (iv) export and save the forum entries to a file for further processing.

[1] https://www.cybermed.hr/

Text Preprocessing. The content of health forums showed similar characteristics to mobile messaging or social media content with included emoticons, images, superfluous texts, advertisements, etc. The curation step filtered out images and all texts outside of the actual content written by forum users. The textual content was further adapted for normalization purposes by removing any characters outside of a predefined regular expression `[a-zA-ZćĆčČžŽšŠđĐ]` with whitespace characters, e.g., symbols, special characters, emojis, numeric expressions, etc., as well as collapsing and normalizing multiple whitespace characters existent due to the previous rule-based filter. Sentences, in particular greetings, which contained personal identifiers, such as forum usernames, were removed with rules, due to their ambiguous or nonsensical naming conventions. Other types of personal identifiers within health forums were kept without changes.

Data Annotation. The annotation of each text entry was done by a domain expert through manual dictionary lookups to support the categorization of the texts. The four categories that were annotated consisted of: (a) short forms, i.e., all variants of abbreviations and acronyms, (b) lexical variations, e.g., spelling mistakes, mistypings, non-standard variations of terms, dialect variations of the same terms, etc., (c) brand names, i.e., medication and drug names, product names based on their brands, e.g., Aspirin, the brand name used to document the intake of that medication with the drug substance or active substance "acetylsalicylic acid", (d) proper names, i.e., locations, people, or personnel referenced in health forum entries, i.e., a treatment rehabilitation center as a geographic location that is referenced during documentation. Two versions of the dataset were created, once for a binary classification to distinguish between lexical entities and non-lexical entities (NLEs) on main categorical non-lexical content (Main-NLC) level, and a second iteration applying the same methodology for a multi-label classification task, to differentiate between the four categories named above on a subcategorical NLC (Sub-NLC) level.

Annotated Dataset. The dataset was split to create three sets for training, validation, and testing. On sentence level, 80% of sentences were used for training, and each 10% of sentences for validation and testing. The annotated NLEs were transformed into a specific format needed for sequence modeling, called BIO-labeling (beginning-inside-outside), and saved in the CoNLL format, which was established by the same-named conference (Conference on Natural Language Learning). Tokens not pertaining to NLEs were annotated with outside labels, and beginning and inside labels were annotated and allocated to single or multi-token NLEs. The tokenization for this investigation

was done through whitespace character tokenization. In Table 4.1, an example of an annotated sentence in BIO-labeling format can be found.

Table 4.1: Excerpt from an annotation example for NLEs in the training dataset, which translates to "*Yesterday, the pulmonologist prescribed Flixotide [...]*" Reproduced from Kugic et al. [74] with permission from publisher Institut za hrvatski jezik (Institute for the Croatian Language).

Tokens	Annotation	Description
Jucer	B-LVR	lexical variant, misspelling/mistyping
je	O	lexical entry
pulmolog	O	lexical entry
oveo	O	lexical entry
Flixotide	B-BNE	pharmaceutical brand name

Model Architecture. A NER approach was implemented to classify the varieties of NLEs in medical texts. Only four NLE categories were annotated and focused on with reference to a domain dictionary of medical Croatian terms. Two language models were applied, a multi-lingual BERT (Bidirectional Encoder Representations from Transformers) [30] language model[2] encompassing more than 100 languages as a baseline, and a focused language specific ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) [69] model[3] covering Croatian, Bosnian, Slovene, and Montenegrin language variations [68]. In the published investigation [73], both token and entity level performance were evaluated, with precision, recall, and F1-measure for exact and partial boundary matches, calculated with MUC-5 (Message Understanding Conference) [78] metrics.

Model Fine-Tuning. The fine-tuning of the base models BERT and ELECTRA were implemented with the HuggingFace framework [79]. Mid-training evaluations enabled early stopping during the fine-tuning of the language models. The metric evaluation loss was chosen, and the fine-tuning process compared the evaluation loss between mid-training evaluations. If at those points, the difference between the calculated values was lower than 0.01, then the fine-tuning was stopped and the model saved. Otherwise training would continue until this case occurred.

Model Evaluation. The evaluation of the labeled test set was performed on two levels, token and entity level, for two categorical analyses, main and sub-categorical

[2]https://huggingface.co/google-bert/bert-base-multilingual-cased
[3]https://huggingface.co/classla/bcms-bertic

analyses of non-lexical content (NLC). Due to the whitespace tokenization and BIO-labeling of the text spans in the dataset, both categories were utilized for a detailed error analysis on token and entity level. The entity level would either correspond to partial or exact matches. The implementation was supported by a Python library that implemented the well-established MUC-5 metrics [78]. The subsequently performed error analysis was manually performed in analyzing all false positive and false negative annotations.

4.1.3 Results and Discussion

On subcategorical entity level, the language specific ELECTRA model outperformed the BERT-based model with an F1-measure of 0.928 for partial matches, and 0.914 for exact matches, in comparison to BERT partial matches with an F1-measure of 0.901, and exact matches of 0.886. Slight improvements for the main categorical level across both language models could be found, with an F1-measure in the range of 0.914 to 0.930, although these still follow the above mentioned trends. Results on token level, along with confusion matices, were published in the corresponding paper [73].

Table 4.2: Performance metrics for BERT and ELECTRA models on entity level.

Model	Category	Mode	Precision	Recall	F1-measure
BERT	Main-NLC	exact	0.921	0.921	0.915
BERT	Main-NLC	partial	0.935	0.936	0.930
BERT	Sub-NLC	exact	0.891	0.895	0.886
BERT	Sub-NLC	partial	0.906	0.910	0.901
ELECTRA	Main-NLC	exact	0.934	0.925	0.925
ELECTRA	Main-NLC	partial	0.946	0.938	0.938
ELECTRA	Sub-NLC	exact	0.913	0.923	0.914
ELECTRA	Sub-NLC	partial	0.927	0.938	0.928

The aim to detect NLEs in a data-driven manner was shown through this experiment to be possible. The prerequisites of obtaining a sufficiently large labeled dataset sometimes require too many annotation hours for this approach to be viable in clinical practice. Other solutions, such as mining for similar contextual expressions from existent datasets, as well as introducing methods that would continuously integrate annotations by human annotators, would in turn decrease the work load for annotators. The latter approach would focus more on contextually similar text spans, rather than

annotating entire text segments with commonly known expressions. Distinguishing between lexical and non-lexical entities usually depends on context. Possible reasons for the slight performance boost with the ELECTRA model compared to the BERT model could be related to the language specificity and/or on architecture of the ELECTRA model. As the ELECTRA model was only pre-trained on Bosnian, Croatian, Serbian, and Montenegrin texts, and used an architecture that could have enabled the model to better detect NLE token variants.

4.2 Detection of NLEs with Large Language Models

4.2.1 Background and Significance

As a continuation of the previous experiment in section 4.1, the same task was attempted to be performed with LLMs. The idea was to see if the contextual understanding of the LLM suffices, to recognize and annotate given text spans into the four subcategories correctly, as well as to investigate how fine-tuning would impact performance.

LLMs with their ability to contextually understand a prompt and synthesize an answer to the given task and/or question requires a lot of training data. Additionally, a lot of NLP components for LLMs need to work together seamlessly to generate relevant responses with regard to the initial starting prompt. The LLMs available for use, such as LLMs by the company OpenAI, have fine-tuned their models for certain general tasks. Other tasks, not fine-tuned for specific use cases, might perform poorly with only the innate ability of the language model to contextually understand and semantically disambiguate a prompt.

Hu et al. [80] analyzed the application of two GPT models (GPT-3.5 and GPT-4) for extraction of problems, treatments, tests from clinical notes. The identification of vaccine adverse events with diverse prompt engineering methods was assessed to gauge the usability and effectiveness of these LLMs for clinical NER. Depending on the type of information detected and extracted from clinical narratives, and prompt engineering, a range of F1-measures between 0.301 to 0.861 were recorded. The comparison of GPT models to clinical pre-trained language models fell closely short of the BioClinicalBERT [61] baseline, which still outperformed the LLMs, though as a very small amount

of training samples was needed to gain high performance results, this seemed accessible for future applications. Another investigation towards entity recognition was performed by Keloth et al. [81] on biomedical datasets, where similarly to all LLM-based tasks, instead of implementing a sequence labeling task, similar to previous investigation in section 4.1, a generative implementation with LLMs was pursued. Baselines consisting of PubMedBERT [82], GPT-3.5, GPT-4, and PMC-LLamMA [83] were compared to Llama-1-7b and Llama-2-7b models. The best performances were recorded for diverse datasets, for the fine-tuned BioNER-Llama-2-7b and BioNER-Llama-1-7b model, as well as for the fine-tuned approach via PubMedBERT. The results achieve high F1-measures of 0.880 for strict matches for a biomedical disease corpus, and chemical and gene-based datasets had similar high performances. The baselines show furthermore that non-fine-tuned models perform poorly for a generative entity recognition task.

4.2.2 Materials and Methods

Prompting. Annotation of datasets, generation and labeling of NLEs require a lot of human annotation effort, and while LLMs might not be able to replace annotators, these systems and models might be able to support them in such tasks. To understand LLM base performances, the models were tested in a zero-shot manner, i.e., no examples were supplied during the application of the LLMs through prompting.

Choice of LLMs. Foundation models[4] consist of two groups: (i) closed source LLMs, where only the finalized model is made available, and (ii) open source LLMs, where the generation of the model and reuse can be fully traced. Examples of closed source models are OpenAI models, which are named GPT models, e.g., GPT-3.5, GPT-4, etc., and open source models allow for the LLM to be accessed and analyzed from a source code perspective. Examples for open source models would be the different variations of Llama models, e.g., Llama-2-7b-chat. The curated large language models from different sources need to aggregate large amounts of data and fine-tune the model to reach the required size and parameters to qualify to be a LLM. For this experiment, we chose the popular and best performing type of models, which were the GPT models by the company OpenAI, according to various reports [86, 87]. Due to cost considerations in

[4]According to Wornow et al. [84] and Bommasani et al. [85], foundation models in NLP are "*machine learning models capable of performing many different tasks after being trained on large, typically unlabeled datasets*". A typical example for a foundation model is a LLM.

fine-tuning the model with the indicated amount of data, the GPT-3.5-turbo model was applied.

Fine-Tuning LLMs. In the fine-tuning step of the LLMs, the same type of prompt was applied, however, fine-tuning in this LLM context means the following: The same annotated corpus of Croatian health forum entries from section 4.1 was reused for this experiment, i.e., the dataset split into training, validation, and test sets stayed the same. A subset of the training set was processed for generating a fine-tuned model, which meant prompting with the description of the set task, the data (sentences to be processed) from the training set, as well as adding the expected correct answer with regard to the given prompt. The size of the chosen subsets of the training datasets were chosen based on practical reasons, i.e., evaluation time of annotator and cost considerations, as well as to see how many sentences would be needed to reach the same performance or improved performance in comparison to the previous experiment described in section 4.1 with BERT and ELECTRA models.

Step 1: Aggregation and Transformation of Dataset

The sentences in the BIO-labeling format, reused as gold standard annotation, needed to be transformed into a JSON (JavaScript Object Notation) formatted document. Each JSON query contained, extracted from the training set, the prompt with the included sentence (without labels), and in another field, the expected correct answer with the labels in the correct format, but without included line breaks, as this would have made the formatting incorrect and obsolete. As mentioned above, two subsets of the training set were created, one set contained 100 sentences, while the second set consisted of 1,000 sentences.

Step 2: Creation of Fine-tuned Models

For the fine-tuning step, the dashboard from OpenAI, i.e., the online graphic user interface, was used. The base model (the fine-tuned LLM) was selected from a dropdown. Specifically, GPT-3.5-turbo was selected as base model for both fine-tuning processes. Then, the training data could be uploaded or selected, if it was already uploaded to the dashboard. Validation data could also be selected, supplied and uploaded, or no validation could be chosen. If no validation set is supplied, as was the case in the fine-tuning of these two models, the model fine-tuning was adjusted based on the supplied training, i.e., fine-tuning, dataset. Naming and seed controls were available, and while naming suffixes were added, no seed controls were entered for fine-tuning. Default val-

ues for hyperparameters (batch size, learning rate multiplier, number of epochs) were used.

With the entered datasets and settings, the model fine-tuning process applied the training datasets, and calculated, based on internally generated splits of the training set as validation set, the amount of needed epochs and batch sizes. The smaller training set of 100 sentences contained 98,283 tokens, had a batch size of 1, and trained for 3 epochs. The larger dataset of 1,000 sentences, consisted of 957,021 tokens, had a batch size of 2, and trained for 3 epochs. Figure 4.1 and Figure 4.2 show the training loss calculated during the fine-tuning of the models.

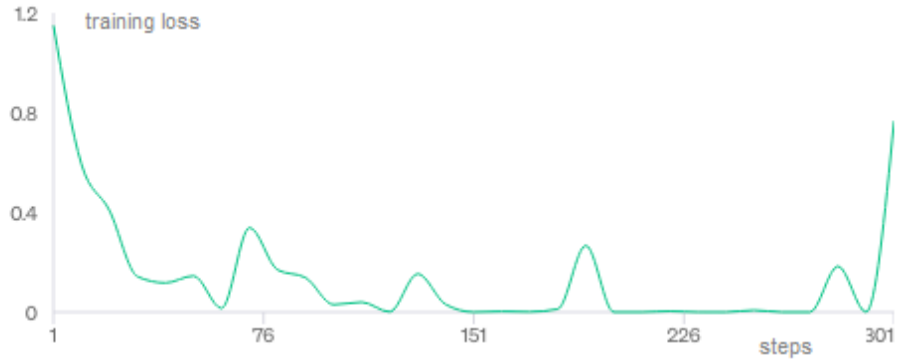

Figure 4.1: Training loss calculated during the fine-tuning of the foundation LLM GPT-3.5-turbo with 100 sentences.

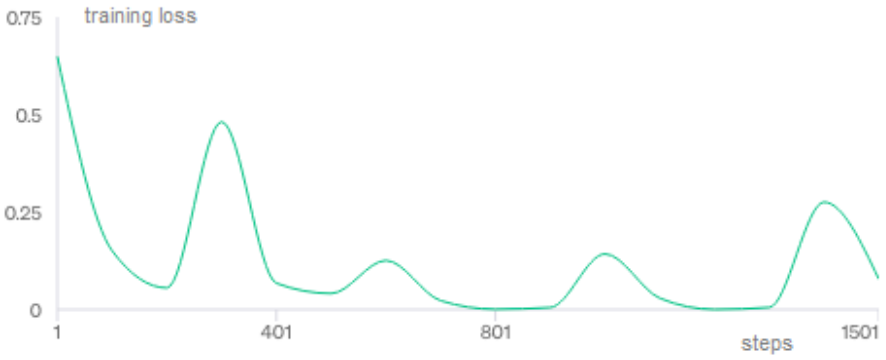

Figure 4.2: Training loss calculated during the fine-tuning of the foundation LLM GPT-3.5-turbo with 1,000 sentences.

The fine-tuned models were saved to the users' profile and made available to be used with the application programming interface (API) key for users that had the corresponding model name.

Step 3: Application of the Fine-Tuned LLMs

The prompting for testing with the fine-tuned models was done via the API. The test set of the previous experiment was applied, just with two different fine-tuned LLMs that had two different sized datasets for fine-tuning. For an additional baseline, as well as to gauge how accurate the recognition of the LLM with only the given prompt and dataset, the GPT-3.5-turbo model was used without fine-tuning. The prompt consisted of a detailed description of the text spans to be annotated, with further output instructions for the LLM. The latter consisted of limitations in explanations, formatting requirements, and data transformation limitations. This approach relied on the LLMs innate contextual understanding to fulfill the set task and accomplish the detection of NLEs with the LLM model, simply through prompting the model, i.e., working under the assumption that the model understood the concepts of abbreviations, acronyms, lexical variations, brand names, and proper names, as well as parsing the contextual information present to detect NLEs.

4.2.3 Results and Discussion

For each experiment in Table 4.3, the experiment name, if fine-tuning was performed, the count of sentences for fine-tuning from the training set, and performance metrics, precision, recall, and F1-measure were supplied. The results shown in Table 4.3 indicate that by only applying the LLM without fine-tuning, in a zero-shot scenario, an F1-measure of 0.48 underperformed quite extensively in comparison to the baseline results with the BERT and ELECTRA models, reaching 0.88 and 0.91 respectively. Through use of fine-tuning and supplying 100 sentences to the LLM, the results improved by a large margin to 0.82 in F1-measure. By using less than a third of the sentences from the training set as fine-tuning set for the last experiment with 1,000 sentences in total, the F1-measure almost doubled in comparison to the non-fine-tuning approach, and achieved a value of 0.90 in F1-measure.

In comparison to the baseline models, this performance outperformed the baseline BERT results, but underperformed just slightly for F1-measure and outperforms in precision, when analyzing the baseline ELECTRA results. The baseline ELECTRA model as a language specific model that includes only four Slavic languages still performed comparably well in comparison to a multilingual LLM with a lot more examples to learn from in training. Based on the amount of sentences required to train or fine-

Table 4.3: Performance metrics for BERT and ELECTRA models, in comparison to prompting for NLEs with foundation, as well as fine-tuned, models, on entity level. Reproduced from Kugic et al. [74] with permission from publisher Institut za hrvatski jezik (Institute for the Croatian Language).

Experiment	Fine-tuning	Count Sent.	Precision	Recall	F1-measure
Baseline BERT	yes	3,663	0.89	0.89	0.88
Baseline ELECTRA	yes	3,663	0.91	0.92	0.91
Prompting GPT-3.5	no	0	0.78	0.48	0.48
Prompting FT GPT-3.5	yes	100	0.88	0.78	0.82
Prompting FT GPT-3.5	yes	1,000	0.94	0.87	0.90

tune each model, the assumption would be that comparable results with LLMs in comparison to the traditional NER approach could be achieved by only requiring less than a third of the dataset from language models.

Chapter 5

Expansion of Lexicons with Non-Lexical Entities

A lexicon, dictionary or terminology all represent rich language resources for any language processing task. No lexicons are complete and should always be in constant revision, as languages constantly change, evolve and new words, also known as neologisms, are created, or languages adapt in social and cultural situations that can influence lexicon creation and expansion [88]. In the medical field, similar changes occur with reference to domain dictionaries and/or other medical terminologies, e.g., SNOMED CT. Additionally, clinical terms and phrases can be adapted to local language variations and common abbreviations, such as using acronyms for hospital providers, e.g., "BHB" ("Barmherzige Brüder Krankenhaus" (name of a convent hospital in Graz, Austria)).

Two investigations for the expansion of lexicons were performed. First, an embeddings-based representation of clinical diagnoses formulated as problem list entries were searched via a k-nearest neighbor search to identify synonyms, hyponyms, and hypernyms, with reference to the seed term for possible term candidates to expand the German domain dictionary on diagnoses terms [89, 90]. The aim was to first create an embedding space with millions of existent n-gram variations of disease terms in our dataset, with a language model that applied the BERT technology to align semantically similar terms. Then, the embedding space was indexed to make it searchable, and finally a nearest neighbor search was performed to find semantically similar terms. The term candidates were evaluated by a domain expert. Secondly, and as a follow up, LLMs were prompted to explore the possibility of a hierarchical directed search

for term candidates, i.e., to determine new term candidates for addition to the domain dictionary based on a class, such as synonyms, in combination with specific seed terms supplied to the LLM, via a prompt [91].

5.1 Co-occurrence Analysis and Embedding Spaces

5.1.1 Background and Significance

Vector representations of texts help identify similar contexts in other texts through various mathematical functions. Language models assist in this process by generating vector representations of texts, i.e., embeddings, and processing them to find semantically similar terms within the embedding space. The goal was to combine a co-occurrence analysis to identify statistically similar terms according to their p-values, and use those as input for the embeddings-based approach. These two mentioned approaches, co-occurrence analysis and embeddings, were often used separately, but in this investigation were combined for representative results.

Co-occurrence analyses were applied on lexicons and corpora to find statistical significant term candidates. Yalçin [92] used this method to determine the scope of terms and phrases applied in context of gambling for a certain age group. Statistical significant terms enabled a semantic analysis of the corpus. Other investigations combined the use of co-occurrence analyses with tensor learning for the identification of time-sensitive markers to determine the onset of symptoms through clinical narratives [93]. Even gene-disease associations could be identified with this approach in biomedical texts by Watford et al. [94].

Embeddings are the state-of-the-art approach to find semantically similar terms to create an embedding representation of a specific dataset, which contains contextualized and non-contextualized terms. This representation schema would then be harvested to find semantically similar, e.g., synonymous term candidates, with reference to the seed term. Fan et al. [95] applied this method to find term candidates for dietary supplements in word embeddings, trained on a corpus of clinical notes. Wang et al. [96] utilized this approach for a comparative analysis of word embeddings across four domains that comprised of news, biomedical corpora, Wikipedia, and clinical notes. Gu

et al. [97] investigated the possibility to create embeddings from online health forum entries to extract synonymous terms based on the semantic distance in the vector space.

In 2022, Kreuzthaler et al. [98] aimed to extract synonyms, hypernyms, and hyponyms via a co-occurrence analysis, and showed the usability of the method with one ICD-10 code, *I25.3*, which corresponds to "*Herz-(Wand-)Aneurysma*" (myocardial aneurysma). Combining this method in a two-step process with an embeddings approach, the results were a statistically motivated extraction of term candidates for each ICD-10 code for a 1–5-gram distribution.

5.1.2 Materials and Methods

Dataset. The source dataset consisted of de-identified diagnosis-centered problem list entries written by clinicians in the course of diagnosis documentation at discharge. In Austria, clinicians are required to enter the ICD-10 codes for the reimbursement of healthcare services, financial management and resource planning. At KAGes ("Steiermärkische Krankenanstaltengesellschaft"), an Austrian Hospital Network, this is supported by a disease encoding tool, which accepts a query and then displays ICD-10 codes with their official German descriptions. An additional free text field is editable. This option is widely used to document as much information as possible (including therapies and data), which often results in an extreme reduction in text, the abbreviation of technical terms, and the use of context-dependent descriptors. These entries are then automatically copied into the first draft of a patient's discharge summary. These characteristics present in a large collection of approx. 1.9 million unique de-identified problem list entries, and form a unique dataset for secondary uses, where each unique entry is connected to a manual coding by clinicians.

Text Pre-processing. The raw text inputs correspond to the three described columns from the dataset description. To be able to use this kind of textual data and transform it into an embedding space, some pre-processing was needed. Existent research by Chai [99] suggested that uniform pre-processing would improve and streamline results. The type of pre-processing was informed based on the applied language model, and consisted of replacing any characters not found in the regular expression `[a-zA-ZćĆčČžŽšŠđĐ]` with whitespace characters, and collapsed consecutive whitespaces afterwards, so that each token is separated by only one whitespace.

Generation of Seed Terms via Co-occurrence Analysis. With the two value pairs, the ICD code and the problem list entry, the creation of contingency tables enabled finding statistically significant n-gram representations from the problem list entries with reference to the ICD-10 code. With a p-value lower than 0.01 and 0.95 confidence intervals, n-grams qualified for seed term selection. Ranked by ascending p-values, the top ranked n-grams were selected as seed terms.

Generation of Embedding Space. To create the embedding space, the SapBERT (Self-Alignment Pretraining for Biomedical Entity Representations) language models [66, 100] were utilized. These types of language models aim to align synonymous expressions during a pre-training phase, and had shown to be effective in comparison to state-of-the-art methods in previously conducted work by the authors. For each n-gram ($0 < n < 6$), an embedding space was created with Faiss [101] for indexing, and this was performed with two different language models. A base cross-lingual language model [100] and an updated language model, which consisted of the enriched base language model via fine-tuning with synonym-pair lists connected to the ICD-10 codes. Both language models were indexed, searched via an k-nearest neighbor search, and compared with the same input seed terms. This meant that as part of this investigation a total of 10 embedding spaces were created.

Search for Term Candidates. A k-nearest neighbor search was combined with rules to find term candidates. Depending on the seed term, the casing information, n-gram composition, and ICD-10 code, the found term candidates underwent further filtering steps. Term candidate exclusion occurred based on ICD-10 codes, if the term candidate was identical to the seed terms, and rule-based casing comparisons to exclude uncertainty statements. Furthermore, the n-gram count of the seed term automatically selected the same n-gram composition from the created ten embedding spaces.

Classification of Term Candidates. A domain expert analyzed the found term candidates, and assigned classifications for *type of content* and a *hierarchical classification*. For *type of content*, the idea was to investigate any preference of the embedding space with the needed indexing to find lexical or non-lexical term candidates. The lexical assignment was gauged on the basis of manual lexical lookup operations, and how difficult the localization of a term candidate would be with reference to a medical domain dictionary. With this gauge as the focus, any term candidates with short forms, misspellings, or mixed languages found in the term candidates, with reference to the German medical domain, received the classification "*non-lexical*". For *hierarchical*

classification, seven main categories with reference to the ICD-10 code in question were assigned for each found term candidate. The seven classes were *synonym*, *synonym with extension*, *hyponym*, *hyponym with extension*, *hypernym*, *hypernym with extension*, and *incorrect*. Referencing the ICD-10 code, the extension of a term candidate refers to additional information being provided beyond the ICD-10 code, although still being part of the given ICD code. The class "incorrect" pertains to term candidates that do not match the ICD-10 code for which the term candidates were extracted for.

Evaluation. The classification of the domain expert forms the basis for the evaluation. The metric precision at k and mean average precision at k (MAP@k) were selected as evaluation metrics on class level. The reduction of hierarchical classifications to three main classes, hypernyms, hyponyms, and synonyms, allowed for more accurate information retrieval assessment, i.e., classifications with extensions were in that case then collapsed to their main class. For example, hypernym with extension was for an overall assessment joined to belong to the hypernym classification.

5.1.3 Results and Discussion

In Table 5.1, the MAP@k performance results for each term were reported. When combining all main classes (synonyms, hyponyms, hypernyms) to assess the overall information extraction performance, a drop in performance for MAP@k, k = 10 between language models could be found, cf. the base language model reached 0.865, while the updated language model performance dropped to 0.849. In contrast, the synonym extraction performance increased with the fine-tuned updated language model in comparison to the base model, while hyponym and hypernym performance dropped slightly, which was also mirrored by the semantic distances recorded for each term candidate from the embedding space. Type of content analyses were performed with statistical significance testing with a chi-square test with the hypothesis that no difference between lexical and non-lexical term candidates extracted from the embedding space would be found. The chi-square test confirmed this assumption that no distinction could be found between the seed terms and term candidate classifications for type of content, i.e., lexicality. Even though these embedding spaces were built with language models, the models might not represent the contextual information of non-lexical term candidates perfectly, as seen from the performance results. Still, NLEs would not be

at an disadvantage when being processed in this manner, and this further does not influence the extraction of new term candidates in any way.

Table 5.1: Average precision at k calculated per ICD code for each language model per class (Syn, Hypo, Hyper) and across all classes (All). Additionally, mean average precision at k (MAP) is calculated across all ICD codes. Reproduced from Kugic et al. [90] with permission from publisher Elsevier.

Average Precision @ k								
	Base LM				Updated LM			
ICD	All	Syn	Hypo	Hyper	All	Syn	Hypo	Hyper
I25.1	0.565	0.021	0.544	0.000	0.511	0.000	0.511	0.000
I25.3	1.000	0.000	1.000	0.000	1.000	0.293	0.707	0.000
I64	1.000	0.851	0.149	0.000	1.000	1.000	0.000	0.000
J44.1	0.683	0.092	0.000	0.591	0.836	0.509	0.000	0.327
J22	1.000	0.000	0.000	1.000	1.000	0.065	0.000	0.935
C34.0	1.000	0.439	0.561	0.000	1.000	0.531	0.469	0.000
F00.1	1.000	0.262	0.000	0.738	1.000	0.131	0.000	0.869
K59.0	1.000	0.044	0.956	0.000	1.000	0.086	0.914	0.000
E14.5	0.857	0.000	0.847	0.010	1.000	0.000	1.000	0.000
N19	0.545	0.545	0.000	0.000	0.291	0.291	0.000	0.000
C18.7	0.869	0.492	0.377	0.000	0.697	0.469	0.229	0.000
MAP	0.865	0.250	0.403	0.213	0.849	0.307	0.348	0.194

5.2 Co-occurrence Analysis and Large Language Models

5.2.1 Background and Significance

A possible solution to keep clinical lexicons up-to-date would be, for instance, to find statistically significant term candidates via a co-occurrence analysis, and extract synonyms, hypernyms, and hyponyms, with the application of LLMs. As token n-grams in medical and clinical lexicons can be particularly difficult to understand and contextualize, to find semantically similar terms, the application of a foundation model could be useful in this regard. In this follow-up investigation, a directed extraction of synonyms, hypernyms, and hyponyms was implemented. The primary way to find semantically

similar terms would be embeddings, as explained in section 5.1. A directed extraction would have also been possible with embeddings; however, while German disease terminology resources and catalogues exist, such as ICD-10, NLEs are predominantly not represented as part of those resources. Specifically, a large portion of annotation hours would have been needed to create a corpus of clinical terms annotated into various classes, and then fine-tune the embedding spaces with this information to then in turn search for semantically similar terms. Some related works without LLMs to perform directed term extractions are the following ones: Koleck et al. [102] applied a tool called NimbleMiner [103], a word2vec [104] implementation. This method was used to develop a specialized vocabulary for symptom mentions from the supplied EHR notes with the application of SNOMED CT and the UMLS Metathesaurus "synonyms" category. The symptom detection performance ranged from 0.80 to 0.96 for five different SNOMED CT symptom concepts, and detected also term candidates with misspellings, abbreviations, unique token n-gram combinations. Zhang et al. [105] combined structure embeddings from a graph convolutional network and semantic embeddings from a pre-trained language model to extract Chinese synonymous and hyponymous terms. This combined model achieved 0.84 for Hits@5, a metric that describes the percentage of the best ranked text spans according to the terminology base for term candidate retrieval [106, 107]. A similar application of graph convolutional networks and recurrent neural networks to augment existent ontologies were performed by Nath et al. [108], and a comparison with vectors from pre-trained language models.

5.2.2 Materials and Methods

Dataset, Pre-processing, and Co-occurrence Analysis. The same dataset was used as in section 5.1.2. The pre-processing followed the same guidelines, and the generation of seed terms with the co-occurrence analysis was performed as described. The count of seed terms per ICD code was set to five, instead of ten, to reduce annotation hours.

Prompt Engineering. To extract term candidates with LLMs, a prompt was formulated, so that five semantically similar terms were generated by the LLM. With the OpenAI API, the GPT-4 model was used for term candidate generation. The system requirements and prompts were formulated in German. The system requirements explained that the model should act as a expansion tool for medical terminologies to find

relevant terms, cf. in German: "*Dieses System soll für die Erweiterung von medizinischer Terminologien relevante Terme ausfindig machen.*" The English prompt translation corresponds to: "*This system should find relevant terms for the expansion of medical terminologies.*" The prompt design consisted of instructions for term generations with output requirements performed in a two step process. First, a generation of five possible clinical term candidates with reference to a class, such as hyponyms, synonyms, and hypernyms, was queried for. Output guidelines made sure to mitigate explanations, and identical term candidates in comparison to the seed term. The latter case often was found during initial test runs of the method. For example, one of the prompts used in German to generate synonymous terms for the multi-term expression acute exacerbation of a chronic obstructive lung disease: "*Generiere 5 mögliche synoynme Ausdrücke des folgenden klinischen Ausdrucks: Akute Exazerbation einer COLD. Die synonymen Ausdrücke sollen ohne weitere Erklärungen aufgelistet werden und mit einem Semikolon getrennt werden. Es soll keine numerische Aufzählung der Terme ausgegeben werden. Der angegebene klinische Ausdruck soll nicht als Output ident ausgegeben werden.*" The English prompt translation corresponds to: "*Generate 5 possible synonymous expressions for the following clinical term: Acute exacerbation of COLD. The synonymous expressions should be listed without further explanation and separated by a semicolon. No numerical enumeration of the terms should be output. The specified clinical seed term should not be output as term candidate.*" In a second step, the terms were queried for relevance with reference to the seed term, and prompted to be ordered according to their relevance and probability of correctness, i.e., most likely and probable term candidates, with lesser likely candidates below. Previous work by Lin et al. [109] showed that GPT models can output confidence percentages, i.e., uncertainty, with reference to the generated output by the model, which are in a way probability estimations by the model for the self generated answers. To accomplish this, the term candidates were supplied to the LLM based on the first prompt, and detailed instructions for output of the data were given to make sure that an automated processing of terms could take place, e.g., specific formatting and output of terms with the ranked numbers 1 through 5 given before each term. The same procedure was followed for each term candidate and class. The system guidelines stayed the same as in the first step, and an example prompt for the second step in German for the extraction of synonymous term candidates of the same term candidate as above would be the following: "*Folgende synonyme Ausdrücke wurden für den klinischen Ausdruck "Akute Exazerbation einer COLD" gefunden, welche mit Strichpunkt getrennt angeführt werden: Akute Verschlechterung*

einer COPD; Akuter Schub einer chronisch obstruktiven Lungenerkrankung; Akutphase einer chronisch obstruktiven Atemwegserkrankung; Akut aufgetretene Verschlimmerung einer chronischen Bronchitis; Akuter Anfall einer obstruktiven Atemwegserkrankung. Erstelle ein Ranking nach absteigender Wahrscheinlichkeit, welches gefundenen Ausdrücke, die am wahrscheinlichsten ein Synonym zum klinischen Ausdruck "Akute Exazerbation einer COLD" darstellen, eine sehr hohe Platzierung (bspw. Platz 1) bekommen. Alle anderen Terme sollen mit abnehmender Wahrscheinlichkeit folgen, bis zur 5. Platzierung. Keine Veränderung von Termen oder neues Hinzufügen von Termen ist gestattet. Keine Erklärungen oder weitere Beschreibungen sollen außer dem Ranking ausgegeben werden. Die Ausgabe soll pro Zeile aus der numerischen Platzierung, und dem klinischen Ausdruck bestehen in diesem Format: Platzierung. Ausdruck" The English prompt translation corresponds to: "*The following synonymous terms were found for the clinical term "acute exacerbation of COLD", which are listed separately with a semicolon: Acute exacerbation of COPD; Acute episode of chronic obstructive pulmonary disease; Acute phase of chronic obstructive airway disease; Acute exacerbation of chronic bronchitis; Acute attack of obstructive airway disease. Create a ranking according to descending probability, which found terms are most likely a synonym for the clinical term "acute exacerbation of COLD", and should be given a very high ranking (e.g. 1st place). All other terms should follow with decreasing probability, up to the 5th place. No changes to terms or new additions of terms are permitted. No explanations or further descriptions should be output apart from the ranking. The output per line should consist of the numerical ranking and the clinical expression in this format: Ranking. Expression*"

Evaluation. Annotation of the terms was performed by a domain expert. Each term candidate was analyzed with reference to the seed term, to determine the correct extraction of term candidates per class, i.e., synonyms, hypernyms, hyponyms. Based on the additional ranking information extrapolated from the LLM output, two metrics were used for the quantification of the results: accuracy and mean average precision (MAP)@k, for k= 1 and 5. For the accuracy performance, a 95% confidence interval was calculated and reported in conjunction with the average accuracy.

5.2.3 Results and Discussion

The performance results for the LLM term candidate generation task for each class under investigation can be found in Table 5.2. The performance values indicated that the LLM GPT-4 could extract term candidates with reference to a clinical seed term, where particularly hypernyms performed the best with an accuracy of 0.776, and a MAP@5 of 0.801. With reference to the confidence intervals, statistical significance was found in term class comparisons, which means hypernyms performed the best, followed by synonyms with 0.604 in accuracy, and then hyponyms with 0.428. The LLM-based re-ranking performed similarly well, as can be seen from the MAP@k reported results that also follow the indicated performance trend for each class.

Table 5.2: Performance metrics for each class (synonym, hypernym, hyponym) under investigation with accuracy, reported with a 95% confidence interval, and mean average precision MAP@k, for k = 1 and 5. Reproduced from Kugic et al. [91] with permission from publisher IOS Press.

Class	Accuracy [95% CI]	MAP@1	MAP@5
Synonym	0.604 [0.540 - 0.665]	0.680	0.723
Hypernym	0.776 [0.719 - 0.826]	0.780	0.801
Hyponym	0.428 [0.366 - 0.492]	0.500	0.507

Chapter 6

Disambiguation of Non-Lexical Entities

Acronyms have a particular characteristic that more or less are shared across all NLEs, in which their ambiguity is interlocked with the context this subtype of NLEs appear in. Additionally, the way acronyms are formed in clinical narratives, the letters the acronyms are comprised of, appear in the same sequence with letters appearing between them, when comparing the short form with the long form. Another characteristic of acronyms, or abbreviations, is their length, and the interchangeability of long forms and short forms, i.e., the shorter the acronyms, the harder the disambiguation of acronyms. All those characteristics are partly the reason why the following two investigations were focused on acronyms, in conjunction with short form sense inventories that are either curated for the German and Portuguese language or reused for the English language, all based on clinical narrative datasets.

This section focused on two investigations for acronym disambiguation. First [110], through text mining web results from online searches in a data-driven, rule-based, and heuristic approach, long forms for short forms were found, and compared to the use of the LLM GPT-3.5. Second [111], as a follow-up to the first investigation, the LLM method was further analyzed with other LLMs. Variables, such as language and context length, were varied to understand the capabilities of LLMs more clearly.

6.1 Text Mining for Acronym Disambiguation

6.1.1 Background and Significance

In 2020, a systematic review on social media datasets [112] showed how symptom information could be extracted and processed with text mining approaches. It explained the use of text mining, i.e., rule and pattern-based statical analyses of text, and natural language processing, i.e., use of content, context and phrasing patterns, to find matching results. Furthermore, the review grouped clinical content categories analyzed in the articles, and their symptom concepts that had been investigated in the sources. These non-standard variations of clinical terms, such as "brain fog" instead of "cognitive dysfunction", were partly reasons, why manual extraction of symptom information would be preferable over automated processes, as such NLEs were not mapped or found in clinical concept lexicons or dictionaries. Menaha and Jayanthi [113] conducted a survey of text mining methods to find acronym-expansion pairs across text and web documents to disambiguate mentions. The summarized methods applied for this NLP task were heuristic approaches, ML methods, such as SVM, Hidden Markov models, and Conditional Random Fields (CRF). Link et al. [114] explored a semi-supervised method named CASEml, which incorporated context information for binary acronym disambiguation in clinical narratives. An accuracy range of 0.70 to 0.95 was achieved for three two-character acronyms. Further related works for short form processing are summarized in the performed scoping systematic review on short forms [48].

6.1.2 Materials and Methods

Data. From the specialties, dermatology, cardiology, and oncology, a random selection of de-identified German clinical narratives was made. Based on those clinical narratives, a small collection of acronyms with their syntactic context were extracted to form the basis of this investigation. With a rule-based method, which found acronyms of two to seven letters in length, with at least two letters in uppercase, without the exclusion of digits as part of the acronym, 143 text spans were extracted. 43 text spans and were used for training and establishing the methodology, while 100 text spans were applied for testing. Additional metadata information in the form of specialty information was saved for each text span.

Web Mining. With the aim to disambiguate acronyms through web mining, a corpus of texts from the web needed to be curated. Through an API to the search engine BING, a query with the acronym and the context was performed. The results pages, as generally expected for search engines, consisted of headlines followed by short texts that summarize the contents of each linked website. All these texts were collected based on the search API parameters, with German as the target language and a limitation to a maximum of 50 hits per query. To enhance and enlarge the text corpus being created, as well as avoid duplicate entries, an offset of 50 was introduced, for additional search results. The obtained web text corpus was created for each search, more accurately for each acronym, with each result page then being downloaded and cleaned. The cleaning and normalization of the text corpus aimed to establish a normalized text base, from which possible acronym resolution candidates could be mined. First, the whole text corpus was tokenized, and tokens containing non-Latin letters were flagged as not viable options for term candidate selection. Second, the n-gram frequency was computed. The n-grams ordered by decreasing frequency, allowed for the processing of the most often appearing n-grams in the dataset, and any n-gram that appeared more than twice, was flagged as a viable candidate for a long form, and an acronym candidate score was calculated for each long form. This automatically curated term candidate lexicon for each acronym in the training set was then evaluated manually by a domain expert. Depending on certain characteristics of long forms, a plausibility score would be automatically calculated and assigned to each long form, based on varying factors that influence the correctness of terms. The maximum plausibility score for a long form was 1. Those factors were implemented as features for each term candidate. The importance of each feature was determined by the domain expert, and the features consisted of compression, term candidate length imbalance comparison to the assigned short form, Levenshtein edit distance, casing, stop word occurrence, and placement of neighboring tokens. The main driving factor here was to increase precision and accuracy through the introduced features and the interconnected plausibility score. These features were implemented, detected, and calculated with rules, regular expressions, and supported functions from Python libraries.

Conversational Agent. The disambiguation of acronyms with a conversational agent, the LLM "GPT-3.5-turbo", was accomplished by prompting for the resolution with the acronym and its context. The API from OpenAI was applied, and the query sent to the LLM was posed as a question, followed by general system information and output requirements. System information comprised of additional information

for the model, i.e., the main objective of this model should be to disambiguate clinical acronyms, and act as an acronym disambiguation tool. Output requirements instructed the model to output the long form version of the acronym in a JSON format. Both short form and long form versions of the queried acronym were required to be in the answer, so that a systematic extraction of results was possible. For example, one of the clinical narrative text spans corresponded to "*negativ HNAP frei pupillen rund mittelweit*" (negative HNAP free pupils round medium-wide). With the prompt for GPT-3 in German, the aim was to disambiguate the acronym HNAP, i.e., "*Was bedeutet "HNAP" im klinischen Kontext "negativ HNAP frei pupillen rund mittelweit"? Die Expansion des gesuchten Akronyms soll als JSON Format zurückgeliefert werden, und soll nur aus dem Akronym und der Expansion bestehen. Eine Erklärung des klinischen Textes ist nicht notwendig.*" The English prompt translation corresponds to: "*What does "HNAP" mean in the clinical context "negative HNAP free pupil round medium-wide"? The expansion of the searched acronym should be returned as JSON format and should only consist of the acronym and the expansion. An explanation of the clinical text is not needed.*"

Pre-processing. During the establishment of the text mining approach, false information related to the search often appeared due to the context around the acronym in the query. The search consisted of the context and additional metadata information concatenated with whitespace characters, during the web mining of the acronym. As part of the search for web content, the context often consisted of digits, lab results or additional information, in which the context was, as expected, filled with jargon-based expressions, lists of numeric or tabular information as part of the narratives, and other clinical concepts as short forms. Especially digits and symbols hindered the accuracy for the type of information that should be found with the given search for web mining. Based on that observation, the digits and symbols were replaced by whitespace characters, followed by normalizing the distances between the tokens. In an effort to make both approaches, i.e., web mining and conversational agent approach, comparable, the same pre-processing of the context was also applied for the LLM API query.

Term Candidates. The validity of each term candidate was determined by the domain expert. To allow some flexibility, non-standard spelling variations of clinical terms or long forms were permitted, such as "oe" instead of "ö", or "c" instead of "k", etc. Non-German words as part of the term candidate were allowed as well, such as Latin or English. The translations of the acronym long form in other languages, how-

ever, needed to follow the acronym sequence to be flagged as viable term candidate. Letters similar to acronyms, such as roman numerals as part of n-gram resolutions, were never expanded.

Evaluation. For evaluation, both approaches selected either the highest ranked long form candidate for a given acronym or the generated long-form version from the LLM, i.e., only one long form per method and acronym were evaluated. The domain expert analyzed the test set, and annotated all resolutions with reference to their context with "correct" or "incorrect". Precision, recall, and F1-measure were used as metrics for the evaluation of the results. Statistical significance was calculated with the Chi-square hypothesis tests to assess, if a difference between the methods could be found.

6.1.3 Results and Discussion

Text mining with the search engine BING reached 0.488 in F1-measure. The introduction of a threshold with reference to the calculated plausibility score allowed an increase in precision to 0.75, while the F1-measure worsened as a result to 0.179. The conversational agent, the LLM, reached 0.679 in F1-measure, and outperformed all other methods in this comparison with statistical significance.

Table 6.1: Performance metrics for acronym-expansion via BING and ChatGPT. Reproduced from Kugic et al. [110] with permission from publisher IOS Press.

Experiment	Precision	Recall	F1-score
BING: no threshold	0.535	0.449	0.488
BING: threshold >0.1	0.530	0.440	0.481
BING: threshold >0.5	0.750	0.101	0.179
ChatGPT – GPT-3.5-turbo	0.740	0.627	0.679

Both methods showcased only acceptable results. The main issue with the resolution of clinical acronyms seemed to be related to the non-domain relevant results found during the building of the corpora. Most resolutions ended up not medically relevant, and adjusting for these instances with search filters or relevant search terms, such as "*clinical*" or "*medical*", did not statistically influence the results. Some of the same issues were also found for the LLM-based approach. Referring back to the example prompt with the acronym HNAP from the methods section, the model output consisted of four possibilities for the resolution of the acronym, and not one of them was

correct, i.e., "*Hepatitis A/B/C Negativ, HIV Anitkörper Negativ, Antiphospholipid-Antikörper Negativ, Paraneoplastische Syndrome Negativ*" (Hepatitis A/B/C negative, HIV antibody negative, antiphospholipid antibody negative, paraneoplastic syndromes negative). The correct resolution would have been "*Hirnnervenaustrittspunkte*" (cranial nerve exit points). From this small scale experiment, the assumption would be that the LLM is the better alternative for acronym disambiguation in comparison to a web mining approach.

6.2 Large Language Models for Acronym Disambiguation

As a continuation of the previous work on web mining and the comparative analysis to LLMs [110], acronym disambiguation was explored not only with a larger dataset, but in different languages, with a variety of LLMs, and prompt adjustments [111].

6.2.1 Background and Significance

From the introduction of LLMs to consumers in 2021, cf. ChatGPT, research explored various ways to apply this methodology to medical research questions, and to see its impact and usability for its application in medicine. Since then, other LLMs have been made available, and the researchers' appointed task would be to test their veracity and applicability for specific research questions. A view on a particular LLM interface, ChatGPT, by Liu et al. [115] summarized potential applications of LLMs in the medical field, such as clinical decision support, question-answering for medical queries, writing and analyses of medical documentation to support clinicians in their day-to-day tasks. Future research directions by the authors indicated real-time monitoring, personalized medical treatments, as well as remote health care or home monitoring, and possible integrations to support interoperability between health systems.

At the time of the investigation, LLMs had not been explored at length for acronym disambiguation tasks. From the related works, only one investigation had applied LLMs for acronym disambiguation for clinical narratives from the English CASI dataset, described in more detail in section 2.1.3. In particular, the investigation by Agrawal et al. [47] denoted as clinical sense disambiguation applied three datasets in English

with prompting for the resolutions with InstructGPT [116] on the basis of the GPT-3 model. The combination of a zero-shot approach (no examples given as part of the prompt design), with embeddings made it possible to expand the given acronyms in clinical narratives with an accuracy of 0.86. Additionally, Liu et al. [117] explored LLMs for the disambiguation of acronyms. Even though the prompt for the resolution of the acronym was performed in a zero-shot, one-shot and three-shot shot manner, the prompt did include a list of possible answers, which would be best comparable to a multiple choice question and answer task to choose the correct sense. This method was tested on 1-5 notes for each of the 41 abbreviations with the LLMs, GPT-3.5 and GPT-4. Open source LLMs, such as Mixtral and a BERT-based model, BioBERT [31], were applied for the same tasks. In a zero-shot approach, the LLM GPT-4 outperformed the other LLMs, and achieved a macro F1-measure of 0.95 for 1 note, and 0.74 for 5 notes. Further related works for short form processing are summarized in the performed systematic scoping review on short forms [48].

6.2.2 Materials and Methods

Dataset. Three languages were used for the selection of datasets: English, German, and Portuguese. The choice of languages and datasets were selected on the basis of the domain experts' fluency in those languages. Two extracts of the CASI dataset were used for English, and three datasets were curated for Portuguese and German from manually de-identified clinical narratives. In particular, SemClinBr [51] for the Portuguese dataset, and German clinical narratives from cardiology, oncology and dermatology departments at KAGes, a hospital network provider in Austria were used. Patient identifiers seldom occurred within the clinical narrative text spans. In the manual review for de-identification, personal patient identifiers were replaced by tags, as the same de-identification process was implemented with the CASI dataset. The naming conventions for the datasets inform about the language and dataset size, i.e., "17k" and "500", each refer to distinct text span counts, and "3A" refers to a dataset, which only consists of three distinct acronyms. The curation process consisted of a rule-based extraction of acronyms and their context from clinical narratives, where the mean text span length was set to 100, while the text span length for the CASI datasets was not shortened and processed in full. For additional context information, metadata were extracted for each text span, and depending on the dataset, the type of metadata information changed due to inaccessibility of the same context information per dataset

and text span. The metadata information comprised of: section header information for the CASI datasets, specialty allocation, such as cardiology, for the German datasets, and annotated signs, symptoms, and disorders from the Portuguese dataset.

Pre-processing. It was assumed that a decreasing performance of the LLM could be attributed, in the previous investigation [110], to the extensive pre-processing that took place to make the previous methods comparable to text mining. To mitigate any dampening effects from our experimental setup, no pre-processing of the clinical texts took place in these experiments.

Prompt Engineering. The language of the prompts was selected based on the source dataset languages, i.e., the instructions given to solve the disambiguation task were formulated for German clinical narratives, in German, and the same procedure was followed for Portuguese and English. The prompt itself consisted of the task that needed to be solved, e.g., resolution of the specific acronym according to the given context, and output requirements, e.g., no further explanations, output formatting requirements, e.g., "short form, long form", etc. The specific wording of the prompts can be viewed in the published manuscript [111]. Four LLMs were tested for acronym disambiguation: GPT-3.5, GPT-4, Llama-2-7b-chat, Llama-2-70b-chat.

Evaluation. A domain expert evaluated all resolutions with reference to the context, and annotated each resolution with "correct/incorrect". These manual evaluations by the domain expert were especially needed as no gold standard for the datasets was available, apart from the CASI dataset. The metric accuracy and a 95% confidence interval were calculated to compare the various datasets and prompt combinations, i.e., inclusion or exclusion of metadata, and the application of different language models, e.g., Llama-2-7b-chat. These evaluation measures were indicators for statistical significance, i.e., if an overlap of confidence intervals between methods was recorded, then the comparison between methods would not be statistically significant. An additional automated evaluation of more than 17,000 acronym resolutions was performed with the GPT-3.5 model with a separate matching algorithm, consisting of regular expression, string similarity, and mapping tables to compare resolutions from the LLM with the long form from the sense inventory. To explain, a larger subset of more than 17,000 text spans and their short forms from the CASI dataset were prompted with the same method to establish a more nuanced baseline performance comparison that was not limited by the three-acronym dataset selection. The automated evaluation post-processed the resolutions with rule-based normalization techniques, and compared resolutions to

manually created mapping tables, to see if the gold standard resolution recorded as part of the CASI dataset matched the LLM resolution. To ascertain the correctness of the automated matching, 500 annotations were additionally manually checked, and any inconsistencies computed as error rate.

6.2.3 Results and Discussion

The performance metrics for the disambiguation of acronyms can be seen in Table 6.2 and Table 6.3. For the *English 17k* dataset, the GPT-3.5 model obtained an accuracy of 0.91, with a calculated error rate of 0.014. For the other datasets overall, the *English 3A* dataset achieved an accuracy of 0.98 with the GPT-4 model, which was better in comparison to the GPT-3.5 or the Llama models, with statistical significance based on the confidence intervals.

Table 6.2: Overall accuracy scores for prompting GPT-3.5 and GPT-4 models per dataset (with 0.95 confidence intervals) with two different prompt combinations (PCs), i.e. prompt, context and metadata (MD) variations. Reproduced from Kugic et al. [111] with permission from publisher Oxford University Press.

Datasets	PC (i) GPT-3.5 + context	PC (ii) GPT-3.5 + context + MD	PC (i) GPT-4 + context	PC (ii) GPT-4 + context + MD
English 3A	0.85 [0.82, 0.89]	0.88 [0.84, 0.91]	0.97 [0.95, 0.98]	0.98 [0.97, 0.99]
German 3A	0.41 [0.36, 0.45]	0.37 [0.33, 0.42]	0.65 [0.61, 0.69]	0.59 [0.54, 0.63]
German 100	0.74 [0.64, 0.82]	0.72 [0.62, 0.81]	0.86 [0.78, 0.92]	0.85 [0.76, 0.91]
Portuguese 500	0.74 [0.70, 0.78]	0.76 [0.72, 0.80]	0.88 [0.85, 0.91]	0.89 [0.86, 0.91]

Table 6.3: Overall accuracy scores for prompting Llama-2-7b-chat and Llama-2-70b-chat models per dataset (with 0.95 confidence intervals) with two different prompt combinations (PCs), i.e., prompt, context and metadata (MD) variations. Reproduced from Kugic et al. [111] with permission from publisher Oxford University Press.

Datasets	PC (i) Llama-2-7b-chat + context	PC (ii) Llama-2-7b-chat + context + MD	PC (i) Llama-2-70b-chat + context	PC (ii) Llama-2-70b-chat + context + MD
English 3A	0.73 [0.69, 0.77]	0.72 [0.68, 0.76]	0.69 [0.65, 0.73]	0.70 [0.66, 0.74]
German 3A	0.02 [0.01, 0.03]	0.04 [0.03, 0.06]	0.08 [0.06, 0.11]	0.10 [0.08, 0.13]
German 100	0.34 [0.25, 0.44]	0.34 [0.25, 0.44]	0.41 [0.31, 0.51]	0.45 [0.35, 0.55]
Portuguese 500	0.16 [0.13, 0.19]	0.15 [0.12, 0.18]	0.28 [0.24, 0.32]	0.29 [0.25, 0.33]

Llama models only reached suitable results for *English 3A* with an accuracy of 0.73, while German and Portuguese datasets would not be able to be used in this constellation in clinical practice, even on premise. Especially in the case of Llama models for German and Portuguese, the prompt and output requirements were not followed, and responses were filled with hallucinations and incorrect predictions, such as generating non-existent words for the resolution of acronyms. Examples included "*Microsoft Contin*", "*Automatische Puff-Leitung*", "*Herz-Tasternalis*", or "*Tibutation*". The addition of metadata as part of the prompts did not produce statistically significant results based on the confidence intervals, i.e., it had no effect.

Chapter 7

Lifestyle-related Risk Factors and Non-Lexical Entities

Both smoking behavior and alcohol consumption belong to two topics, which are lifestyle-related risk factors for patients health, and these also belong to the more general topic of Social Determinants of Health (SDOH). According to the World Health Organization, SDOH are "*non-medical factors that influence health outcomes [...] the conditions in which people are born, grow, work, live, and age, and the wider set of forces and systems shaping the conditions of daily life*" [118, 119]. Lybarger et al. [120] analyzed the applicability of NLP to extract SDOH from clinical narratives to augment EHR datasets with structured variables from diversely documented text representations. Their method was based on a multi-label BERT transformer model that identified relevant text spans, and labeled these according to five event types, i.e., alcohol, drug, tobacco, employment, and living situation. The SDOH method achieved an F1-measure of 0.86, and was able to augment EHRs with missing structured SDOH variables. In 2023, Lybarger et al. [121] reported on the outcomes of the 2022 shared task on extracting SDOH from clinical narratives organized by the George Mason University and the University of Washington. The organizers supplied training and test sets for participating teams to compare information extraction, generalizability, and learning transfer for SDOH events, e.g., alcohol, drug, tobacco, employment, etc. Pre-trained language models were utilized by the best performing teams, whereas a sequence-to-sequence approach reached first place across all sub-tasks. Specifically for information extraction, 0.90 in F1-measure constituted as the best performance result across all teams.

A systematic review by Patra et al. [122] summarized the application of NLP for the extraction of SDOH, in which both alcohol and smoking status were well represented. The review showed that ML methods can be successfully applied for SDOH information retrieval. Furthermore, clinical documentation of SDOH often relied on rule-based or semi-automated systems for lexicon creation because of varied documentation practices or noisy datasets [122], which might be an indicator for the presence of NLEs. Particularly, processing clinical narratives without specialized handling of NLEs (for identification, expansion, and disambiguation) was of interest, to investigate ML models ability to contextually learn and understand documentation practices in a data-driven way. For example, the term "C4 Abusus" (alcohol abuse) in German clinical narratives can hinder interpretation and extraction of information, if the referenced text span is not found in lexicons.

In two investigations as part of this dissertation, text classification of smoking and alcohol status [123, 124] was performed. Various text classification methods to train ML models were implemented. Multi-class classification schemata were created in both instances through a bottom-up analysis of clinical narratives to establish a terminology schema for each risk factor independently with reference to SNOMED CT.

7.1 Alcohol Status Classification

7.1.1 Background and Significance

Alcohol consumption is problematic due to its impact on health. Immediate and long-term side effects cause many diseases and deteriorate the prognosis of existing illnesses, which is why alcohol consumption is both seen as a primary and secondary risk factor. The differentiation between types of alcohol consumption status is needed for a nuanced interpretation of patients as SDOH play a role in treatment [118].

The following investigations should give an overview of alcohol status classification: Alzoubi et al. [125] implemented a system for text classification with a bag-of-words and keyword search to identify sentences relevant for alcohol consumption, while accounting for negation and temporal contexts. The implementation was modeled as a multi-class classification problem on document level, i.e., the classes "Non-drinker", "Past drinker", "Current drinker", and "Unknown", were used as classifications by the annotators for 5,000 MIMIC-III discharge summaries [126]. Based on a 5-fold cross-validation, SVM and logistic regression were the best classifiers for alcohol consumption and reached F1-measures between 0.88 to 0.99 between the four classes. Similarly, Lix et al. [127] applied SVM for text classification into three categories: "Current drinker", "Current non-drinker", and "Unknown". With a 10-fold cross-validation, the model created from a gold standard of 2,000 EHRs reached an F1-measure of 0.89 for current drinkers of alcohol and 0.98 for the unknown category. Topaz et al. [128] used the MIMIC-II dataset [129] and the NimbleMiner [103] system to classify alcohol and substance abuse on a curated gold standard dataset of 1,610 discharge summaries. This system achieved an average F1-measure of 0.84 for both categories, and outperformed the Convolutional Neural Network (CNN) [130, 131] and cTAKES [22] baselines with 0.81 and 0.83, respectively.

7.1.2 Materials and Methods

Data. Using RegEx (regular expressions[1]), 100 characters to the left and right of the signal word were extracted from clinical narratives from the three clinical specialties cardiology, dermatology, and oncology. In total, a 200-character text span length was

[1] full regular expression published in the corresponding publication [123]

extracted with the signal word occurring in the middle, i.e., a total of 1,429 text spans constituted the whole dataset. To de-identify clinical narratives, identifiers were replaced with realistic substitutes. To create a gold standard, a bottom-up annotation schema was created by the annotators to assign one of the following six classes to each span: "Current drinker of alcohol", "Current non-drinker of alcohol", "Ex-problem drinker", "Disorder caused by alcohol", "Problem drinker", "Alcohol consumption unknown". The classification were derived from SNOMED CT, and the counts per class can be found in Table 7.1. The 200-character text spans were classified into one of the six classes without additional context information. The alcohol consumption documentation varied in the dataset, e.g., the signal expression in text spans consisted of consumed drinks per day, alcohol abstinence documented with temporal values, or even disorders related to alcohol consumption. Examples for the class "Disorder caused by alcohol" were expressions related to "Alkoholdemenz" (alcohol dementia), "Alkoholdelir" (alcohol delirium), "Alkoholabusus" (alcohol abuse), and "Leberzirrhose" (liver cirrhosis). One annotator annotated the whole dataset, while the second annotator reannotated 20% of the dataset to ascertain data quality. The annotators demonstrated strong agreement with an inter-rater agreement (Cohen's kappa κ) of 0.9.

Table 7.1: SNOMED CT value sets and class distributions. Reproduced from Kugic et al. [123] with permission from publisher IEEE.

Class	SCTID	Fully Specified Name	Counts
0	219006	Current drinker of alcohol	197
1	105542008	Current non-drinker of alcohol	259
2	286857004	Ex-problem drinker	100
3	719848005	Disorder caused by alcohol	249
4	228281002	Problem drinker	374
5	160580001	Alcohol consumption unknown	250

Pre-processing. Due to the application of an uncased language model, the input text was normalized. As part of this normalization, any uppercase letters were converted to lowercase, symbols were converted to whitespace characters, consecutive whitespace characters were collapsed, with the aim to only have alphanumeric characters, including diacritics, in the text span. For the text span length variation aspect of this study to investigate how much length would be needed to correctly classify the alcohol consumption status, the length of the span was always shortened from both sides, and then

padded with whitespace characters. Each variation of text span length for fine-tuning the language model had the same character length.

Model Fine-Tuning. The uncased German BERT language model[2] was applied. The dataset was split into 80 – 10 – 10 (training – validation – test set) partitions. Early stopping was implemented to stop the fine-tuning process, as soon as no further information could be learned from the given dataset. In this instance, this was defined as conducting mid-training evaluations, and if three evaluations in succession did not decrease the evaluation loss by at least 0.01, training was concluded and the trained model was saved for application and testing. For a baseline comparison, the German fastText [132] model [3] was applied.

7.1.3 Results and Discussion

The baseline fastText model reached a macro F1-measure of 0.82. The BERT model performed marginally better in comparison to the baseline, and achieved a macro F1-measure of 0.85 for the maximum context length of 100, while the context length of 60 had a performance of 0.83. The performance metrics per class for the best performing model can be found in Table 7.2.

Table 7.2: Performance Metrics per Class for the BERT model at context length of 100. ©2022 IEEE. Reproduced from Kugic et al. [123] with permission from publisher IEEE.

Class	Precision	Recall	F1-measure
0	0.93	0.68	0.79
1	0.83	0.94	0.88
2	0.92	0.92	0.92
3	1.00	0.78	0.88
4	0.81	0.84	0.82
5	0.78	0.86	0.82
macro avg	0.88	0.84	0.85

Pre-processing and contextualized expressions about alcohol consumption status could have impacted the performance. Symbols and uppercase letters in certain instances

[2]https://huggingface.co/dbmdz/bert-base-german-uncased
[3]https://fasttext.cc/docs/en/crawl-vectors.html

are indicators in clinical narratives, e.g., symbols, such as a plus or minus symbol, can symbolize the existence or absence of substances or the negation of a given state, which would not have been contextually present in the text spans processed by the language models. The limited sample size of 1,143 sentences could have impacted the results. The application of a domain-specific language model might additionally offer a performance boost.

7.2 Smoking Status Classification

7.2.1 Background and Significance

Tobacco smoking is the act of inhaling and exhaling the smoke of burning tobacco, which contains numerous harmful chemicals, including nicotine, tar, and carbon monoxide. It is a leading health risk factor globally, significantly increasing the risk of various diseases, especially cancers, cardiovascular and respiratory diseases. Smoking weakens the immune system, harms nearly every organ, and shortens life expectancy. It remains one of the most preventable causes of disease and death worldwide [133]. In 2022, smoking and tobacco consumption were included in social determinants of health as a field related to chronic disease outcomes. The health problems associated with smoking would need to be addressed in a population-wide manner to introduce preventive measures to decrease tobacco consumption. The World Health Organization (WHO) published a European Health Report for 2021 that considered the introduction and increase of e-cigarettes challenging [134]. Although e-cigarettes may initially appear less harmful than traditional cigarettes, their popularity has emerged among younger generations, who often begin using them early and eventually transition to regular tobacco and cigarettes [135]. The analysis of smoking status documentation in EHRs had shown that in 80% of EHRs problems occurred (data inconsistencies, missing data, outdated information, etc.) that led to underestimations for preventive cancer screenings. Additionally, the possibility to automate this process through NLP would largely positively impact classification and prediction tasks in the health care sector [136]. Various research investigations with reference to extracting and classifying EHR variables or clinical narratives with ML and/or deep learning approaches were conducted. Yang et al. [137] employed rules to extract smoking-related data, such as packs per day, from a small dataset of 200 clinical narratives. The rule-based system

achieved a strict F1-measure of 0.94. Ruckdeschel et al. [138] aimed to select patients from clinical narratives for lung cancer detection with low-dose computed tomography. A deep learning sentence classification model via BlueBERT [139] for smoking status classification was combined with rule-based chronological processing of sentences, which favored the smoking status classification last stated in clinical narratives. The F1-measure for correctly identifying patients for cancer screening was 0.88. Bae et al. [140] implemented a keyword search and expansion with embeddings for bilingual clinical narratives in English and Korean, compared to a SVM baseline. The embeddings approach reached an F1-measure of 0.90, and compared to the baseline, it was an improvement of 1.8%.

7.2.2 Materials and Methods

Data. To create a German corpus of smoking-related snippets from clinical narratives, a rule-based expression was designed with the help of an ETL (Extract Transform Load) expert to find the various documentation types for multi-class smoking status mentions. In total, 7,242 unique de-identified snippets were extracted from the clinical narrative corpora, and classified into six classes by a physician based on the context of the snippet. These six classes, designed and created with reference to the ontology SNOMED CT, comprised of "Ex-smoker", "(Current) Smoker", "Non-smoker", "Never smoked tobacco", "Current or past smoker", and "Tobacco smoking consumption unknown". The class "Current or past smoker" was not represented in the reference ontology, which meant assigning two SNOMED CT codes through post-coordination to represent that value, i.e., the SNOMED CT code for the term "Smoker" and a temporal context value were combined in this instance. Example text spans for a "Current or past smoker" classification would be "*[...] seit 5 Jahren Nikotin-Karenz, fragl. Compliance [...]*" (nicotine-free for 5 years, questionable compliance), and "*[...] Nikotinell Pflaster bei Bedarf [...]*" (Nikotinell patches if required). In these and similar text spans, the annotators could not distinguish based on the clinical narrative whether the patient had stopped smoking.

Pre-processing. Text spans were only adapted to remove line breaks, but were not preprocessed in any other way. NLEs in any form, such as misspellings, mistyping, or linguistic variations were not changed. Depending on the applied methods for com-

Table 7.3: SNOMED CT value set and class distributions for smoking-related mentions. Reproduced from Kugic et al. [124] with permission of publisher Springer Nature.

Class	SCTID	Preferred Term	Counts
0	8517006	Ex-smoker	2,182
1	77176002	Smoker	4,255
2	8392000	Non-smoker	27
3	266919005	Never smoked tobacco	432
4	410511007; 77176002	Current or past smoker	85
5	266927001	Tobacco smoking consumption unknown	261

parison, different text representations were utilized, e.g., label encoders for LSTM and CNN models.

Model Specifics. Four ML methods, SVM, CNN, LSTM, and BERT models, were used to classify the smoking status: SVM as the classical ML method for the baseline approach, and three different variations of deep learning, where the CNN and LSTM approaches were compared with a transformer architecture via a pre-trained clinical language model, medBERT.de [141]. For each method, a nested 10x5 cross-validation approach was implemented with hyperparameter tuning and 10 different random state variations. The metric accuracy ranked hyperparameters, and selected the best ones for each method. The dataset was split into an 80% training set and 20% test set with the train_test_split function from scikit-learn [142].

Evaluation. Due to the dataset imbalance across classes, the evaluation of each method was reported with weighted average precision, recall, and F1-measure. Each calculation of random states reported slightly different performance measures due to the dataset split variations, which is why the mean, standard error, and confidence intervals for those metrics were used.

7.2.3 Results and Discussion

In Table 7.4, the results for smoking status classification were listed. The best performance was recorded by the BERT model-based approach with a mean F1-measure of 0.97. Consecutively, CNN, SVM, and LSTM followed with 0.94, 0.89, and 0.85, respectively. Trends for F1-measure were seen in recall and precision, as well. For

hyperparameter tuning results and detailed error analyses, see Kugic et al. [124]. The models demonstrated high performance on this specific dataset. However, further steps are necessary to externally validate the trained models using a separate dataset from a different clinic or specialty. This would allow for comparative results and help mitigate selection bias, as the current dataset was limited to clinical narratives from only three departments within a single hospital network.

Table 7.4: Mean performance metrics for SVM, CNN, LSTM, BERT models on the test data reported with precision, recall and F1-measure. Reproduced from Kugic et al. [124] with permission of publisher Springer Nature.

Classifier	**Metrics**	**Mean ± SE**	**95% CI**
SVM	Precision	0.894 ± 0.002	$[0.889 - 0.900]$
	Recall	0.894 ± 0.003	$[0.888 - 0.900]$
	F1-measure	0.891 ± 0.003	$[0.885 - 0.897]$
CNN	Precision	0.951 ± 0.003	$[0.944 - 0.957]$
	Recall	0.940 ± 0.002	$[0.934 - 0.945]$
	F1-measure	0.942 ± 0.002	$[0.937 - 0.948]$
LSTM	Precision	0.866 ± 0.004	$[0.856 - 0.875]$
	Recall	0.845 ± 0.005	$[0.834 - 0.856]$
	F1-measure	0.850 ± 0.006	$[0.838 - 0.862]$
BERT	Precision	0.973 ± 0.002	$[0.970 - 0.976]$
	Recall	0.972 ± 0.002	$[0.970 - 0.975]$
	F1-measure	0.973 ± 0.002	$[0.969 - 0.976]$

Chapter 8

Discussion

8.1 Scientific Knowledge Gain

Do specialized methods for NLEs (non-lexical entities) improve clinical text processing?

All specialized approaches, which improve readability and reduce ambiguity of clinical narratives, would prove beneficial in clinical natural language processing (NLP). The reduction of ambiguity, improved text normalization, applicability and adaptability for fine-tuning tasks with reference to the applied language model enhance the accuracy of NLP systems. Additionally, clinical decision-making, cohort building, and clinical routine tasks would benefit from these applications. For example, even clinicians struggle to disambiguate jargon expressions and abbreviations in clinical narratives while maintaining and reviewing patient records, and 71% of general practitioners reported that ambiguous or unexplained abbreviations could not be disambiguated in-situ [143]. Applied machine learning (ML) methods in healthcare information systems would ideally support clinicians, and benefit patients and clinicians alike, without negatively impacting workload or hindering the workflow.

However, processing NLEs in clinical narratives introduces challenges that impact resource allocation and computational requirements. Methods implemented to process NLEs can be broadly categorized into three groups: shallow approaches, deep learning techniques, and methods based on large language models (LLMs). Each of these varies in terms of complexity, accuracy, and computational workload, both during training

and application. Modeling NLEs means balancing three key parameters: the quality of annotated datasets, the approach to modeling and fine-tuning, and the hardware resources needed to train and deploy models effectively. High-quality datasets, such as gold standards, are essential but often scarce, influencing the performance of models. Annotation of new datasets can be done, though the trade-off, in this case, lies within manual annotations performed by domain experts, i.e., it requires expert resources. Additionally, deep and LLM-based methods tend to demand substantial computational resources, while shallow approaches can perform often exceptionally, but only through time-intensive manual adaptions. Processing NLEs improves clinical text processing, but the modeling approach should be tailored to the specific task for each implementation for efficient resource allocation. For the identification of NLEs, ML models did reach state-of-the-art (SOTA) results with Croatian health forums with F1-measures in the range of 0.91 to 0.93 [73]. Similar results were even recorded with the application of LLMs for Named Entity Recognition (NER) [74]. For the expansion of lexicons with NLEs, data-driven ML methods play an important role, especially when representations of text spans are not found within domain lexicons. For example, the expansion of disease terminology with co-occurrence analysis and embedding spaces [90] showed that extracting similar expressions in embeddings without labeled hierarchical classifications, i.e., term classes, such as synonyms, hypernyms, and hyponyms, underperformed per term class. It was expected that the performance by class would not reach SOTA results, i.e., the modeling of the research task did not aim to extract specific term class classifications as no similar annotated dataset in German with term class hierarchies was available. Consequently, if NLEs are not found in domain lexicons, particularly due to the high term variability in health forums or clinical narratives, the inclusion of NLEs as part of NER tasks can boost performance [73, 144]. The investigation with embeddings [90] did however show that NLEs did not impact the extraction and processing of term candidates in the embedding spaces in comparison to lexical term candidates. For the disambiguation of NLEs, in particular acronyms, the benefit of specialized NLE processing methods generally would improve the readability and clarity of clinical narratives and reduce ambiguity [41]. A comparison between text mining and LLMs [110] indicated that LLMs would outperform heuristic, rule-based approaches. Additionally, NLEs can impact processing performance if specialized methods were not applied [123, 124], and pre-processing clinical narratives, e.g., to replace NLEs with their lexical variants, could be one option to improve performance results.

Do LLMs yield better results than traditional ML methods?

With the release of ChatGPT [70], many investigations covered the potential use of LLMs, particularly regarding applications in healthcare and medical research. While generative outputs of LLMs existed prior to GPT-3, the applicability and functionality in processing the context of LLM prompts seemed revolutionary. Further iterations of LLMs encompassed huge quantities of textual data, and harness that information to answer and pass entrance exams, generate context-appropriate texts, summarize or solve complex tasks [145]. The list of functions and qualities showcased many potential use cases and opportunities, though it is more and more obvious that LLMs alone cannot realize those potential use cases. The possibilities for solving specific tasks in healthcare are promising; however, the explainability of results remains problematic due to the black-box design of LLMs. Partly responsible for this are hallucinations, i.e., false, misleading, or nonsensical information generated by LLMs that appear plausible but lack a factual basis [146]. Understanding hallucinations, i.e., differentiating between facts and fabricated information, comprehending the reasoning behind generated outputs and decisions made, as well as replicating the same LLM functions on premise, are needed to ensure patient safety and patient privacy. To guarantee patient privacy, LLMs should run on local servers using open-source models, such as Llama-2, to prevent patient data from being sent to the cloud [147]. These models must undergo rigorous performance testing for clinical NLP tasks to reduce hallucinations and minimize patient harm, therefore increasing patient safety. Additionally, bias in clinical narratives and in LLMs can lead to inaccurate or unfair clinical recommendations, especially for underrepresented populations. In 2024, a review by Yang et al. [148] elaborated on the possible origins of bias in artificial intelligence models. The authors summarized that at every stage of ML modeling, bias can occur, and various types of bias exist, such as representation, aggregation, or measurement bias, which all cause biased datasets and biased ML systems. Prior to debiasing methods, recognizing bias in ML models is often the first step to be able to counter any ill effects for patients [149].

Overall, the performances of all implemented investigations as part of this dissertation can be viewed in two ways, from the perspective of the cornerstones of NLE processing, and from a more general view on deep and shallow learning strategies versus currently popular generative strategies to perform NLP tasks. In all three cornerstones, identification, expansion and disambiguation of NLEs, a comparison of SOTA approaches with LLMs were implemented, although dataset restrictions did not always allow the use

of the same datasets for LLM application to test both methodologies. With reference to NLE processing, traditional methods consist of rules, regular expression, statistics and data-driven methods. Generative modeling via LLMs hold great potential, but the performance does not yet make the current SOTA methods obsolete, and LLMs often underperform in comparison to SOTA methods. Examples include the comparative results for the identification of NLEs, where the LLM achieved comparable results to the traditional baseline BERT method with only a third of the supplied training dataset for fine-tuning [74]. This approach exemplified that even prompting in English with a Croatian dataset performed poorly. Due to the complexity of the LLM prompt, and input and output requirements, the use of fine-tuning was necessary to reach SOTA results. Similarly, the application of LLMs achieved performance measures of 0.98 in accuracy for the disambiguation of acronyms in English clinical narratives through prompting [111]. With the application of the GPT-4 model, an accuracy range of 0.91 to 0.98 could be reached for the English dataset [111]. Traditional ML methods have shown to be equally effective for the same task, cf. Hosseini et al. [45] reached 0.96 with dataset balancing techniques and a BiLSTM model. LLMs contextually generally have a good understanding of prompts, and how to synthesize appropriate responses to those prompts. Although, even LLMs as generalist models are fine-tuned to perform well for selective tasks, such as coding or writing support, question answering, etc. [70] The negative implications of LLMs are the unpredictable answers synthesized by the models, which further propagates and strengthens the view of LLMs as black boxes. Possible factors influencing results could be language dependent, tracing back to the types of datasets and sizes used for training LLMs, and tokenization dependent, which might mean that tokenization of LLM prompts needs to be adjusted depending on domain and dataset requirements for better results [150].

Furthermore, in reviewing the traditional ML investigations for information retrieval and text classification of lifestyle risk factors, data-driven methods perform exceptionally well. A specialized BERT model reached 0.973 in F1-measure for smoking status classification [124]. With the same method, applied to alcohol status classification, with a different dataset and different language model, an F1-measure of 0.85 could be reached [123]. For alcohol consumption status, a follow-up investigation might be needed, to see whether the more specialized medBERT.de [141] language model, as applied in smoking status classification, would boost performance. The alcohol status text classification task furthermore was meant to investigate how much context from clinical narratives would be needed for processing to reach ideal results. The best per-

formance was allocated to the experiment with the most context length available. The same conclusion could also be made with prompting LLMs for information retrieval tasks, such as disambiguation of acronyms, where more context seemed to be one factor that increased the likelihood towards better results.

Are there differences in performance when comparing languages?

To assure the performance of ML tasks, a language model matching the source text language with reference to domain dictionaries would be needed [151]. For example, processing general English text with a general ML model trained on general texts would be a suitable match of language model, domain and source texts. The existence of an overlap between the linguistic properties of a particular dataset and a language model to be used in a data-driven manner would probably increase the performance of the set research task. Depending on the type of data used to create the pre-trained language model, domain adaptability might be impacted, as data creates the foundation on which the model is based [13]. Due to privacy concerns and data accessibility, foundation models are built with generalized sentences publicly available to the research community [85], e.g., Wikipedia articles, PubMed abstracts, etc. The linguistic characteristics of Wikipedia and PubMed texts often do not match the content found in clinical narratives, and therefore might decrease performance results [4, 151], i.e., clinical narratives often comprise no grammatical structure, short sentences, non-standard content based on a multitude of medical domains, etc.

Since 2023, models have started to become available to the research community, which are based on clinical narratives in German, e.g., medBERT.de [141] or BioGottBERT [10]). These were built with transformer architectures [5], and have shown consistent results in text classification tasks. Based on the performed investigation for smoking status classification [124], a language model pre-trained on the same domain as the source dataset, cf. medBERT.de [141], and further fine-tuned for a classification task, outperforms CNN [130, 131], LSTM [18] and SVM [39] baselines. The modeling to solve smoking status classification was very specifically tailored to the problem statement, and generalizability on a separate dataset in German was not tested. Documentation practices vary and the application of the same fine-tuned model across other datasets in German would lead to a decrease in performance, if the model is not updated with another more comprehensive dataset in the same manner.

Specifically for LLMs, performance seemed to decrease after changing dataset and prompt language [111] from English to German and Portuguese in the acronym dis-

ambiguation task. The application of GPT models showed a significant decrease in accuracy based on the confidence intervals between languages. In small scale tests, prompting for acronym resolutions with either German or English for the instructions did not seem to make a significant impact. In the application of the Llama-2 models with the same prompts, the results for English were significantly lower in comparison to GPT results, but still applicable with an accuracy range of 0.70 to 0.73. Results for languages other than English for Llama models were unusable ranging from 0.02 to 0.45 in performance.

8.2 Conclusion and Outlook

Non-lexical entities (NLEs) can cause miscommunication or misunderstandings that can impact patient care. In particular, the way clinical texts are phrased and written can introduce more ambiguity and uncertainty for the reader in cognitively understanding the information. The ambiguity of NLEs, particularly those with multiple meanings, can hinder comprehension for clinicians, who rely on domain understanding, and for laypeople, who may struggle with unfamiliar clinical language. With natural language processing (NLP), clinical narratives can be analyzed and used as datasets to perform various tasks through the application of machine learning (ML) approaches. Due to the pervasiveness of NLEs, these challenging aspects can be seen as documentation traits by clinicians, and while documentation provides added value in a clinical setting, NLEs also are seen as a hindrance for processing clinical narratives effectively. The reuse of electronic health records (EHRs) offers a high value to train accurate and applicable ML models to solve clinical routine and biomedical research tasks. These ML tasks span multiple disciplines and methodologies with the goal to extract, classify, enrich, and/or summarize the datasets and information fed into these applications, e.g., for clinical decision support, cohort building, disambiguation of NLEs, etc.

The processing of NLEs was investigated, and for the three cornerstones in NLE processing, identification of NLEs, expansion of lexicons, and disambiguation of NLEs, research questions were formulated to understand the challenges and provide solutions. The identification of NLEs with named entity recognition through sequence labeling was performed, expansion of lexicons were achieved through embeddings-based representations, and disambiguation was implemented through text mining. For each cornerstone, additional experiments showcased the applicability of LLMs. One sub-

type of NLEs, short forms, were investigated in more depth to assess the best NLP processing methods for recognizing short forms, expanding short forms to their possible long forms, and correctly selecting or disambiguating the long form version of a short form. Finally, information extraction tasks for lifestyle-related risk factors made it possible to implement text classifications of clinical narratives with NLEs. These implementations would be used in the context of cohort building, to find patients for cancer screenings, or to populate structured fields in EHRs with relevant data when such information is initially absent.

Future work will focus on refining identification, lexicon expansion, and disambiguation of NLEs to further improve the accuracy of NLP applications in clinical contexts. Leveraging LLMs for advanced text mining and embedding-based methods could enhance the ability to process and interpret ambiguous clinical narratives, particularly in terminology expansion. Additionally, integrating domain-specific knowledge with LLMs could optimize the interpretation of NLEs, enabling more precise and context-aware applications in clinical decision support and information extraction.

Bibliography

[1] Moy A. J., Hobensack M., Marshall K., Vawdrey D. K., Kim E. Y., Cato K. D., and Rossetti S. C. Understanding the perceived role of electronic health records and workflow fragmentation on clinician documentation burden in emergency departments. *Journal of the American Medical Informatics Association*, 30(5): 797–808, May 2023. ISSN 1527-974X. doi: 10.1093/jamia/ocad038. URL https://doi.org/10.1093/jamia/ocad038.

[2] Moy A. J., Schwartz J. M., Chen R., Sadri S., Lucas E., Cato K. D., and Rossetti S. C. Measurement of clinical documentation burden among physicians and nurses using electronic health records: a scoping review. *Journal of the American Medical Informatics Association*, 28(5):998–1008, May 2021. ISSN 1527-974X. doi: 10.1093/jamia/ocaa325. URL https://doi.org/10.1093/jamia/ocaa325.

[3] Del Rio-Bermudez C., Medrano I. H., Yebes L., and Poveda J. L. Towards a symbiotic relationship between big data, artificial intelligence, and hospital pharmacy. *Journal of Pharmaceutical Policy and Practice*, 13(1):75, November 2020. ISSN 2052-3211. doi: 10.1186/s40545-020-00276-6.

[4] Kreuzthaler M., Brochhausen M., Zayas C., Blobel B., and Schulz S. Linguistic and ontological challenges of multiple domains contributing to transformed health ecosystems. *Frontiers in Medicine*, 10, March 2023. ISSN 2296-858X. doi: 10.3389/fmed.2023.1073313. URL https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2023.1073313/full. Publisher: Frontiers.

[5] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., and Polosukhin I. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[6] Nerella S., Bandyopadhyay S., Zhang J., Contreras M., Siegel S., Bumin A., Silva B., Sena J., Shickel B., Bihorac A., Khezeli K., and Rashidi P. Transformers and large language models in healthcare: A review. *Artificial Intelligence in Medicine*, 154:102900, August 2024. ISSN 09333657. doi: 10.1016/j.artmed.2024.102900. URL https://linkinghub.elsevier.com/retrieve/pii/S0933365724001428.

[7] Wu S., Roberts K., Datta S., Du J., Ji Z., Si Y., Soni S., Wang Q., Wei Q., Xiang Y., Zhao B., and Xu H. Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, 27(3):457–470, March 2020. ISSN 1527-974X. doi: 10.1093/jamia/ocz200. URL https://doi.org/10.1093/jamia/ocz200.

[8] Merriam-Webster . Definition of DICTIONARY, December 2024. URL https://www.merriam-webster.com/dictionary/dictionary.

[9] Kersloot M. G., Putten van F. J. P., Abu-Hanna A., Cornet R., and Arts D. L. Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies. *Journal of Biomedical Semantics*, 11(1):14, November 2020. ISSN 2041-1480. doi: 10.1186/s13326-020-00231-z.

[10] Lentzen M., Madan S., Lage-Rupprecht V., Kühnel L., Fluck J., Jacobs M., Mittermaier M., Witzenrath M., Brunecker P., Hofmann-Apitius M., Weber J., and Fröhlich H. Critical assessment of transformer-based ai models for german clinical notes. *JAMIA Open*, 5(4):ooac087, 11 2022. ISSN 2574-2531. doi: 10.1093/jamiaopen/ooac087. URL https://doi.org/10.1093/jamiaopen/ooac087.

[11] Chute C. G. Clinical classification and terminology: Some history and current observations. *Journal of the American Medical Informatics Association*, 7(3): 298–303, 05 2000. ISSN 1067-5027. doi: 10.1136/jamia.2000.0070298. URL https://doi.org/10.1136/jamia.2000.0070298.

[12] Zeng Q. T., Tse T., Divita G., Keselman A., Crowell J., Browne A. C., Goryachev S., and Ngo L. Term identification methods for consumer health vocabulary development. *J Med Internet Res*, 9(1):e4, Mar 2007. ISSN 1438-8871. doi: 10.2196/jmir.9.1.e4. URL http://www.jmir.org/2007/1/e4/.

[13] Spasic I. and Nenadic G. Clinical Text Data in Machine Learning: Systematic Review. *JMIR medical informatics*, 8(3):e17984, March 2020. ISSN 2291-9694. doi: 10.2196/17984.

[14] Zeng Q., Tse T., Divita G., Keselman A., Crowell J., Browne A., Goryachev S., and Ngo L. Term Identification Methods for Consumer Health Vocabulary Development. *Journal of Medical Internet Research*, 9(1):e606, March 2007. doi: 10.2196/jmir.9.1.e4. URL https://www.jmir.org/2007/1/e4. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[15] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–270, January 2004. ISSN 1362-4962. doi: 10.1093/nar/gkh061.

[16] Faris H., Faris M., Habib M., and Alomari A. Automatic symptoms identification from a massive volume of unstructured medical consultations using deep neural and BERT models. *Heliyon*, 8(6):e09683, June 2022. ISSN 2405-8440. doi: 10.1016/j.heliyon.2022.e09683. URL https://www.sciencedirect.com/science/article/pii/S2405844022009719.

[17] Antoun W., Baly F., and Hajj H. AraBERT: Transformer-based Model for Arabic Language Understanding. In Al-Khalifa H., Magdy W., Darwish K., Elsayed T., and Mubarak H., editors, *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France, May 2020. European Language Resource Association. ISBN 979-10-95546-51-1. URL https://aclanthology.org/2020.osact-1.2.

[18] Hochreiter S. and Schmidhuber J. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

[19] Sogandi F. Identifying diseases symptoms and general rules using supervised and unsupervised machine learning. *Scientific Reports*, 14(1):17956, August 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-69029-8. URL https://www.nature.com/articles/s41598-024-69029-8. Publisher: Nature Publishing Group.

[20] Xu H., Stetson P. D., and Friedman C. A Study of Abbreviations in Clinical Notes. *AMIA Annual Symposium Proceedings*, 2007:821–825, 2007. ISSN 1942-597X. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655910/.

[21] Kim Y., Hurdle J., and Meystre S. M. Using UMLS lexical resources to disambiguate abbreviations in clinical text. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2011:715–722, 2011. ISSN 1942-597X.

[22] Savova G. K., Masanz J. J., Ogren P. V., Zheng J., Sohn S., Kipper-Schuler K. C., and Chute C. G. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 09 2010. ISSN 1067-5027. doi: 10.1136/jamia.2009.001560. URL https://doi.org/10.1136/jamia.2009.001560.

[23] Awaysheh A., Wilcke J., Elvinger F., Rees L., Fan W., and Zimmerman K. A review of medical terminology standards and structured reporting. *Journal of Veterinary Diagnostic Investigation*, 30(1):17–25, January 2018. ISSN 1040-6387, 1943-4936. doi: 10.1177/1040638717738276. URL https://journals.sagepub.com/doi/10.1177/1040638717738276.

[24] Nuopponen A. 23. Terminological Concept Systems. In Humbley J., Budin G., and Laurén C., editors, *Languages for Special Purposes*, pages 453–468. De Gruyter, October 2018. ISBN 978-3-11-022801-4. doi: 10.1515/9783110228014-023. URL https://www.degruyter.com/document/doi/10.1515/9783110228014-023/html.

[25] Noll R., Frischen L. S., Boeker M., Storf H., and Schaaf J. Machine translation of standardised medical terminology using natural language processing: A scoping review. *New Biotechnology*, 77:120–129, November 2023. ISSN 1871-6784. doi: 10.1016/j.nbt.2023.08.004. URL https://www.sciencedirect.com/science/article/pii/S1871678423000432.

[26] Sarker A. LexExp: a system for automatically expanding concept lexicons for noisy biomedical texts. *Bioinformatics (Oxford, England)*, 37(16):2499–2501, August 2021. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaa995.

[27] Levenshtein V. I. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union,

1966. URL https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf. Issue: 8.

[28] Sarker A. and Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics*, 53:196–207, February 2015. ISSN 1532-0480. doi: 10.1016/j.jbi.2014.11.002.

[29] Koroleva A., Kamath S., and Paroubek P. Measuring semantic similarity of clinical trial outcomes using deep pre-trained language representations. *Journal of Biomedical Informatics*, 100:100058, 2019. ISSN 15320464. doi: 10.1016/j.yjbinx.2019.100058. URL https://linkinghub.elsevier.com/retrieve/pii/S2590177X19300575.

[30] Devlin J., Chang M.-W., Lee K., and Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein J., Doran C., and Solorio T., editors, *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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[31] Lee J., Yoon W., Kim S., Kim D., Kim S., So C. H., and Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, February 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz682. URL https://doi.org/10.1093/bioinformatics/btz682.

[32] Beltagy I., Lo K., and Cohan A. SciBERT: A Pretrained Language Model for Scientific Text. In Inui K., Jiang J., Ng V., and Wan X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL https://aclanthology.org/D19-1371.

[33] Carpenter K. A. and Altman R. B. Using GPT-3 to Build a Lexicon of Drugs of Abuse Synonyms for Social Media Pharmacovigilance. *Biomolecules*, 13(2):387, February 2023. ISSN 2218-273X. doi: 10.3390/biom13020387. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9953178/.

[34] Wu Y., Denny J. C., Rosenbloom S. T., Miller R. A., Giuse D. A., and Xu H. A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries. *AMIA Annual Symposium Proceedings*, 2012:997–1003, November 2012. ISSN 1942-597X. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3540461/.

[35] Aronson A. R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings. AMIA Symposium*, pages 17–21, 2001. ISSN 1531-605X.

[36] Friedman C., Shagina L., Socratous S. A., and Zeng X. A WEB-Based Version of MedLEE: A Medical Language Extraction and Encoding System. *Proceedings of the AMIA Annual Fall Symposium*, page 938, 1996. ISSN 1091-8280. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2233000/.

[37] Friedman C., Hripcsak G., DuMouchel W., Johnson S. B., and Clayton P. D. Natural language processing in an operational clinical information system. *Natural Language Engineering*, 1(1):83–108, 1995. doi: 10.1017/S1351324900000061.

[38] Wu Y., Denny J. C., Rosenbloom S. T., Miller R. A., Giuse D. A., Song M., and Xu H. A Preliminary Study of Clinical Abbreviation Disambiguation in Real Time. *Applied Clinical Informatics*, 06(02):364–374, 2015. ISSN 1869-0327. doi: 10.4338/ACI-2014-10-RA-0088. URL http://www.thieme-connect.de/DOI/DOI?10.4338/ACI-2014-10-RA-0088.

[39] Cortes C. and Vapnik V. Support-vector networks. *Machine Learning*, 20(3): 273–297, September 1995. ISSN 1573-0565. doi: 10.1007/BF00994018. URL https://doi.org/10.1007/BF00994018.

[40] Mowery D. L., South B. R., Christensen L., Leng J., Peltonen L.-M., Salanterä S., Suominen H., Martinez D., Velupillai S., Elhadad N., Savova G., Pradhan S., and Chapman W. W. Normalizing acronyms and abbreviations to aid patient understanding of clinical texts: ShARe/CLEF eHealth Challenge 2013, Task 2. *Journal of Biomedical Semantics*, 7(1):1–13, December 2016. ISSN 2041-1480. doi: 10.1186/s13326-016-0084-y. URL https://jbiomedsem.biomedcentral.com/articles/10.1186/s13326-016-0084-y. Number: 1 Publisher: BioMed Central.

[41] Suominen H., Salanterä S., Velupillai S., Chapman W. W., Savova G., Elhadad N., Pradhan S., South B. R., Mowery D. L., Jones G. J. F., Leveling J., Kelly

L., Goeuriot L., Martinez D., and Zuccon G. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In Forner P., Müller H., Paredes R., Rosso P., and Stein B., editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 212–231, Berlin, Heidelberg, 2013. Springer. ISBN 978-3-642-40802-1. doi: 10.1007/978-3-642-40802-1_24.

[42] Lafferty J. D., McCallum A., and Pereira F. C. N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, June 2001. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-778-1.

[43] Adams G., Ketenci M., Bhave S., Perotte A., and Elhadad N. Zero-Shot Clinical Acronym Expansion via Latent Meaning Cells. In *Proceedings of Machine Learning Research*, 2020.

[44] Moon S., Pakhomov S., and Melton G. Clinical Abbreviation Sense Inventory, October 2012. URL http://conservancy.umn.edu/handle/11299/137703. Accepted: 2012-10-31T19:58:41Z.

[45] Hosseini M., Rasekh A. H., and Keshavarzi A. Improving clinical abbreviation sense disambiguation using attention-based Bi-LSTM and hybrid balancing techniques in imbalanced datasets. *Journal of Evaluation in Clinical Practice*, 30(7): 1327–1336, October 2024. ISSN 1365-2753. doi: 10.1111/jep.14041.

[46] Wen Z., Lu X. H., and Reddy S. MeDAL: Medical Abbreviation Disambiguation Dataset for Natural Language Understanding Pretraining. In Rumshisky A., Roberts K., Bethard S., and Naumann T., editors, *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 130–135, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.clinicalnlp-1.15. URL https://aclanthology.org/2020.clinicalnlp-1.15.

[47] Agrawal M., Hegselmann S., Lang H., Kim Y., and Sontag D. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.130. URL https://aclanthology.org/2022.emnlp-main.130.

[48] Kugic A., Martin I., Modersohn L., Pallaoro P., Kreuzthaler M., Schulz S., and Boeker M. Processing of Short-Form Content in Clinical Narratives: Systematic Scoping Review. *Journal of Medical Internet Research*, 26:e57852, September 2024. ISSN 1438-8871. doi: 10.2196/57852. URL https://www.jmir.org/2024/1/e57852.

[49] Hardman W., Banks M., Davidson R., Truran D., Ayuningtyas N. W., Ngo H., Johnson A., and Pollard T. SNOMED CT Entity Linking Challenge, 2023. URL https://physionet.org/content/snomed-ct-entity-challenge/1.0.0/.

[50] Johnson A. E. W., Bulgarelli L., Shen L., Gayles A., Shammout A., Horng S., Pollard T. J., Hao S., Moody B., Gow B., Lehman L.-w. H., Celi L. A., and Mark R. G. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, January 2023. ISSN 2052-4463. doi: 10.1038/s41597-022-01899-x. URL https://www.nature.com/articles/s41597-022-01899-x. Publisher: Nature Publishing Group.

[51] Oliveira L. E. S. E., Peters A. C., Da Silva A. M. P., Gebeluca C. P., Gumiel Y. B., Cintho L. M. M., Carvalho D. R., Al Hasan S., and Moro C. M. C. SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks. *Journal of Biomedical Semantics*, 13(1):13, December 2022. ISSN 2041-1480. doi: 10.1186/s13326-022-00269-1. URL https://jbiomedsem.biomedcentral.com/articles/10.1186/s13326-022-00269-1.

[52] Moon S., Pakhomov S., Liu N., Ryan J. O., and Melton G. B. A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources. *Journal of the American Medical Informatics Association*, 21(2):299–307, March 2014. ISSN 1067-5027, 1527-974X. doi: 10.1136/amiajnl-2012-001506. URL https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2012-001506.

[53] Beckers H. *Abkürzungslexikon medizinischer Begriffe: einschl. Randgebiete*. Fachterminologie Medizin. Verlag Arzt + Information, Köln, 8. ergänzte auflage edition, 2015. ISBN 978-3-9807384-7-7.

[54] Wikipedia . Liste medizinischer Abkürzungen, February 2025. URL https://de.wikipedia.org/w/index.php?title=Liste_medizinischer_Abk%C3%BCrzungen&oldid=253036258. Page Version ID: 253036258.

[55] Gjuran-Coha A. Terminologizacija jezika medicinske struke. *medicina fluminensis*, 47(1), 2011.

[56] Fabijanić I. and Malenica F. Abbreviations in English medical terminology and their adaptation to Croatian. *European Journal of Bioethics*, 4(7), 2013.

[57] Kocijan K., Kurolt S., and Mijić L. Building Croatian Medical Dictionary from Medical Corpus. *Rasprave Instituta za hrvatski jezik*, 46(2):765–782, October 2020. ISSN 1331-6745, 1849-0379. doi: 10.31724/rihjj.46.2.17. URL https://hrcak.srce.hr/clanak/356595. Publisher: Institut za hrvatski jezik i jezikoslovlje.

[58] Čamić K. *Elektronička izdanja Leksikografskog zavoda Miroslav Krleža*. info:eu-repo/semantics/masterThesis, University of Zagreb. Faculty of Humanities and Social Sciences. Department of information and Communication sciences, September 2021. URL https://urn.nsk.hr/urn:nbn:hr:131:157813.

[59] Karasman I. S. Radovi Leksikografskog zavoda Miroslav Krleža. 2008.

[60] Djordjević S. P. *Medical Dictionary: Serbian and Croatian-English*. Jordana Pub, Banning, Calif, 2009. ISBN 978-0-9764480-0-6.

[61] Alsentzer E., Murphy J., Boag W., Weng W.-H., Jindi D., Naumann T., and McDermott M. Publicly Available Clinical BERT Embeddings. In Rumshisky A., Roberts K., Bethard S., and Naumann T., editors, *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL https://aclanthology.org/W19-1909.

[62] Zubiaga A. Natural language processing in the era of large language models. *Frontiers in Artificial Intelligence*, 6, January 2024. ISSN 2624-8212. doi: 10.3389/frai.2023.1350306. URL https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2023.1350306/full. Publisher: Frontiers.

[63] Wang L., Wan Z., Ni C., Song Q., Li Y., Clayton E., Malin B., and Yin Z. Applications and Concerns of ChatGPT and Other Conversational Large Language Models in Health Care: Systematic Review. *Journal of Medical Internet Research*, 26(1):e22769, November 2024. doi: 10.2196/22769. URL https://www.jmir.org/2024/1/e22769. Company: Journal of Medical Internet

Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[64] Ling C., Zhao X., Lu J., Deng C., Zheng C., Wang J., Chowdhury T., Li Y., Cui H., Zhang X., Zhao T., Panalkar A., Mehta D., Pasquali S., Cheng W., Wang H., Liu Y., Chen Z., Chen H., White C., Gu Q., Pei J., Yang C., and Zhao L. Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey, March 2024. URL http://arxiv.org/abs/2305.18703. arXiv:2305.18703 [cs].

[65] Li H. Language models: past, present, and future. *Communications of the ACM*, 65(7):56–63, July 2022. ISSN 0001-0782, 1557-7317. doi: 10.1145/3490443. URL https://dl.acm.org/doi/10.1145/3490443.

[66] Liu F., Shareghi E., Meng Z., Basaldella M., and Collier N. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, June 2021.

[67] Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L., and Stoyanov V. Unsupervised Cross-lingual Representation Learning at Scale. In Jurafsky D., Chai J., Schluter N., and Tetreault J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

[68] Ljubešić N. and Lauc D. BERTić - the transformer language model for Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42, Kiyv, Ukraine, April 2021. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2021.bsnlp-1.5.

[69] Clark K., Luong M.-T., Le Q. V., and Manning C. D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators, March 2020. URL http://arxiv.org/abs/2003.10555. arXiv:2003.10555 [cs].

[70] Radford A., Narasimhan K., Salimans T., and Sutskever I. Improving Language Understanding by Generative Pre-Training. 2018.

[71] Touvron H., Martin L., Stone K., Albert P., Almahairi A., Babaei Y., Bashlykov N., Batra S., Bhargava P., Bhosale S., Bikel D., Blecher L., Ferrer C. C., Chen M., Cucurull G., Esiobu D., Fernandes J., Fu J., Fu W., Fuller B., Gao C., Goswami V., Goyal N., Hartshorn A., Hosseini S., Hou R., Inan H., Kardas M., Kerkez V., Khabsa M., Kloumann I., Korenev A., Koura P. S., Lachaux M.-A., Lavril T., Lee J., Liskovich D., Lu Y., Mao Y., Martinet X., Mihaylov T., Mishra P., Molybog I., Nie Y., Poulton A., Reizenstein J., Rungta R., Saladi K., Schelten A., Silva R., Smith E. M., Subramanian R., Tan X. E., Tang B., Taylor R., Williams A., Kuan J. X., Xu P., Yan Z., Zarov I., Zhang Y., Fan A., Kambadur M., Narang S., Rodriguez A., Stojnic R., Edunov S., and Scialom T. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. URL http://arxiv.org/abs/2307.09288. arXiv:2307.09288 [cs].

[72] Page M. J., McKenzie J. E., Bossuyt P. M., Boutron I., Hoffmann T. C., Mulrow C. D., Shamseer L., Tetzlaff J. M., Akl E. A., Brennan S. E., Chou R., Glanville J., Grimshaw J. M., Hróbjartsson A., Lalu M. M., Li T., Loder E. W., Mayo-Wilson E., McDonald S., McGuinness L. A., Stewart L. A., Thomas J., Tricco A. C., Welch V. A., Whiting P., and Moher D. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, 2021. doi: 10.1136/bmj.n71. URL https://www.bmj.com/content/372/bmj.n71.

[73] Kugic A., Schulz S., and Kreuzthaler M. Identification of Non-Lexical Content in Croatian Health Forum Entries. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 4328–4335, Istanbul, Turkiye, December 2023. IEEE. ISBN 9798350337488. doi: 10.1109/BIBM58861.2023.10386069. URL https://ieeexplore.ieee.org/document/10386069/.

[74] Kugic A., Kreuzthaler M., and Schulz S. Annotation of Non-Lexical Entities in Croatian Health Forum Entries With Large Language Models. In *Book of Abstracts of the XXI EURALEX International Congress*, pages 147–151, Cavtat, Croatia, 2024. Institut za hrvatski jezik. ISBN 978-953-7967-74-1.

[75] Heryawan L., Sugiyama O., Yamamoto G., Khotimah P. H., Santos L. H. O., Okamoto K., and Kuroda T. A Detection of Informal Abbreviations from Free

Text Medical Notes Using Deep Learning. *European Journal of Biomedical Informatics*, 2020. URL https://www.ejbi.org/abstract/a-detection-of-informa l-abbreviations-from-free-text-medical-notes-using-deep-learning-5680.html. Publisher: European Journal of Biomedical Informatics.

[76] Huang X., Zhang E., and Koh Y. S. Supervised Clinical Abbreviations Detection and Normalisation Approach. In Nayak A. C. and Sharma A., editors, *PRICAI 2019: Trends in Artificial Intelligence*, Lecture Notes in Computer Science, pages 691–703, Cham, 2019. Springer International Publishing. ISBN 978-3-030-29894-4. doi: 10.1007/978-3-030-29894-4_55.

[77] Kaplar A., Stošović M., Kaplar A., Brković V., Naumović R., and Kovačević A. Evaluation of clinical named entity recognition methods for Serbian electronic health records. *International Journal of Medical Informatics*, 164:104805, August 2022. ISSN 1872-8243. doi: 10.1016/j.ijmedinf.2022.104805.

[78] Chinchor N. and Sundheim B. MUC-5 Evaluation Metrics. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*, 1993. URL https://aclanthology.org/M 93-1007.

[79] Wolf T., Debut L., Sanh V., Chaumond J., Delangue C., Moi A., Cistac P., Rault T., Louf R., Funtowicz M., Davison J., Shleifer S., Platen von P., Ma C., Jernite Y., Plu J., Xu C., Scao T. L., Gugger S., Drame M., Lhoest Q., and Rush A. M. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos .6.

[80] Hu Y., Chen Q., Du J., Peng X., Keloth V. K., Zuo X., Zhou Y., Li Z., Jiang X., Lu Z., Roberts K., and Xu H. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1812–1820, September 2024. ISSN 1527-974X. doi: 10.1093/jamia/ocad259. URL https://doi.org/10.1093/jamia/ocad259.

[81] Keloth V. K., Hu Y., Xie Q., Peng X., Wang Y., Zheng A., Selek M., Raja K., Wei C. H., Jin Q., Lu Z., Chen Q., and Xu H. Advancing entity recognition in

biomedicine via instruction tuning of large language models. *Bioinformatics*, 40 (4):btae163, April 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae163. URL https://doi.org/10.1093/bioinformatics/btae163.

[82] Gu Y., Tinn R., Cheng H., Lucas M., Usuyama N., Liu X., Naumann T., Gao J., and Poon H. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, January 2022. ISSN 2691-1957, 2637-8051. doi: 10.1145/3458754. URL http://arxiv.org/abs/2007.15779. arXiv:2007.15779 [cs].

[83] Wu C., Lin W., Zhang X., Zhang Y., Xie W., and Wang Y. PMC-LLaMA: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843, September 2024. ISSN 1527-974X. doi: 10.1093/jamia/ocae045. URL https://doi.org/10.1093/jamia/ocae045.

[84] Wornow M., Xu Y., Thapa R., Patel B., Steinberg E., Fleming S., Pfeffer M. A., Fries J., and Shah N. H. The shaky foundations of large language models and foundation models for electronic health records. *NPJ digital medicine*, 6(1):135, July 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00879-8.

[85] Bommasani R., Hudson D. A., Adeli E., Altman R., Arora S., Arx S. v., Bernstein M. S., Bohg J., Bosselut A., Brunskill E., Brynjolfsson E., Buch S., Card D., Castellon R., Chatterji N., Chen A., Creel K., Davis J. Q., Demszky D., Donahue C., Doumbouya M., Durmus E., Ermon S., Etchemendy J., Ethayarajh K., Fei-Fei L., Finn C., Gale T., Gillespie L., Goel K., Goodman N., Grossman S., Guha N., Hashimoto T., Henderson P., Hewitt J., Ho D. E., Hong J., Hsu K., Huang J., Icard T., Jain S., Jurafsky D., Kalluri P., Karamcheti S., Keeling G., Khani F., Khattab O., Koh P. W., Krass M., Krishna R., Kuditipudi R., Kumar A., Ladhak F., Lee M., Lee T., Leskovec J., Levent I., Li X. L., Li X., Ma T., Malik A., Manning C. D., Mirchandani S., Mitchell E., Munyikwa Z., Nair S., Narayan A., Narayanan D., Newman B., Nie A., Niebles J. C., Nilforoshan H., Nyarko J., Ogut G., Orr L., Papadimitriou I., Park J. S., Piech C., Portelance E., Potts C., Raghunathan A., Reich R., Ren H., Rong F., Roohani Y., Ruiz C., Ryan J., Ré C., Sadigh D., Sagawa S., Santhanam K., Shih A., Srinivasan K., Tamkin A., Taori R., Thomas A. W., Tramèr F., Wang R. E., Wang W., Wu B., Wu J., Wu Y., Xie S. M., Yasunaga M., You J., Zaharia M., Zhang M., Zhang T.,

Zhang X., Zhang Y., Zheng L., Zhou K., and Liang P. On the Opportunities and Risks of Foundation Models, July 2022. URL http://arxiv.org/abs/2108.07258. arXiv:2108.07258 [cs].

[86] Ramachandran G. K., Fu Y., Han B., Lybarger K., Dobbins N., Uzuner O., and Yetisgen M. Prompt-based Extraction of Social Determinants of Health Using Few-shot Learning. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 385–393, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.clinicalnlp-1.41. URL https://aclanthology.org/2023.clinicalnlp-1.41.

[87] Ben Abacha A., Yim W.-w., Adams G., Snider N., and Yetisgen M. Overview of the MEDIQA-Chat 2023 Shared Tasks on the Summarization & Generation of Doctor-Patient Conversations. In Naumann T., Ben Abacha A., Bethard S., Roberts K., and Rumshisky A., editors, *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 503–513, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.clinicalnlp-1.52. URL https://aclanthology.org/2023.clinicalnlp-1.52.

[88] Kurtyigit S., Park M., Schlechtweg D., Kuhn J., and Walde Schulte im S. Lexical Semantic Change Discovery. In Zong C., Xia F., Li W., and Navigli R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6985–6998, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.543. URL https://aclanthology.org/2021.acl-long.543.

[89] Kugic A., Pfeifer B., Schulz S., and Kreuzthaler M. Data-Driven Identification of Clinical Real-World Expressions Linked to ICD. *Studies in Health Technology and Informatics*, 302:827–828, May 2023. ISSN 1879-8365. doi: 10.3233/SHTI230279.

[90] Kugic A., Pfeifer B., Schulz S., and Kreuzthaler M. Embedding-based terminology expansion via secondary use of large clinical real-world datasets. *Journal of Biomedical Informatics*, page 104497, September 2023. ISSN 1532-0464. doi: 10.1016/j.jbi.2023.104497. URL https://www.sciencedirect.com/science/article/pii/S1532046423002186.

[91] Kugic A., Schulz S., and Kreuzthaler M. Term Candidate Generation to Enrich Clinical Terminologies with Large Language Models. In Mantas J., Hasman A., Demiris G., Saranto K., Marschollek M., Arvanitis T. N., Ognjanović I., Benis A., Gallos P., Zoulias E., and Andrikopoulou E., editors, *Studies in Health Technology and Informatics*. IOS Press, August 2024. ISBN 978-1-64368-533-5. doi: 10.3233/SHTI240509. URL https://ebooks.iospress.nl/doi/10.3233/SHTI240509.

[92] Yalçin R. Research on Gambling in Young People: A Co-Occurrence Analysis. *Journal of Gambling Studies*, 39(2):531–539, May 2022. ISSN 1573-3602. doi: 10.1007/s10899-022-10131-9. URL https://link.springer.com/10.1007/s10899-022-10131-9.

[93] Nesaragi N., Patidar S., and Aggarwal V. Tensor learning of pointwise mutual information from EHR data for early prediction of sepsis. *Computers in Biology and Medicine*, 134:104430, July 2021. ISSN 00104825. doi: 10.1016/j.compbiomed.2021.104430. URL https://linkinghub.elsevier.com/retrieve/pii/S0010482521002249.

[94] Watford S. M., Grashow R. G., De La Rosa V. Y., Rudel R. A., Friedman K. P., and Martin M. T. Novel application of normalized pointwise mutual information (NPMI) to mine biomedical literature for gene sets associated with disease: Use case in breast carcinogenesis. *Computational Toxicology*, 7:46–57, August 2018. ISSN 24681113. doi: 10.1016/j.comtox.2018.06.003. URL https://linkinghub.elsevier.com/retrieve/pii/S246811131830029X.

[95] Fan Y., Pakhomov S., McEwan R., Zhao W., Lindemann E., and Zhang R. Using word embeddings to expand terminology of dietary supplements on clinical notes. *JAMIA Open*, 2(2):246–253, March 2019. ISSN 2574-2531. doi: 10.1093/jamiaopen/ooz007. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6904105/.

[96] Wang Y., Liu S., Afzal N., Rastegar-Mojarad M., Wang L., Shen F., Kingsbury P., and Liu H. A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics*, 87:12–20, November 2018. ISSN 1532-0464. doi: 10.1016/j.jbi.2018.09.008. URL https://www.sciencedirect.com/science/article/pii/S1532046418301825.

[97] Gu G., Zhang X., Zhu X., Jian Z., Chen K., Wen D., Gao L., Zhang S., Wang F., Ma H., and Lei J. Development of a Consumer Health Vocabulary by

Mining Health Forum Texts Based on Word Embedding: Semiautomatic Approach. *JMIR medical informatics*, 7(2):e12704, May 2019. ISSN 2291-9694. doi: 10.2196/12704.

[98] Kreuzthaler M., Pfeifer B., and Schulz S. Terminology Expansion via Co-occurrence Analysis of Large Clinical Real-World Datasets. In *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*, pages 01–02, Rochester, MN, USA, June 2022. IEEE. ISBN 978-1-66546-845-9. doi: 10.1109/ICHI54592.2022.00124. URL https://ieeexplore.ieee.org/document/9874543/.

[99] Chai C. P. Comparison of text preprocessing methods. *Natural Language Engineering*, 29(3):509–553, May 2023. ISSN 1351-3249, 1469-8110. doi: 10.1017/S1351324922000213. URL https://www.cambridge.org/core/journals/natural-language-engineering/article/comparison-of-text-preprocessing-methods/43A20821D65F1C0C4366B126FC794AE3.

[100] Liu F., Vulić I., Korhonen A., and Collier N. Learning domain-specialised representations for cross-lingual biomedical entity linking. In *Proceedings of ACL-IJCNLP 2021*, pages 565–574, August 2021.

[101] Johnson J., Douze M., and Jégou H. Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, July 2021. ISSN 2332-7790. doi: 10.1109/TBDATA.2019.2921572. URL https://ieeexplore.ieee.org/document/8733051. Conference Name: IEEE Transactions on Big Data.

[102] Koleck T. A., Tatonetti N. P., Bakken S., Mitha S., Henderson M. M., George M., Miaskowski C., Smaldone A., and Topaz M. Identifying Symptom Information in Clinical Notes Using Natural Language Processing. *Nursing Research*, 70(3):173–183, May 2021. ISSN 1538-9847, 0029-6562. doi: 10.1097/NNR.0000000000000488. URL https://journals.lww.com/10.1097/NNR.0000000000000488.

[103] Topaz M., Murga L., Bar-Bachar O., McDonald M., and Bowles K. NimbleMiner: An Open-Source Nursing-Sensitive Natural Language Processing System Based on Word Embedding. *Computers, informatics, nursing: CIN*, 37(11):583–590, November 2019. ISSN 1538-9774. doi: 10.1097/CIN.0000000000000557.

[104] Mikolov T., Chen K., Corrado G., and Dean J. Efficient Estimation of Word Representations in Vector Space, September 2013. URL http://arxiv.org/abs/1301.3781. arXiv:1301.3781 [cs].

[105] Zhang J., Zhang Z., Zhang H., Ma Z., Ye Q., He P., and Zhou Y. From electronic health records to terminology base: A novel knowledge base enrichment approach. *Journal of Biomedical Informatics*, 113:103628, January 2021. ISSN 15320464. doi: 10.1016/j.jbi.2020.103628. URL https://linkinghub.elsevier.com/retrieve/pii/S1532046420302562.

[106] Pang N., Zeng W., Tang J., Tan Z., and Zhao X. Iterative entity alignment with improved neural attribute embedding. In *DL4KG@ ESWC*, pages 41–46, 2019.

[107] Wang Z., Lv Q., Lan X., and Zhang Y. Cross-lingual knowledge graph alignment via graph convolutional networks. In Riloff E., Chiang D., Hockenmaier J., and Tsujii J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 349–357, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1032. URL https://aclanthology.org/D18-1032/.

[108] Nath N., Lee S.-H., McDonnell M. D., and Lee I. The quest for better clinical word vectors: Ontology based and lexical vector augmentation versus clinical contextual embeddings. *Computers in Biology and Medicine*, 134:104433, July 2021. ISSN 00104825. doi: 10.1016/j.compbiomed.2021.104433. URL https://linkinghub.elsevier.com/retrieve/pii/S0010482521002274.

[109] Lin S., Hilton J., and Evans O. Teaching Models to Express Their Uncertainty in Words. *Transactions on Machine Learning Research*, June 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=8s8K2UZGTZ.

[110] Kugic A., Kreuzthaler M., and Schulz S. Clinical Acronym Disambiguation via ChatGPT and BING. *Studies in Health Technology and Informatics*, 309:78–82, October 2023. ISSN 1879-8365. doi: 10.3233/SHTI230743.

[111] Kugic A., Schulz S., and Kreuzthaler M. Disambiguation of acronyms in clinical narratives with large language models. *Journal of the American Medical Informatics Association*, 31(9):2040–2046, September 2024. ISSN 1527-974X. doi: 10.1093/jamia/ocae157. URL https://doi.org/10.1093/jamia/ocae157.

[112] Dreisbach C., Koleck T. A., Bourne P. E., and Bakken S. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *International journal of medical informatics*, 125:37,

February 2019. doi: 10.1016/j.ijmedinf.2019.02.008. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6438188/.

[113] Menaha R. and Jayanthi V. A Survey on Acronym–Expansion Mining Approaches from Text and Web. In Satapathy S. C., Bhateja V., and Das S., editors, *Smart Intelligent Computing and Applications*, pages 121–133, Singapore, 2019. Springer. ISBN 9789811319211. doi: 10.1007/978-981-13-1921-1_12.

[114] Link N. B., Huang S., Cai T., Sun J., Dahal K., Costa L., Cho K., Liao K., Cai T., and Hong C. Binary acronym disambiguation in clinical notes from electronic health records with an application in computational phenotyping. *International Journal of Medical Informatics*, 162:104753, June 2022. ISSN 1386-5056. doi: 10.1016/j.ijmedinf.2022.104753. URL https://www.sciencedirect.com/science/article/pii/S1386505622000673.

[115] Liu J., Wang C., and Liu S. Utility of ChatGPT in Clinical Practice. *Journal of Medical Internet Research*, 25(1):e48568, June 2023. doi: 10.2196/48568. URL https://www.jmir.org/2023/1/e48568. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[116] Ouyang L., Wu J., Jiang X., Almeida D., Wainwright C. L., Mishkin P., Zhang C., Agarwal S., Slama K., Ray A., Schulman J., Hilton J., Kelton F., Miller L., Simens M., Askell A., Welinder P., Christiano P., Leike J., and Lowe R. Training language models to follow instructions with human feedback, March 2022. URL http://arxiv.org/abs/2203.02155. arXiv:2203.02155.

[117] Liu Y., Melton G. B., and Zhang R. Exploring Large Language Models for Acronym, Symbol Sense Disambiguation, and Semantic Similarity and Relatedness Assessment. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2024:324–333, 2024. ISSN 2153-4063.

[118] Morelli V. Social Determinants of Health: An Overview for the Primary Care Provider. *Primary Care*, 50(4):507–525, December 2023. ISSN 1558-299X. doi: 10.1016/j.pop.2023.04.004.

[119] Organization W. H. Social determinants of health, 2024. URL https://www.who.int/health-topics/social-determinants-of-health.

[120] Lybarger K., Dobbins N. J., Long R., Singh A., Wedgeworth P., and others . Leveraging natural language processing to augment structured social determinants of health data in the electronic health record. *Journal of the American Medical Informatics Association*, 30(8):1389–1397, August 2023. ISSN 1527-974X. doi: 10.1093/jamia/ocad073. URL https://doi.org/10.1093/jamia/ocad073.

[121] Lybarger K., Yetisgen M., and Uzuner Ö. The 2022 n2c2/UW shared task on extracting social determinants of health. *Journal of the American Medical Informatics Association: JAMIA*, 30(8):1367–1378, July 2023. ISSN 1527-974X. doi: 10.1093/jamia/ocad012.

[122] Patra B. G., Sharma M. M., Vekaria V., Adekkanattu P., Patterson O. V., Glicksberg B., Lepow L. A., Ryu E., Biernacka J. M., Furmanchuk A., George T. J., Hogan W., Wu Y., Yang X., Bian J., Weissman M., Wickramaratne P., Mann J. J., Olfson M., Campion T. R., Weiner M., and Pathak J. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *Journal of the American Medical Informatics Association: JAMIA*, 28(12):2716–2727, November 2021. ISSN 1527-974X. doi: 10.1093/jamia/ocab170.

[123] Kugic A., Potjan L. M., Hammer L. M., Schulz S., and Kreuzthaler M. Alcohol Status Standardization from Clinical Real World Data with Transformer Architectures. In *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*, pages 233–238, June 2022. doi: 10.1109/ICHI54592.2022.00043. ISSN: 2575-2634.

[124] Kugic A., Abdulnazar A., Knezovic A., Schulz S., and Kreuzthaler M. Smoking Status Classification: A Comparative Analysis of Machine Learning Techniques with Clinical Real World Data. In Finkelstein J., Moskovitch R., and Parimbelli E., editors, *Artificial Intelligence in Medicine*, volume 14844, pages 182–191. Springer Nature Switzerland, Cham, 2024. ISBN 978-3-031-66537-0 978-3-031-66538-7. doi: 10.1007/978-3-031-66538-7_19. URL https://link.springer.com/10.1007/978-3-031-66538-7_19. Series Title: Lecture Notes in Computer Science.

[125] Alzoubi H., Ramzan N., Alzubi R., and Mesbahi E. An Automated System for Identifying Alcohol Use Status from Clinical Text. In *2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, pages 41–46, Southend, United Kingdom, August 2018. IEEE. ISBN 978-1-5386-4904-6. doi: 10.1109/iCCECOME.2018.8658578. URL https://ieeexplore.ieee.org/document/8658578/.

[126] Johnson A. E. W., Pollard T. J., Shen L., Lehman L.-w. H., Feng M., Ghassemi M., Moody B., Szolovits P., Anthony Celi L., and Mark R. G. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035, May 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35. URL https://www.nature.com/articles/sdata201635. Publisher: Nature Publishing Group.

[127] Lix L., Munakala S. N., and Singer A. Automated Classification of Alcohol Use by Text Mining of Electronic Medical Records. *Online Journal of Public Health Informatics*, 9(1), May 2017. ISSN 1947-2579. doi: 10.5210/ojphi.v9i1.7648. URL http://journals.uic.edu/ojs/index.php/ojphi/article/view/7648.

[128] Topaz M., Murga L., Bar-Bachar O., Cato K., and Collins S. Extracting Alcohol and Substance Abuse Status from Clinical Notes: The Added Value of Nursing Data. *MEDINFO 2019: Health and Wellbeing e-Networks for All*, pages 1056–1060, 2019. doi: 10.3233/SHTI190386. URL https://ebooks.iospress.nl/doi/10.3233/SHTI190386. Publisher: IOS Press.

[129] Saeed M., Villarroel M., Reisner A. T., Clifford G., Lehman L.-W., Moody G., Heldt T., Kyaw T. H., Moody B., and Mark R. G. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. *Critical care medicine*, 39(5):952–960, May 2011. ISSN 0090-3493. doi: 10.1097/CCM.0b013e31820a92c6. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3124312/.

[130] LeCun Y., Boser B., Denker J. S., Henderson D., Howard R. E., Hubbard W., and Jackel L. D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, December 1989. ISSN 0899-7667. doi: 10.1162/neco.1989.1.4.541. URL https://ieeexplore.ieee.org/abstract/document/6795724. Conference Name: Neural Computation.

[131] Lecun Y., Bottou L., Bengio Y., and Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. ISSN 00189219. doi: 10.1109/5.726791. URL http://ieeexplore.ieee.org/document/726791/.

[132] Grave E., Bojanowski P., Gupta P., Joulin A., and Mikolov T. Learning word vectors for 157 languages. In Calzolari N., Choukri K., Cieri C., Declerck T., Goggi S., Hasida K., Isahara H., Maegaard B., Mariani J., Mazo H., Moreno A., Odijk J., Piperidis S., and Tokunaga T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1550.

[133] Lu W., Aarsand R., Rylance S., Schotte K., Han J., Lebedeva E., Tsoy E., Bill W., Halpin D., Rivera P. M., Fong K., Kathuria H., Gappa M., Lam D. C., Tursan D'Espaignet E., and Sohal S. S. *Tobacco and chronic obstructive pulmonary disease (COPD)*. World Health Organization, November 2023. ISBN 978-92-4-008445-2. URL https://www.who.int/publications-detail-redirect/9789240084452.

[134] World Health Organization , editor. *European health report 2018: more than numbers - evidence for all*. Regional Office for Europe, Copenhague, 2018. ISBN 978-92-890-5343-3. URL https://iris.who.int/handle/10665/279904.

[135] WHO . *The European Health Report 2021. Taking stock of the health-related Sustainable Development Goals in the COVID-19 era with a focus on leaving no one behind*. World Health Organization, March 2022. ISBN 978-92-890-5754-7. URL https://www.who.int/europe/publications/i/item/9789289057547.

[136] Kukhareva P. V., Caverly T. J., Li H., Katki H. A., Cheung L. C., Reese T. J., Del Fiol G., Hess R., Wetter D. W., Zhang Y., Taft T. Y., Flynn M. C., and Kawamoto K. Inaccuracies in electronic health records smoking data and a potential approach to address resulting underestimation in determining lung cancer screening eligibility. *Journal of the American Medical Informatics Association*, 29(5):779–788, April 2022. ISSN 1527-974X. doi: 10.1093/jamia/ocac020. URL https://academic.oup.com/jamia/article/29/5/779/6529026.

[137] Yang X., Yang H., Lyu T., Yang S., Guo Y., Bian J., Xu H., and Wu Y. A Natural Language Processing Tool to Extract Quantitative Smoking Status from Clinical Narratives. In *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–2, November 2020. doi: 10.1109/ICHI48887.2020.9374369. URL https://ieeexplore.ieee.org/document/9374369. ISSN: 2575-2634.

[138] Ruckdeschel J. C., Riley M., Parsatharathy S., Chamarthi R., Rajagopal C., Hsu H. S., Mangold D., and Driscoll C. Unstructured Data Are Superior to Structured Data for Eliciting Quantitative Smoking History From the Electronic Health Record. *JCO clinical cancer informatics*, 7:e2200155, February 2023. ISSN 2473-4276. doi: 10.1200/CCI.22.00155.

[139] Peng Y., Yan S., and Lu Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In Demner-Fushman D., Cohen K. B., Ananiadou S., and Tsujii J., editors, *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5006. URL https://aclanthology.org/W19-5006.

[140] Bae Y. S., Kim K. H., Kim H. K., Choi S. W., Ko T., Seo H. H., Lee H.-Y., and Jeon H. Keyword Extraction Algorithm for Classifying Smoking Status from Unstructured Bilingual Electronic Health Records Based on Natural Language Processing. *Applied Sciences*, 11(19):8812, January 2021. ISSN 2076-3417. doi: 10.3390/app11198812. URL https://www.mdpi.com/2076-3417/11/19/8812. Number: 19 Publisher: Multidisciplinary Digital Publishing Institute.

[141] Bressem K. K., Papaioannou J.-M., Grundmann P., Borchert F., Adams L. C., Liu L., Busch F., Xu L., Loyen J. P., Niehues S. M., Augustin M., Grosser L., Makowski M. R., Aerts H. J., and Löser A. medbert.de: A comprehensive german bert model for the medical domain. *Expert Systems with Applications*, 237:121598, 2024. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2023.121598. URL https://www.sciencedirect.com/science/article/pii/S0957417423021000.

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

[143] Schwarz C. M., Hoffmann M., Smolle C., Eiber M., Stoiser B., Pregartner G., Kamolz L., and Sendlhofer G. Structure, content, unsafe abbreviations, and completeness of discharge summaries: A retrospective analysis in a University Hospital in Austria. *Journal of Evaluation in Clinical Practice*, 27(6):1243–1251, December 2021. ISSN 1356-1294, 1365-2753. doi: 10.1111/jep.13533. URL https://onlinelibrary.wiley.com/doi/10.1111/jep.13533.

[144] Leaman R., Khare R., and Lu Z. Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics*, 57:28–37, October 2015. ISSN 1532-0464. doi: 10.1016/j.jbi.2015.07.010. URL https://www.sciencedirect.com/science/article/pii/S1532046415001501.

[145] Tian S., Jin Q., Yeganova L., Lai P.-T., Zhu Q., Chen X., Yang Y., Chen Q., Kim W., Comeau D. C., Islamaj R., Kapoor A., Gao X., and Lu Z. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493, 01 2024. ISSN 1477-4054. doi: 10.1093/bib/bbad493. URL https://doi.org/10.1093/bib/bbad493.

[146] Huang L., Yu W., Ma W., Zhong W., Feng Z., Wang H., Chen Q., Peng W., Feng X., Qin B., and Liu T. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55, January 2025. ISSN 1046-8188. doi: 10.1145/3703155. URL https://dl.acm.org/doi/10.1145/3703155.

[147] Jonnagaddala J. and Wong Z. S.-Y. Privacy preserving strategies for electronic health records in the era of large language models. *npj Digital Medicine*, 8(1): 1–3, January 2025. ISSN 2398-6352. doi: 10.1038/s41746-025-01429-0. URL https://www.nature.com/articles/s41746-025-01429-0. Publisher: Nature Publishing Group.

[148] Yang Y., Lin M., Zhao H., Peng Y., Huang F., and Lu Z. A survey of recent methods for addressing AI fairness and bias in biomedicine. *Journal of Biomedical Informatics*, 154:104646, June 2024. ISSN 1532-0464. doi: 10.1016/j.jbi.2024.104646. URL https://www.sciencedirect.com/science/article/pii/S1532046424000649.

[149] Brender T. D., Celi L. A., and Cobert J. M. Clinical Notes as Narratives: Implications for Large Language Models in Healthcare. *Journal of General Internal*

Medicine, October 2024. ISSN 1525-1497. doi: 10.1007/s11606-024-09093-y. URL https://doi.org/10.1007/s11606-024-09093-y.

[150] Shah K., Xu A. Y., Sharma Y., Daher M., McDonald C., Diebo B. G., and Daniels A. H. Large Language Model Prompting Techniques for Advancement in Clinical Medicine. *Journal of Clinical Medicine*, 13(17):5101, January 2024. ISSN 2077-0383. doi: 10.3390/jcm13175101. URL https://www.mdpi.com/2077-0383/13/17/5101. Number: 17 Publisher: Multidisciplinary Digital Publishing Institute.

[151] Laparra E., Mascio A., Velupillai S., and Miller T. A Review of Recent Work in Transfer Learning and Domain Adaptation for Natural Language Processing of Electronic Health Records. *Yearbook of Medical Informatics*, 30(1):239–244, August 2021. ISSN 0943-4747, 2364-0502. doi: 10.1055/s-0041-1726522. URL http://www.thieme-connect.de/DOI/DOI?10.1055/s-0041-1726522. Publisher: Georg Thieme Verlag KG.