Dissertation

Computational Semantics for Intelligent Digital Health Applications

submitted by

Akhila Naz Kuppassery Abdulnazar

for the Academic Degree of

Doctor of Philosophy
(PhD)

at the

Medical University of Graz

Institute for Medical Informatics, Statistics and Documentation
CBmed - Center for Biomarker Research in Medicine

under the Supervision of

Markus Eduard Kreuzthaler
Ass.-Prof. Priv.-Doz. Dipl.-Ing. Dr.scient.med.

2025

Declaration of Academic Integrity

I hereby confirm that the present 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, 22nd April 2025 Akhila Naz Kuppassery Abdulnazar

Disclosure

Parts of the dissertation are already published and contain the literal text of:

(1) **Abdulnazar A**[1,2], Kugic A[1], Schulz S[1], Stadlbauer V[2,3], Kreuzthaler M[1]. O2 supplementation disambiguation in clinical narratives to support retrospective COVID-19 studies. BMC Med Inform Decis Mak 2024;24:29. https://doi.org/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.

Markus Kreuzthaler supervised the workflow, provided feedback, and corrected the final draft. Vanessa Stadlbauer-Köllner triggered the problem motivation. Amila Kugic and Stefan Schulz annotated the dataset and corrected the final draft. The manuscript is published under the terms of the Creative Commons Open-Access `CC BY 4.0` licence.

(2) **Abdulnazar A**[1,2], Roller R[3], Schulz S[1], Kreuzthaler M[1]. Unsupervised SapBERT-based bi-encoders for medical concept annotation of clinical narratives with SNOMED CT. DIGITAL HEALTH 2024. https://doi.org/10:20552076241288681,2024.

[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]German Research Center for Artificial Intelligence (DFKI), Berlin, Germany.

Markus Kreuzthaler and Roland Roller supervised the methodology workflow, provided feedback, and corrected the final draft. Stefan Schulz annotated the dataset and corrected the final draft. The manuscript is published under the terms of the Creative Commons Open-Access `CC BY-NC-ND 4.0` licence.

③ **Abdulnazar A**[1,2], Roller R[3], Schulz S[1], Kreuzthaler M[1]. Large language models for clinical text cleansing enhance medical concept normalization. IEEE Access 2024;12:147981–90. https://doi.org/10.1109/ACCESS.2024.3472500.

[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]German Research Center for Artificial Intelligence (DFKI), Berlin, Germany.

Markus Kreuzthaler and Roland Roller supervised the methodology workflow, provided feedback, and corrected the final draft. Stefan Schulz triggered the problem motivation, annotated the dataset, and corrected the final draft. The manuscript is published under the terms of the Creative Commons Open-Access `CC BY-NC-ND 4.0` licence.

In the role as the creator and original author of these publications, permission for their inclusion in this dissertation is granted under the terms of the Creative Commons Open-Access licences `CC BY 4.0` and `CC BY-NC-ND 4.0`. All co-authors have granted their approval for the inclusion of these works in this dissertation.

Additional Publications

Veeranki S, **Abdulnazar A**, Kramer D, Kreuzthaler M, Lumenta D. Multi-label text classification via secondary use of large clinical real-world data sets. Sci Rep 2024;14:26972. https://doi.org/10.1038/s41598-024-76424-8.

Abdulnazar A, Kreuzthaler M, Schulz S, Prietl B, Herbsthofer L. Tumor board visualization: Integrating clinical and laboratory insights. Stud Health Technol Inform 2024;316:1750–1. https://doi.org/10.3233/SHTI240767.

Kugic A, **Abdulnazar A**, Knezovic A, Schulz S, Kreuzthaler M. Smoking status classification: A comparative analysis of machine learning techniques with clinical real world data. Lecture notes in computer science 2024;14844:182–91.

Abdulnazar A, Kreuzthaler M, Roller R, Schulz S. SapBERT-based medical concept normalization using SNOMED CT. Stud Health Technol Inform 2023;302:825–6. https://doi.org/10.3233/SHTI230278.

Schulz S, **Abdulnazar A**, Kreuzthaler M. Clustering similar diagnosis terms. Stud Health Technol Inform 2023;302:837–8. https://doi.org/10.3233/SHTI230284.

Abdulnazar A, Schulz S, Kreuzthaler M. Smoking status normalization with cross-encoders and SNOMED CT. Stud Health Technol Inform 2025. Accepted for publication.

Use of AI in this Dissertation

Between November 2024 and April 2025, the language optimization process utilized AI tools, including ChatGPT (version 4o, OpenAI) and DeepSeek Chat (DeepSeek). More details about the tool can be found at the following URL: ChatGPT (https://openai.com/chatgpt), DeepSeek (https://www.deepseek.com).

Acknowledgement

First and foremost, I am deeply grateful for the strength and guidance that have supported me throughout this journey. This invaluable source of inspiration and resilience has been pivotal in making this work possible.

Next, I would like to express my sincere appreciation to my supervisors, Markus Kreuzthaler and Stefan Schulz. Your unwavering support, wisdom, and patience have been the foundation of my academic and personal growth. Thank you for believing in me and my abilities, encouraging me to pursue this journey, and providing unwavering motivation and support.

My heartfelt gratitude also goes to Roland Roller, from the German Research Center for Artificial Intelligence (DFKI). Collaborating with you has been an incredibly enriching experience. Your expertise, encouragement, and guidance have profoundly shaped the course of my research. Thank you so much for your support, especially during my research stay and the preparation of my dissertation. Your investment of time and energy has been invaluable.

I would also like to sincerely thank my thesis committee members, Rada Hussein, and Martin Boeker. Your insightful feedback and constructive criticism were invaluable in refining my work. Your thoughtful suggestions pushed me to see my research from new perspectives and helped me strengthen this dissertation.

I sincerely thank CBmed GmbH for providing the funding and unwavering support that made this research possible. A heartfelt thanks also goes to the team within the Institute of Medical Informatics, Statistics, and Documentation at the Medical University of Graz, whose collaborative efforts provided invaluable guidance and the data that formed the foundation of this work. I sincerely thank the AMBRA PhD School at the Medical University of Graz; the opportunities and support provided by this program have been

instrumental in making this dissertation possible. To all my colleagues, thank you for the countless discussions and shared ideas, and for being an essential part of this journey. Your expertise and commitment have been integral to the success of this project, and I am deeply grateful for your support. Your camaraderie made the long days shorter and the tough times easier.

To my family, words will never be enough to express my gratitude. To my mother and father, your endless love, encouragement, and sacrifices have been the fuel that kept me going. Your support and belief in me have kept me going through this journey. To my brother, thank you for always being my rock and for the constant support and laughter that kept me grounded. And to my dearest, thanks for your support and for making this final stretch so joyful.

To my friends, thank you for being my support system. You always knew when I needed a break, a laugh, a listening ear, or a memorable trip to recharge and create cherished moments. Your companionship has lightened the load of this journey and made the difficult times easier to bear.

This thesis is dedicated to all of you for your love, support, and belief in me.

Graz, 22nd April 2025 Akhila Naz Kuppassery Abdulnazar

Contents

1 Introduction **1**

1.1 Electronic Health Records 2

1.1.1 Challenges 4

1.1.2 International Terminologies 6

1.2 Machine Learning for Text Processing 8

1.2.1 Classical Machine Learning Methods 8

1.2.2 Deep Learning Methods 10

1.2.3 Large Language Models 12

1.3 Key Tasks in Text Processing 14

1.3.1 Named Entity Recognition 14

1.3.2 Medical Concept Normalization 14

1.3.3 Relation Extraction 15

1.4 Derived Dissertation Objectives 16

2 Investigations **18**

2.1 O2 Supplementation Disambiguation in Clinical Narratives to Support Retrospective COVID-19 Studies 18

2.1.1 Summary 18

2.1.2 Discussion 20
2.1.3 Contribution 21
2.2 Unsupervised SapBERT-based Bi-Encoders for Medical Concept Annotation of Clinical Narratives with SNOMED CT 22
2.2.1 Summary 22
2.2.2 Discussion 24
2.2.3 Contribution 26
2.3 Large Language Models for Clinical Text Cleansing Enhance Medical Concept Normalization 26
2.3.1 Summary 26
2.3.2 Discussion 28
2.3.3 Contribution 31
2.4 Gain of Knowledge 31
2.4.1 Summarized Contribution 31
2.4.2 Supportive Investigations 32

3 Conclusion and Outlook **36**

Bibliography **38**

Publications **47**

P.1 O2 Supplementation Disambiguation in Clinical Narratives to Support Retrospective COVID-19 Studies 48
P.2 Unsupervised SapBERT-based Bi-Encoders for Medical Concept Annotation of Clinical Narratives with SNOMED CT 61
P.3 Large Language Models for Clinical Text Cleansing Enhance Medical Concept Normalization 73

Abbreviations

A

AI Artificial Intelligence.

B

BERT Bidirectional Encoder Representations from Transformer.
Bi-LSTM Bidirectional Long Short-Term Memory Network.

C

CNN Convolutional Neural Network.
COVID-19 Coronavirus Disease 2019.
CRF Conditional Random Fields.

E

EHR Electronic Health Record.

F

Faiss Facebook AI Similarity Search.
FHIR Fast Healthcare Interoperability Resources.

G

GDPR General Data Protection Regulation.
GPT Generative Pretrained Transformer.

H

HIPAA Health Insurance Portability and Accountability Act.

I

ICD-9 International Classification of Diseases - Revision Nine.
ICD-10 International Classification of Diseases - Revision Ten.

K

k-NN k-Nearest Neighbor.

L

LIME Local Interpretable Model-Agnostic Explanations.
LLaMA Large Language Model Meta AI.
LLM Large Language Model.
LOINC Logical Observation Identifiers Names and Codes.
LSTM Long Short-Term Memory.

M

MCA Medical Concept Annotation.
MCN Medical Concept Normalization.
MIMIC Medical Information Mart for Intensive Care.
ML Machine Learning.

N

NER Named Entity Recognition.
NLP Natural Language Processing.

O

OMOP Observational Medical Outcomes Partnership.

P

PaLM Pathways Language Model.
PCA Principal Component Analysis.

R

RAG Retrieval Augmented Generation.

RNN Recurrent Neural Network.

S

SapBERT Self-Alignment Pretraining Bidirectional Encoder Representations from Transformers.
SVM Support Vector Machine.

T

T5 Text-to-Text Transfer Transformer.
TF-IDF Term Frequency-Inverse Document Frequency.
t-SNE t-distributed Stochastic Neighbor Embedding.

U

UMLS Unified Medical Language System.

X

XAI Explainable Artificial Intelligence.

List of Figures

1.1 t-SNE visualization of UMLS entity embeddings: PubMedBERT (left) shows overlap, while PubMedBERT plus SapBERT (right) forms compact clusters after self-alignment [52]. 12

2.1 Graphical abstract summarizing the proposed methodology, including text line preprocessing, binary text classification, vector clustering using t-SNE visualization, error analysis, and explainable AI interpretations [77]. 19

2.2 Graphical abstract illustrating the proposed framework: embedding space creation for SNOMED CT, clinical data preparation, n-gram generation with entity recognition and normalization, and best n-gram selection and evaluation [78]. 23

2.3 Graphical abstract summarizing the methodology: embedding space generation from two German terminologies, clinical narrative cleansing with LLM and error analysis, and MCN via dictionary lookup, bi-encoders, and RAG followed by error analysis [79]. 27

2.4 Screenshot of the graphical user interface, illustrating the surgery report coding support system with text input and label selection features [85]. . . 34

2.5 FusionViewer patient screen: Sidebar for selecting patient data (clinical, laboratory, fusion, graphical, sample quality). The main view shows the patient profile with a timeline and key navigation features. 35

Zusammenfassung

Einleitung. Die Einführung elektronischer Gesundheitsakten (EHRs) hat das Gesundheitswesen durch die Digitalisierung von Patientendaten und die Verbesserung ihrer Zugänglichkeit und Wiederverwertbarkeit revolutioniert. Die Interoperabilität stellt jedoch nach wie vor eine große Herausforderung dar, da unterschiedliche Terminologien, Datenformate und Systemarchitekturen in verschiedenen Einrichtungen einen nahtlosen Datenaustausch behindern. Dieser Mangel an Standardisierung schränkt das Potenzial intelligenter digitaler Gesundheitsanwendungen ein. In dieser Dissertation werden Lösungen zur Verbesserung der Interoperabilität von elektronischen Patientenakten erforscht, indem klassische und Deep-Learning-basierte Textklassifizierungsmethoden angewandt, die Erkennung von benannten Entitäten (NER) und die Normalisierung medizinischer Konzepte (MCN) mithilfe unüberwachter Lernmethoden automatisiert und Transformermodelle und große Sprachmodelle (LLMs) für die Zuordnung klinischer Texte zu Terminologien wie SNOMED CT genutzt werden.

Methoden. Diese Dissertation ist in drei Hauptstudien unterteilt, die sich auf die Verbesserung der Interoperabilität von EHRs konzentrieren. Die erste Studie wendet klassisches maschinelles Lernen (ML) und Deep-Learning-Modelle an, um EHR-Daten im Zusammenhang mit Sauerstoffsupplementierung zu klassifizieren, wobei die Wirksamkeit von ML-Methoden bei der Organisation großer Datenmengen, insbesondere für die COVID-19-Forschung, hervorgehoben wird. Die zweite Studie automatisiert NER und MCN mit unüberwachten Methoden. Es wurde eine Pipeline entwickelt, die klinische Texte tokenisiert, Entitäten erkennt und sie auf SNOMED CT abbildet, wobei SapBERT für die Einbettung medizinischer Begriffe und regelbasiertes Re-Ranking für die Disambiguierung von Entitäten verwendet wird. Dieser Ansatz reduziert die Abhängigkeit von manuell annotierten Daten. In der dritten Studie werden SapBERT und LLMs kombiniert, um MCN zu verbessern und exakte Zuordnungen zu SNOMED CT und UMLS zu ermöglichen. Auf LLMs basierende Algorithmen zur Datenbereinigung und für das Re-Ranking verbesserten die Genauigkeit. Weitere Untersuchungen umfassen SapBERT-basiertes Termclustering,

einen Vergleich von kontextuellen und nicht-kontextuellen Vektorrepräsentationen, hybride Ansätze für die Abbildung des Raucherstatus und die Entwicklung von Explainable AI (XAI) und Visualisierungstools für die Integration und Navigation von Patientendaten.

Ergebnisse. In der ersten der drei genannten Hauptstudien erreichte das Textklassifizierungsmodell einen F1 Score von über 90% bei der Kategorisierung von Aufzeichnungen zur Sauerstoffsupplementierung und bewies damit die Wirksamkeit des gewählten maschinellen Lernansatzes für die domänen- und taskspezifische Aufgabe. Die unüberwachten Lernmethoden der zweiten Studie wiesen eine vielversprechende Performance bezüglich Precision und Recall auf, wodurch die Abhängigkeit von manuell annotierten Daten für die Aufgabe der Termnormalisierung in Zukunft erheblich verringert werden kann. Die dritte Studie bestätigte, dass BERT-basierte Modelle einem traditionellen Lexikonabgleich für die Aufgabe der MCN überlegen sind, mit einer Verbesserung der Erkennungsrate um 91,8% von einem F1 Score von 0.297 auf 0.568. Darüber hinaus verbesserte die Anwendung von LLMs für die Datenbereinigung und das Re-Ranking die Leistung von BERT um 6,8% im F1 Score in der Aufgabenstellung, wobei der Normalisierungsprozess verfeinert und die Anpassung an standardisierte medizinische Terminologien verbessert wurde.

Diskussion. Die kollektiven Ergebnisse dieser Arbeit unterstreichen die entscheidende Rolle von fortgeschrittenen kontextuellen maschinellen Lernmethoden und der Anwendung von LLMs bei der Unterstützung der Interoperabilität von Daten in der elektronischen Patientenakte. Durch die Verbesserung der Genauigkeit, Konsistenz und Automatisierung von MCN trägt diese Forschung zur Entwicklung einer standardisierten Repräsentation von Gesundheitsdaten bezüglich internationaler Terminologien, im speziellen SNOMED CT dar. Diese Fortschritte ermöglichen eine bessere Verwendung dieser Daten im Kontext intelligenter Gesundheitsanwendungen, die nahtlos Daten austauschen, klinische Arbeitsabläufe verbessern und die Patientenversorgung optimieren können. Diese Arbeit unterstreicht die Notwendigkeit kontinuierlicher Innovation in der Gesundheitsinformatik und bietet eine Grundlage für künftige Forschung zur Überbrückung der Interoperabilitätslücken in EHR-Systemen mit Hilfe der Erstellung strukturierter und standardisierter Patientenprofile.

Abstract

Introduction. The introduction of electronic health records (EHRs) has revolutionized healthcare by digitizing patient information and improving its accessibility and reuse. However, interoperability remains a significant challenge due to variations in terminologies, data formats, and system architecture across institutions, hindering seamless data exchange. This lack of standardization limits the potential of intelligent digital health applications. This dissertation explores solutions to enhance EHR interoperability by applying classical and deep learning-based text classification methods, automating named entity recognition (NER) and medical concept normalization (MCN) using unsupervised techniques, and leveraging transformer models and large language models (LLMs) for mapping clinical narratives to terminologies such as SNOMED CT.

Methods. This dissertation is divided into three main studies focused on enhancing EHR interoperability. The first study applies classical machine learning (ML) and deep learning models to classify EHR data related to oxygen supplementation, highlighting the effectiveness of ML methods in organizing large volumes of data, especially for COVID-19 research. The second study automates NER and MCN using unsupervised methods. A pipeline was developed to tokenize clinical narratives, detect entities, and map them to SNOMED CT, employing SapBERT for medical term embeddings and rule-based re-ranking for entity disambiguation. This approach reduces reliance on manually labelled data. The third study combines SapBERT and an LLM to enhance MCN, enabling exact mappings to SNOMED CT and UMLS. LLM-based data cleansing and re-ranking algorithms improved accuracy. Additional investigations include SapBERT-based term clustering, a comparison of contextual vs. non-contextual embeddings, hybrid approaches for mapping smoking status, and the development of Explainable AI (XAI) and visualization tools for patient data integration and navigation.

Results. In the first of the three main studies mentioned, the text classification model achieved an F1 score of over 90% in categorizing oxygen supplementation records, demonstrating the effectiveness of the chosen ML approach for the domain- and task-specific challenge. The unsupervised learning methods in the second study showed promising performance in terms of precision and recall, significantly reducing the dependency on

manually annotated data for this task in the future. The third study confirmed that BERT-based models outperformed traditional lexicon matching for the task of MCN, with a 91.8% improvement in detection rate, raising the F1 score from 0.297 to 0.568. Furthermore, the application of LLMs for data cleaning and reranking enhanced the performance of BERT by 6.8% in F1 score for the task, refining the normalization process and improving alignment with standardized medical terminologies.

Discussion. The collective results of this work underscore the critical role of advanced contextual ML methods and the application of LLMs in supporting data interoperability within EHRs. By improving the accuracy, consistency, and automation of MCN, this research contributes to the development of a standardized representation of health data in relation to international terminologies, particularly SNOMED CT. These advancements enable better utilization of such data in the context of intelligent healthcare applications, which can seamlessly exchange data, enhance clinical workflows, and optimize patient care. This work highlights the need for continuous innovation in health informatics. It provides a foundation for future research to bridge interoperability gaps in EHR systems through the creation of structured and standardized patient profiles.

Chapter 1

Introduction

The digital transformation in healthcare and clinical research has positioned electronic health records (EHRs) as a pivotal source of clinical real-world data and an essential focus for optimization [1]. Unlike the carefully structured datasets designed for clinical trials or administrative purposes, real-world data reverse to a multitude of clinical entities such as conditions, diagnoses, procedures, and outcomes documented during routine care processes. These data often take the form of semi-structured narratives, rich in specialized terminology and context, but pose challenges for effective utilization due to their variability and informality. The predominance of semi-structured and non-standardized narratives within EHRs present significant challenges, including difficulty in extracting, interpreting, and standardizing clinical information. While supporting communication between healthcare professionals, these free-text expressions are often concise, informal, and influenced by regional linguistic variations, limiting their interoperability and reuse. Addressing these issues requires converting unstructured clinical narratives into structured, standardized formats supporting comprehensive patient profiling, facilitating better clinical decision-making, enhancing reporting, improving documentation, and streamlining communication in healthcare [2].

A promising approach to address EHR limitations is the adoption of ontologies with computational semantics [3]. These ontologies enhance data quality and interoperability across healthcare systems, enabling more effective data exchange through standardized terminologies, such as SNOMED CT [4]. This approach supports the FAIR data principles [5] – making data Findable, Accessible, Interoperable, and Reusable – thereby

improving clinical decision-making, research, and quality improvement. Integrating a semantic layer beyond database schema definitions in EHRs will enable personalized interfaces, context-sensitive guidance, error detection, and enhanced usability.

In parallel, advances in machine learning (ML), a key branch of artificial intelligence (AI), have introduced powerful tools to unlock the potential of EHRs. Specifically, Natural Language Processing (NLP) techniques allow the automated extraction, classification, and normalization of medical concept mentions, transforming unstructured text into structured, actionable insights [6]. These methods improve the reusability of patient data and support the creation of personalized solutions in healthcare and research. Finally, visualization is crucial for making complex clinical data interpretable and actionable, enabling direct application to patient care [7].

Following this introduction, the dissertation is structured as follows to provide a detailed walkthrough of the field of research: Section 1.1 examines the complexities of EHRs, looking into the challenges they present and the role played by international terminologies. Against this backdrop, Section 1.2 presents the application of ML for text processing, covering both classical and deep learning methodologies, culminating in an overview of large language models (LLMs). Next, Section 1.3 discusses key tasks in text processing, including named entity recognition (NER), medical concept normalization (MCN), and relation extraction, which are crucial to extracting critical information from EHR text. Section 1.4 shows the research gaps that led to the research questions and objectives guiding the three key core investigations in Chapter 2. Finally, Chapter 3 concludes the dissertation by summarizing key contributions, discussing their implications for research and clinical practice, and outlining future directions.

1.1 Electronic Health Records

The concept of recording patient information has evolved over time. Initially, patient data were captured in physical reports, documents, and summaries, often in disparate formats such as handwritten notes, tables, and diagnostic reports. These documents, while valuable, posed challenges in terms of accessibility, storage, and integration. As a result, healthcare systems began digitizing these records to streamline processes and improve patient care. This led to the development of EHRs, which now serves as a comprehensive

digital repository of patient health information. EHRs contain real-time data, including clinical observations, diagnoses, treatments, laboratory results, medications, allergies, and medical histories, enabling more efficient clinical workflows and coordinated care [8].

The primary purposes of EHRs are to improve healthcare delivery through structured data management, workflow optimization, improved communication, and operational efficiency [9]. EHRs integrate administrative data, such as clinical coding for billing, with clinical information, including medical histories, laboratory results, and diagnostic records, to ensure accurate documentation and informed decision-making. By standardizing and structuring data, EHRs enable advanced techniques like NER and MCN to improve data precision and quality. They streamline workflows by automating order entry and efficiently managing results to improve safety and promote evidence-based practices by guiding clinicians through patient-specific data [10]. Communication is significantly enhanced, as EHRs facilitate seamless data sharing among providers, enabling interdisciplinary coordination for complex cases and supporting public health surveillance [11].

As healthcare practices evolve, secondary use case scenarios of EHRs have also expanded. With data sharing, interoperability has become a critical feature in modern healthcare. Interoperability refers to the ability of different healthcare systems and organizations to exchange patient data in a standardized format. By enabling data sharing, EHRs facilitate a more holistic approach to patient care, where information from multiple healthcare facilities can be consolidated to provide a complete view of a patient's medical history [12]. This improves patient care and data reuse, contributing to population health management and the development of personalized treatment plans. In this context, secondary use cases of EHRs have emerged, particularly in research and education. They support patient education, enable telehealth services, and improve administrative processes by validating insurance eligibility, reducing delays, and streamlining communication, such as drug recalls and chronic disease management programs. For example, cohort building for research studies and the training of ML models using de-identified data from diverse hospitals support advances in clinical knowledge and treatment protocols [13]. Additionally, EHRs facilitate knowledge acquisition, improve educational practices, and enhance the use of real-world data in clinical trials.

However, even with these developments, the full potential of EHRs has not yet been put into practice. Although progress has been made in automating data retrieval and decision

support systems, these functionalities are still primarily at the research level. As a result, new primary purposes are emerging for EHRs, including automated decision support, summarization, speech and writing assistance, and visualization. Operationally, they optimize administrative processes by automating insurance validation, drug recall notifications, and chronic disease management. At the same time, standardized data formats simplify reporting and population health management, reducing costs and enhancing public health initiatives. Together, these functions improve the quality, precision, and accessibility of healthcare services [14]. The integration of AI and ML into EHRs has made it possible to automate clinical decision-making processes by recommending treatments, predicting patient outcomes, and identifying potential complications early. AI-driven systems also offer support for summarizing clinical information, assisting healthcare professionals in understanding patient data more quickly and effectively. Moreover, visualization tools have begun to play a crucial role in improving the interpretation of patient data, allowing clinicians to see trends, relationships, and critical information more clearly [15]. Finally, data sharing between hospitals and healthcare systems has also paved the way for creating personalized mobile health records, where a patient's health information is continuously updated and shared across platforms, enabling more precise and individualized care [16]. Integrating AI and semantic resources into EHRs holds significant promise for improving patient outcomes, advancing healthcare research, and providing personalized, evidence-based care.

1.1.1 Challenges

Despite notable advances in EHR systems, numerous limitations constrain their full potential [17]. These challenges are multifaceted, encompassing both technical and organizational dimensions. Chief among them is the persistent issue of interoperability, which remains a critical barrier to the seamless exchange of health information across institutions. EHR systems often lack technical interoperability—the ability of disparate systems to communicate—and semantic interoperability, ensuring that exchanged data retains its intended meaning across different contexts and platforms [18].

The lack of standardized terminologies and data representations is a primary barrier to semantic interoperability in EHR systems. Localized coding practices, such as hospital-specific abbreviations and non-standard nomenclature, impede consistent data interpreta-

tion across institutions. This challenge is exacerbated by the proprietary and fragmented nature of dominant EHR platforms, which rely on closed architectures that often require custom preprocessing to enable inter-system data exchange. These factors hinder seamless information flow, compromise continuity of care, and obstruct efforts toward integrated, patient-centered healthcare delivery. Although initiatives promoting universal coding standards and interoperability frameworks are underway, progress has been slow. As a result, current EHRs remain limited in their capacity to support meaningful, system-wide data interoperability [18].

In addition to interoperability, data quality remains a critical challenge in EHR systems. Inconsistent documentation—such as non-standard abbreviations, typographical errors, and unstructured clinical notes—complicates clinical interpretation and NLP tasks. Moreover, missing or incomplete data, including uncoded diagnoses and partial lab results, compromise record accuracy and reliability. Temporal discrepancies, such as conflicting timestamps or outdated entries, further hinder longitudinal analyses and continuity of care. These issues collectively reduce the utility of EHR data for clinical and secondary uses, emphasizing the need for high-quality, consistent, and temporally accurate data across healthcare settings [19].

Beyond technical barriers, security and privacy remain critical concerns in EHR systems due to the sensitivity of patient data. Breaches can lead to identity theft and loss of patient trust, compromising care delivery. Although compliance with regulations such as the Health Insurance Portability and Accountability Act (HIPAA)[1] and the General Data Protection Regulation (GDPR)[2] is essential, it is often costly and complex, especially for smaller providers. Balancing data utility with regulatory requirements for de-identification is particularly challenging. Additionally, institutional silos and strict consent protocols hinder data sharing, limiting the availability of comprehensive datasets. Cyber threats necessitate robust safeguards, including encryption, multifactor authentication, and regular audits [19].

A significant challenge in EHR adoption is the workflow disruption it causes. Healthcare professionals transitioning from paper-based systems require extensive training, which can temporarily affect patient care. The added time for data entry, particularly in high-pressure settings, increases administrative burden and reduces face-to-face patient inter-

[1] https://www.hhs.gov/hipaa/index.html
[2] https://gdpr-info.eu/

actions, potentially lowering care quality and clinician satisfaction. Additionally, poorly designed user interfaces can lead to inefficiencies and frustration, compromising patient safety and system effectiveness [20]. Resistance to change, driven by concerns over workflow disruptions and increased workload, further complicates adoption. Inadequate training and alert fatigue worsen inefficiencies. Addressing these issues requires technological improvements, cultural shifts, better system design, ongoing training, and seamless workflow integration. A comprehensive approach encompassing technological, financial, and cultural support is essential to maximize EHR benefits and enhance patient outcomes [21].

Clinical contextualization remains a key challenge in leveraging EHR data, primarily due to ambiguous clinical narratives. Terms with multiple meanings hinder accurate concept extraction, while noisy real-world data, such as redundant or copy-pasted notes, introduce bias and undermine analytical reliability [22].

Despite existing challenges, ML holds significant promise for transforming EHR systems by enhancing decision-making, enabling early disease detection, and facilitating personalized care. However, its integration is impeded by several critical barriers, including the requirement for high-quality datasets, compatibility with clinical workflows, mitigation of algorithmic bias, and the necessity for rigorous validation. Additionally, the substantial implementation costs, particularly those associated with predictive analytics, contribute to adoption disparities, disproportionately affecting smaller healthcare providers [23].

1.1.2 International Terminologies

As highlighted in the previous section, the interoperability, data quality, and security challenges inherent in EHR systems underscore the critical need for standardized terminologies to enable efficient communication and data exchange between different healthcare settings. Adopting and implementing international terminology standards, such as SNOMED CT, LOINC, and ICD-10, can effectively address these key challenges, facilitating the seamless integration of clinical data across disparate systems and ultimately improving patient care. Interoperability, achieved through standardized terminologies, is crucial for promoting care coordination, reducing medical errors, improving patient outcomes, and supporting secondary use of data for research and quality improvement [24].

Among these standards, SNOMED CT is pivotal in enabling uniformity in clinical documentation. This standardized, multilingual clinical terminology includes over 350,000 representational units (SNOMED concepts) for diseases, procedures, medications, organisms, substances, drugs, devices, body parts, etc. Via the Unified Medical Language System (UMLS), which integrates approximately 200 biomedical terminologies, SNOMED CT can be mapped to other coding systems such as ICD-9, ICD-10, ensuring broader interoperability [25].

Complementing these are information models, such as Fast Healthcare Interoperability Resources (FHIR) [26], which provides a modern, modular framework for health data exchange, promoting interoperability through standardized "resources" that represent information about patients, medications, procedures, and other objects and processes. FHIR resources easily integrate with both legacy systems and new applications, ensuring efficient, secure data exchange. FHIR supports using standardized terminologies such as SNOMED CT and LOINC for clinical documentation and decision support, fostering agile development and enhancing interoperability to improve patient care and population health. Similarly, Observational Medical Outcomes Partnership (OMOP) is a standard data model designed for harmonizing healthcare data across various sources to facilitate large-scale research. By standardizing data structures, OMOP supports improved observational data quality and enables meaningful, evidence-based research on clinical practices, drug safety, and disease trends. Integrating terminologies such as SNOMED CT and ICD-10 within OMOP enhances data consistency and promotes collaboration among research institutions, advancing innovation in healthcare [27]. These international standards enhance research capabilities, enabling healthcare organizations to efficiently manage population health, conduct comparative effectiveness studies, and develop innovative healthcare applications that seamlessly integrate with EHR platforms [28]. Additionally, standardized terminologies help maintain consistency in patient care documentation, ultimately improving clinical data quality and reliability.

This dissertation primarily focuses on SNOMED CT. Its hierarchical structure organizes concepts into so-called SNOMED CT hierarchies, allowing healthcare professionals to navigate from broad domains to specialized subclasses. Current translation efforts into languages such as German link approximately 79,000 medical terms to 41,000 SNOMED CT concepts [29].

1.2 Machine Learning for Text Processing

The increasing complexity and volume of data in EHRs have driven the need for advanced analytical techniques, such as ML and NLP, to transform this comprehensive repository of information into actionable insights. These technologies have the potential to drive personalized healthcare solutions by addressing the challenges associated with unstructured clinical data [30]. Although EHR systems have transformed healthcare by providing reliable patient records and enabling efficient information sharing, their full potential remains unrealized due to the overwhelming complexity and scale of the data they manage.

Recent advancements in NLP, such as the creation of GatorTron [31] – a large clinical language model trained on over 90 billion words, including over 82 billion de-identified clinical text entries – illustrate how NLP techniques can unlock the latent potential of EHRs. These methods enable the extraction of meaningful insights from unstructured data, facilitating improved decision-making. For example, Wieland et al. [30] reported that 65% of studies utilizing NLP for processing EHR data employed deep learning and rule-based methods. Additionally, Juhn et al. [32] highlighted the role of NLP in automating chart reviews and standardizing phenotype definitions for research on allergies, asthma, and immunology.

The subsequent sections will delve into the application of ML techniques such as classical, deep learning methods, and LLMs in advancing the capabilities of NLP for healthcare. These approaches aim to bridge the gap between raw clinical data and actionable insights, focusing on information extraction and standardization tasks.

1.2.1 Classical Machine Learning Methods

Classical ML refers to traditional ML techniques that predate deep learning and LLMs. These methods typically involve structured data and feature engineering and require less computational power. They include supervised, semi-supervised, and unsupervised learning, with standard algorithms such as support vector machines (SVMs), random forests, logistic regression, naïve Bayes, and clustering methods like K-Means. Dimensionality reduction techniques, such as principal component analysis (PCA) and t-distributed

stochastic neighbour embedding (t-SNE), are also widely used. Each method has distinct strengths and applications in healthcare, particularly with EHR data.

In healthcare, classical ML demonstrates particular utility for EHR-based applications including disease prediction, clinical phenotyping, and decision support. These methods effectively process both structured and unstructured data - for instance, SVMs and conditional random fields (CRFs) have successfully extracted clinical concepts from unstructured EHR notes [33]. Their rule-free, generalizable approach enables robust analysis of diverse patient records.

SVMs, are highly effective for both binary and multi-class classification tasks. They are broadly used in healthcare informatics to predict diseases or categorize patients based on diagnostic codes. By leveraging optimal hyperplanes to separate classes, SVMs excel at handling high-dimensional EHR data. For example, research conducted by Zhang et al. [34] demonstrated the potential of SVMs, achieving an accuracy of 86.2% for ten cancer types and 97.33% for three cancer types using 400 records per type.

Random forests, an ensemble learning method, construct multiple decision trees to effectively manage high-dimensional data, such as patient records and test results. These algorithms automatically identify important features, making them highly valuable for risk factor identification and prediction. A study by Dubrava et al. [35] used random forests to predict diabetic peripheral neuropathy from EHR data, achieving an accuracy of 89.6% and an AUC of 0.824.

Logistic regression, a widely recognized classification method, is particularly valuable for its interpretability in binary classification tasks. It is frequently used to predict the likelihood of diseases. However, careful feature engineering is often required to address the complexities of clinical data. For instance, Duan et al. [36] developed a distributed logistic regression model to assess the risk of fetal loss due to medication exposure, achieving accuracy comparable to pooled data analysis.

Naïve Bayes classifiers are effective for tasks involving smaller datasets and are particularly useful in text classification applications, such as patient feedback or sentiment analysis. Their simplicity and scalability make them suitable for categorizing clinical notes and outcomes. Andry et al. [37] employed naïve Bayes to predict heart disease using EHR data, enabling the identification and management of cardiovascular risks, such as heart attacks and arrhythmia.

Despite advances in deep learning, classical ML retains clinical relevance due to its interpretability, computational efficiency, and effectiveness with limited data. These attributes prove particularly valuable for clinical decision-making, where model transparency is essential. Classical approaches typically employ cross-validation and standard performance metrics (precision, recall, F1 score), though they require manual feature engineering. Modern approaches (e.g., transformers, LLMs) reduce feature engineering and excel with unstructured data, while classical methods remain valuable for interpretable tasks like health forecasting and decision support [38]. This trade-off motivates continued evaluation of both paradigms in healthcare analytics.

1.2.2 Deep Learning Methods

Deep learning has greatly helped health data processing move into new territory by equipping ML and NLP techniques with automatic and hierarchical learning capabilities. Techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) proved efficient in processing not only structured patient data but also unstructured ones such as images and narratives, while keeping dependence on manual feature engineering at minimal levels [39]. The scalability of deep learning facilitates the analysis of large datasets, making it particularly suitable for tasks such as clinical event prediction, medical image analysis, and text processing. Besides, transfer learning has also become a beneficial approach within healthcare, allowing the fine-tuning of pre-trained models for domain-specific applications and reducing the challenges of working with limited availability of labelled data [40]. With that said, deep learning algorithms are often more accurate than classical methods for clinical data analytics.

Following the discussion on deep learning methods, several architectures have emerged as pivotal tools in healthcare analytics, each offering unique strengths for processing diverse data types. CNNs, traditionally designed for image data, have demonstrated efficacy in healthcare beyond imaging, such as text classification and feature extraction from text. For instance, Suo et al.[41] developed a CNN-based framework to create patient representations for personalized disease prediction, highlighting its adaptability to tasks with limited annotated data. RNNs are specialized for sequential data, such as patient visit records, where their memory capabilities enable temporal pattern recognition. Other studies, e.g., Rasmy et al. [42], have effectively employed RNNs for predicting heart failure

risks, achieving robust performance across healthcare information systems. Advancing RNN architectures, Long Short-Term Memory Networks (LSTMs) address the vanishing gradient problem, and excel in capturing long-term dependencies within sequential data [43, 44]. Their ability to analyse clinical narratives and structured data has been demonstrated by Guo et al. [45], where they forecast cardiovascular health outcomes from extensive patient datasets. Bidirectional LSTMs (Bi-LSTMs) further enhance context comprehension by processing sequences in both forward and backward directions, proving particularly effective for tasks that require a nuanced understanding of clinical notes, such as symptom extraction and diagnostics [46].

More recently, transformer-based architectures, which utilize self-attention mechanisms, have transformed NLP in healthcare by capturing relationships within sequences without recurrence [47]. Models like bidirectional encoder representations from transformers (BERT) and their domain-specific variants have performed exceptionally in extracting meaningful insights from clinical narratives. ClinicalBERT, fine-tuned with clinical notes from MIMIC-III, has played a key role in predicting in-hospital mortality and adverse events [48]. Similarly, BioBERT and BlueBERT, trained with biomedical literature and clinical datasets, excel in NER and MCN tasks, enhancing outcomes prediction and decision support [49, 50]. Other specialized transformer models include SciBERT, tailored for scientific literature, and medBERT.de, optimized for structured clinical data such as medical codes and lab results, significantly improving disease prediction accuracy in resource-constrained settings [51].

Additionally, Self-Alignment Pretraining BERT (SapBERT) utilizes self-alignment pretraining with UMLS Metathesaurus data to improve the representation and linking of synonymous biomedical language expressions, improving information retrieval and analysis in healthcare contexts [52]. Collectively, these architectures have not only advanced predictive capabilities but also broadened the scope of healthcare research and clinical applications. This method constructs triplets of anchor, positive, and negative samples to optimize a metric learning framework, which improves the model's ability to distinguish between similar and dissimilar entities. Including ADAPTER [53], a lightweight module designed to fine-tune pre-trained language models with minimal additional parameters, within SapBERT preserves its efficiency, allowing effective training with fewer parameters. Notably, SapBERT achieves a performance improvement of up to 20% over models such as BioBERT and PubMedBERT [54], making it particularly effective for

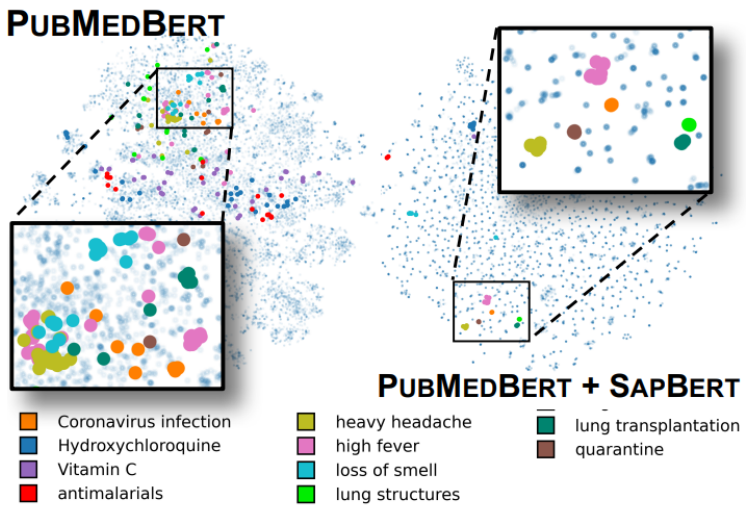

Figure 1.1: t-SNE visualization of UMLS entity embeddings: PubMedBERT (left) shows overlap, while PubMedBERT plus SapBERT (right) forms compact clusters after self-alignment [52].

biomedical entity linking, information retrieval, and the analysis of scientific literature (see Figure 1.1). These advancements in SapBERT have significantly enhanced predictive capabilities, expanding the scope of healthcare research and clinical applications by enabling more accurate extraction of insights from EHR data [52].

1.2.3 Large Language Models

LLMs, advanced ML models typically based on transformer architectures, are designed to generate and interpret human language by predicting contextual word or sequence probabilities. Trained on extensive datasets, they are well-suited for healthcare NLP, particularly in processing complex clinical texts due to their ability to capture long-range dependencies [55]. Unlike task-specific models such as BERT, LLMs perform a broad range of tasks with minimal fine-tuning, demonstrating robust few-shot and zero-shot learning. This adaptability is critical in healthcare, where annotated data are scarce due to privacy constraints. Furthermore, their scalability and capacity for multimodal integration (e.g., text, imaging, lab results, genomics) support comprehensive clinical decision-making [56]. LLMs are already influencing clinical documentation, research, and provider–patient interactions, with substantial potential to transform decision-making

and patient engagement [57]. A focused overview of certain LLMs that are particularly relevant for clinical NLP practices are:

Generative Pre-trained Transformer. The generative pre-trained transformer (GPT) models must be specifically mentioned in this scope. Created by OpenAI, it has many versions, each increasing the capability and application compared to its predecessors. GPT-2 [58] released in 2019, considerably outperformed its earlier version in generating text. Thus, coherent and contextually appropriate responses can be generated, enabling intelligent conversational AI. In 2020, GPT-3 [59] was introduced, having 175 billion parameters, drastically improving understanding and generation of the language. Hence, this technology finds many applications in dialogue systems and content creation. GPT-4 [60], launched in 2023, specializes in generating human-like text with a deep understanding of complex language tasks and combines multimodal inputs such as text and images. Clinical applications include GPT models in generating patient summaries, medical charting applications, and answering clinical queries. These are also useful for creating conversational agents with patients [61].

Other Large Language Models, beyond GPT. These models have also significantly contributed to clinical NLP practices. Pathways Language Model (PaLM), developed by Google, uses a mixture of expert architectures to process diverse language processing tasks efficiently. This makes the model of great help in generating clinical texts and supporting decision-making [62]. Large Language Model Meta AI (LLaMA), released by Meta AI, excels in summarizing clinical documents and extracting information from EHRs [63]. GatorTron, developed for clinical documents and understanding at the University of Florida, extracts structured data from unstructured clinical notes [31]. Text-to-Text Transfer Transformer (T5), also developed by Google, applies a text-to-text framework to various NLP tasks, including summarization and patient education [64]. Lastly, DeepSeek [65] is advancing clinical NLP by enabling broader AI access and leveraging fine-tuned LLM capabilities to improve accuracy and adaptability in real-world healthcare applications. Collectively, these models improve the processing and interpretation of complex clinical language, aiding healthcare providers in enhancing patient care.

1.3 Key Tasks in Text Processing

Integrating classical ML, deep learning, and LLMs has changed the analysis of EHRs, improving clinical and administrative healthcare functions. These are important in patient care improvement, smoothing of functions within the hospital, and informing data-driven decisions [66, 67]. The critical applications include disease prediction and risk stratification, clinical decision support, clinical text classification, and patient outcome prediction. Moreover, information extraction, medical text summarization, and question-answering NLP techniques have boosted it further to interpret and utilize such complex clinical narratives [68]. This suite of NLP applications is fundamental in changing healthcare workflows for speed, accuracy, and personalization. NLP techniques are beneficial for extracting key insights from unstructured EHR narratives. These techniques helped bridge the gap between narrative data and structured data elements; this enhanced automation of several time-consuming tasks improved utility from the clinical perspective [69]. Three of the significant NLP tasks that enable healthcare providers to efficiently extract and effectively utilize information from EHRs are listed below.

1.3.1 Named Entity Recognition

NER is a basic NLP task that identifies and classifies medical terms in clinical text, such as referring to all types of entities referred to in clinical text [70]. These language expressions are then typically tagged and labelled so that NER can convert free-text clinical narratives into structured, usable data. For example, it may automatically identify mentions of diseases such as “diabetes” or drugs like “metformin”. This structured representation will enable clinicians and healthcare systems to retrieve relevant information quickly, allow data to be analysed, and inform decisions. NER plays a critical role in automating administrative tasks, including medical coding, and amplifies clinical research by making textual data available in large volumes [71].

1.3.2 Medical Concept Normalization

MCN can be framed as either a classification, generate-and-rank, or embedding similarity task. In the classification approach, deep learning models are trained to map text

embeddings to numerical labels, often facing challenges due to many output classes. To address this, methods such as type prediction [72], semantic type modelling [73], and prior knowledge from medical terminology systems [74], such as SNOMED CT, are integrated, improving model performance and reducing complexity. Furthermore, MCN models must account for the ambiguity of medical mentions, as the same mention can have different meanings depending on the context. Contextual embeddings and type prediction help disambiguate terms, ensuring each mention is mapped to the correct concept. The generate-and-rank approach generates candidate concepts and then ranks them based on semantic and contextual relevance, utilizing neural networks such as convolutional or transformer-based models. Embedding similarity, another common method, compares terminology terms and mentions from clinical text embeddings through distance metrics, such as cosine similarity, with improvements through triplet loss or external knowledge integration. Combining UMLS knowledge with transformer models and similarity-based approaches significantly improves normalization accuracy, particularly in large-scale and diverse datasets. Once identified, references to medical concepts must be standardized to ensure consistency across different healthcare systems. MCN maps the extracted terms to standardized medical vocabularies, for instance "heart attack" in a clinical text is linked to the ICD-10 code "I21" (Acute Myocardial Infarction). This process ensures that different textual expressions that mean the same are mapped to a single, widely recognized concept [75].

1.3.3 Relation Extraction

Relation extraction extends beyond identifying and normalizing individual entities by discerning their interconnections. For instance, it can link a patient's diagnosis, such as diabetes, to prescribed treatments like insulin or detect adverse drug reactions documented in clinical notes. By capturing these intricate relationships, relation extraction provides healthcare professionals deeper insights into patient conditions and treatment histories. This process is fundamental to constructing knowledge graphs that map interactions between medical concepts, facilitating clinical research, optimizing patient management, and enabling personalized treatment recommendations [76].

1.4 Derived Dissertation Objectives

This *cumulative thesis* focuses on the outcomes of three interconnected projects [77–79] that form a cohesive framework for automating clinical text processing, grounded in the following *research gaps – observations* which lead to the stated research questions and their operationalization into defined experimental objectives:

Observation A. Clinical language exhibits high variability (e.g., abbreviations, synonyms, ambiguous phrasing), complicating accurate information extraction. A comparisons of classical and different deep learning architectures of their efficacy in their clinical tasks (e.g.,NER, MCN) is often lacking specifically in combination with model interpretability, particularly through vector visualization and explainability techniques, which is critical for clinical adoption but remains inadequately addressed.

Research Question A. "How can comparative modeling (classical vs. deep learning) and explainability techniques improve the accuracy and interpretability of clinical data extraction?" The research question was operationalized in the first investigation via a comparison of classical and deep learning models for extracting oxygen therapy data from COVID-19 records, integrating vector visualization to enhance interpretability [77].

Observation B. Manual annotation of clinical narratives with standardized concepts (e.g., SNOMED CT) is resource-intensive, and supervised methods fail to scale due to reliance on labelled data. Current approaches treat NER and MCN as separate tasks, neglecting synergies between them. A unified, label-efficient framework is lacking for joint NER-MCN. This leads to the adoption of unsupervised methods in this context and their reachable performance, in contrast to the more often used supervised model-based approaches.

Research Question B. "Can a unified, unsupervised pipeline leverage semantic similarity to jointly perform NER and MCN without labelled data?" The research question was investigated via an SapBERT-based bi-encoder for joint NER-MCN, eliminating dependency on labeled data and leveraging SNOMED CT ontologies [78].

Observation C. Noisy clinical text (misspellings, ad-hoc abbreviations) degrades MCN performance, and traditional methods struggle with contextual disambiguation. Despite the potential of LLMs, their systematic application to clinical text cleansing (e.g., error correction, abbreviation expansion) and normalization is limited. In addition, the integration of Retrieval-Augmented Generation (RAG) to enhance MCN by dynamically grounding LLMs in curated medical ontologies (e.g., UMLS, SNOMED CT) is of interest.

Research Question C. "How can LLMs and RAG mitigate noise (e.g., misspellings, abbreviations) to enhance MCN accuracy?" The objective in the third investigation was therefore to design an LLM pipeline with RAG to cleanse noisy clinical text (e.g., spelling corrections) and dynamically retrieve ontology concepts for accurate normalization [79].

These investigations are summarized in the next chapter, and their key contributions to answering the stated research questions are presented. Additionally, Chapter 2 contains complementary research efforts, including advancements in vector encoding techniques for concept normalization, re-ranking approaches leveraging cross-encoders, and the integration of visualization tools to enhance encoding processes. Further, the critical role of visualizing structured and unstructured patient data is highlighted. Collectively, these efforts should provide scalable solutions for efficient clinical data processing, improved healthcare management, and broader research applications.

Chapter 2

Investigations

2.1 O2 Supplementation Disambiguation in Clinical Narratives to Support Retrospective COVID-19 Studies

2.1.1 Summary

Accurate oxygen saturation assessment is vital for evaluating COVID-19 severity, particularly in silent hypoxemia cases where patients exhibit no apparent symptoms despite critically low oxygen levels. Although such information is embedded in clinical narratives within EHRs, manual extraction of oxygen supplementation details remains labor-intensive and limits scalable retrospective research. Existing methods fail to automate this process reliably. This study [77] develops and validates an NLP approach to extract and classify oxygen supplementation data from German discharge summaries, thereby enabling efficient cohort identification and reducing physician workload.

To achieve this, we analyzed text lines extracted using regular expressions from anonymized COVID-19 patient discharge summaries written in German. Further, the patients were classified into those who received oxygen supplementation and those who did not. A comparative analysis of various ML algorithms was conducted, ranging from classical to deep learning models. To qualitatively assess semantic encoding, t-SNE was

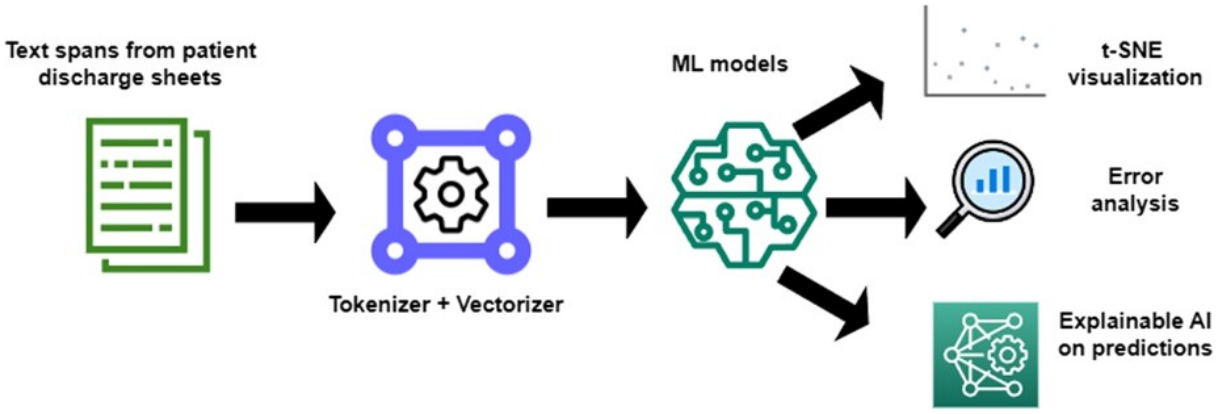

Figure 2.1: Graphical abstract summarizing the proposed methodology, including text line preprocessing, binary text classification, vector clustering using t-SNE visualization, error analysis, and explainable AI interpretations [77].

used to visualize vector clustering. The Local Interpretable Model-agnostic Explanations (LIME) module was used to explain classifier decisions, illustrating how the models arrived at their conclusions.

The results demonstrated that classical and deep learning models achieved similar classification performance, with F-measures between 0.942 and 0.955. However, classical ML methods were faster and less computationally intensive. Analysis of embedding representations indicated notable differences in encoding patterns. Deep learning models formed clearer clusters than classical ML models. t-SNE was used to analyze embedding representations, highlighting the semantic encoding properties of the deep learning models. Furthermore, LIME explanations reveal the most relevant features influencing model decisions at the token level.

In conclusion, this research underscores that classical ML methods can match the performance of deep learning models while offering lower computational costs. LIME interpretation improves the understanding of classification outcomes. This aids automated clinical text processing for COVID-19 research and patient care. For a visual summary, refer to Figure 2.1 for the graphical abstract.

2.1.2 Discussion

Due to abbreviations, synonyms, and ambiguity, clinical language is highly variable, making accurate information extraction challenging. Comparative evaluations of classical and deep learning models, particularly with regard to interpretability and explainability, are limited and hinder clinical adoption. This investigation addresses these gaps by systematically exploring ML-based approaches for classifying oxygen supplementation status in COVID-19 patients using German-language discharge summaries.

The dataset comprises 3,844 anonymised text lines from the KAGes EHR system, an Austrian public hospital network. Two biomedical experts annotated the data and validated it by a third, achieving high inter-annotator agreement (Cohen's Kappa = 0.859). Text lines were labeled as 1 (evidence of oxygen supplementation) or 0 (no evidence), with 3,074 spans for training and 769 for testing. Models such as SVM, random forest, LSTM, Bi-LSTM, and CNN were selected for comparative analysis based on their suitability for classification tasks and handling imbalanced data. These models allow for a comprehensive evaluation of performance metrics such as precision, recall, and F1 score, which are crucial for clinical applications.

The findings underscore the effectiveness of both classical and deep learning models, with F1 scores ranging between 0.942 and 0.955. Notably, classical ML techniques such as random forest demonstrated competitive performance while being computationally more efficient than deep learning models like LSTM and CNN. This result challenges the prevailing assumption that deep learning methods are universally superior, emphasizing that classical models remain viable for specific clinical NLP tasks, particularly in time-sensitive applications.

The t-SNE visualization demonstrated that the dynamic embeddings generated by the CNN model yielded better cluster separability than SVMs static term frequency-inverse document frequency (TF-IDF)representations, highlighting the limitations of traditional feature extraction methods. A key contribution of this study is the use of explainable AI (XAI) module such as LIME to enhance model interpretability. By identifying influential features such as "FiO2", "mit" (German for "with", as in "mit Sauerstoff"), and specific oxygen volume terms, the approach offers transparency into classifier decisions and addresses the "black-box" problem—an essential step toward fostering clinician trust in medical AI.

This study highlights both the value and challenges of using EHR narratives for clinical research, demonstrating their unique ability to capture contextual patient information missing from structured data while facing two key limitations: significant class imbalance (with predominating "no oxygen supplementation" cases that may affect minority class sensitivity despite using precision/recall/F1 metrics) and restricted generalizability from relying on a single hospital system's data. While these narratives provide richer insights for retrospective analysis than structured data alone, future work should address current limitations through advanced techniques like data augmentation or weighted loss functions for class imbalance and multi-center validation studies to enhance generalizability, thereby preserving the method's strengths while overcoming its constraints.

Overall, the findings highlight clinical narratives as a rich source for extracting parameters relevant to cohort selection and epidemiological studies. The results underscore the need for context-aware ML model selection. Given the comparable performance of classical and deep learning models, computational efficiency becomes a key factor for deployment in resource-constrained clinical settings. While deep learning models offer advanced feature representations, their computational demands may outweigh benefits in such environments.

This study also advances the use of explainability in clinical NLP. LIME enhances interpretability, promoting clinical trust, and future work could include additional visualization strategies (e.g., concept distributions, dashboards) to support usability. Despite persistent data access, imbalance, and generalizability issues, the effectiveness of classical models combined with explainability tools provides a strong basis for advancing clinical NLP. Expanding datasets to include multilingual and multi-institutional sources will strengthen model robustness and clinical relevance.

2.1.3 Contribution

The key contributions include:

- an automated oxygen supplementation classification from clinical narratives, crucial for more efficient and accurate COVID-19 retrospective studies.

- a comparative analysis of classical and deep learning models, enhanced by improved model interpretability, contributing to a better understanding of how these models work.

Research Question A: "How can comparative modeling (classical vs. deep learning) and explainability techniques improve the accuracy and interpretability of clinical data extraction?"
Clinical language's inherent variability and ambiguity necessitate robust methods for accurate data extraction. Our hybrid NLP system addresses this challenge by integrating rule-based techniques with ML classification, demonstrating that deterministic rules effectively capture explicit patterns (e.g., "FiO2 50%"). At the same time, ML models resolve contextual ambiguities (e.g., distinguishing between "supplemental oxygen required" and "no oxygen required"). To ensure clinical interpretability, LIME module validates model decisions, fostering trust among end-users. This structured approach transforms unstructured oxygen therapy documentation into reusable formats, facilitating applications in clinical decision support and retrospective research. However, future extensions could enhance interoperability by incorporating standardized terminologies (e.g., SNOMED CT or UMLS) for broader semantic alignment. The findings underscore the viability of hybrid systems in balancing precision and adaptability, though further work is needed to generalize the framework across diverse clinical domains.

2.2 Unsupervised SapBERT-based Bi-Encoders for Medical Concept Annotation of Clinical Narratives with SNOMED CT

2.2.1 Summary

This study [78] addresses the challenge of integrating de-identified clinical narratives from EHRs with standardized medical vocabularies to enhance their effective utilization. While pre-trained on large datasets, traditional models like BERT still require substantial annotated corpora for tasks like NER and MCN. The experiment proposes an unsupervised approach to medical concept annotation (MCA) that structures clinical narratives into

standardized terminologies while automatically generating annotated datasets. By leveraging semantic similarity techniques, the unified pipeline performs both NER and MCN without manual intervention, directly aligning raw text with SNOMED CT concepts.

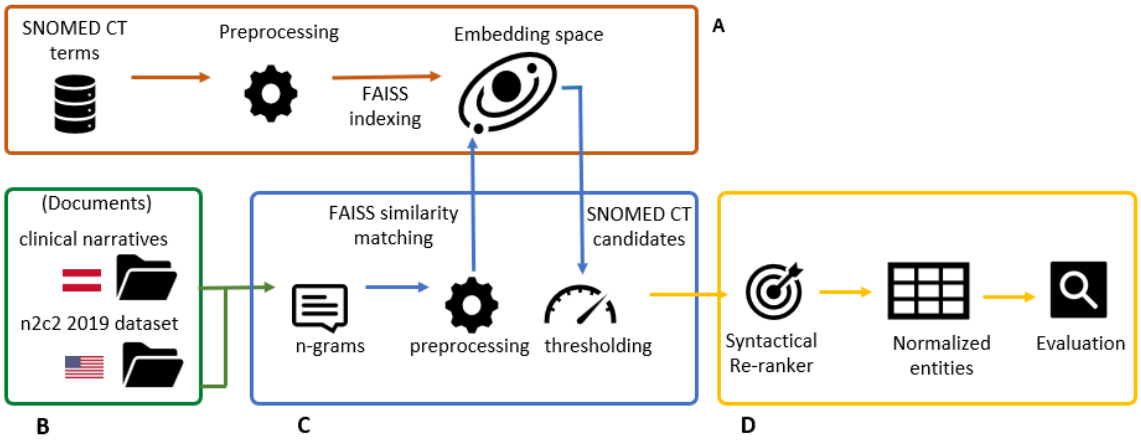

Figure 2.2: Graphical abstract illustrating the proposed framework: embedding space creation for SNOMED CT, clinical data preparation, n-gram generation with entity recognition and normalization, and best n-gram selection and evaluation [78].

The proposed method processes English/German clinical narratives by extracting 1-5 word n-grams as candidate terms. Semantic matching is performed using a Facebook AI similarity search (Faiss[1])-accelerated similarity search between n-gram embeddings and pre-computed SNOMED CT concept embeddings in SapBERTs bi-encoder space. This matching is further refined through syntactic re-ranking to improve alignment accuracy. The unified pipeline simultaneously achieves NER by validating candidate terms against SNOMED CT semantic tags and MCN through standardized concept ID assignment, enabling complete terminology alignment without requiring annotated training data. Key innovations include leveraging SapBERTs self-supervised biomedical knowledge, Faiss-optimized search scalability, and language-agnostic processing validated across both English and German clinical texts.

The results demonstrate that the unsupervised approach achieves an F1 score of 0.765 for MCN in English and 0.557 in German, showing promise in the absence of training data. Notably, the semantic tag “disorder” achieved the highest F1 scores, with 0.871 for English and 0.648 for German datasets in MCN, while the results on MCA of “disorder”

[1]https://Faiss.ai/

were dropping, i.e, 0.839 and 0.696 in English and 0.685 and 0.437 in German dataset for NER and MCN respectively.

This unsupervised approach offers a viable method for preliminary annotation tasks and can potentially serve as a pre-labeling mechanism in manual annotation processes. The study recognizes various ongoing challenges, such as false positives, contextual errors, and the inherent variability of clinical language. Further refinement is required to ensure the method is more precise and applicable in clinical practice. For a visual summary, refer to Figure 2.2 for the graphical abstract.

2.2.2 Discussion

The present study [78] demonstrates the viability of unsupervised MCA for clinical narratives through a SapBERT-based bi-encoder approach, evaluated on both English (n2c2 2019 [52]) and German (hospital EHR) datasets. The English dataset comprising 100 U.S. hospital discharge summaries with 10,919 manually annotated medical concepts mapped to UMLS, of which we focused on 12,760 SNOMED CT-normalized mentions (6,232 training, 6,528 test) across the top 10 semantic categories (covering 95% of concepts); and (2) a newly created German dataset of 10 Austrian hospital EHRs containing 600 SNOMED CT-normalized mentions (97% within the same semantic categories), annotated using INCEpTION[2] following n2c2 guidelines [52]. These diverse, complementary datasets, comprising physician notes and discharge summaries from different healthcare systems and documentation standards, enable robust evaluation of consistent SNOMED CT mapping across varied real-world clinical language and linguistic contexts.

Our methodology addresses a challenging gap in clinical NLP by trying to eliminate the dependency on manually annotated training data, which is particularly valuable for low-resource settings that face challenges of data sensitivity and time constraints for annotation. The proposed approach comprises three steps: (1) n-gram generation from clinical narratives, (2) similarity matching using a bi-encoder model, and (3) syntactical re-ranking to refine candidate terms. The bi-encoder architecture enables direct mapping of clinical text to standardized terminologies without intermediary supervision steps that typically constrain NER systems. SNOMED CT (July 2022 international edition) served

[2]https://inception-project.github.io/

as the reference terminology, with all concepts pre-embedded in SapBERT to enable efficient matching. The framework employs Faiss-accelerated nearest neighbor search to balance semantic accuracy and computational efficiency.

Performance analysis revealed competent English results (MCN F1 score = 0.765), outperforming a conventional n-gram-based dictionary matching baseline, which achieved an F1 score of 0.512 on the same dataset. As a benchmark, the baseline method employed a straightforward pipeline: (1) n-gram generation from clinical text and (2) exact string matching against SNOMED CT concepts without semantic or contextual processing. While computationally efficient, this approach exhibited significant limitations, notably higher false positives and an inability to resolve syntactic variations or polysemy. The substantial gap (F1 score = 0.253) between the baseline and our SapBERT bi-encoder underscores the limitations of rule-based dictionary methods in clinical NLP, particularly for context-sensitive categories like "qualifier value" (baseline F1 score of 0.217 vs. proposed F1 score of 0.621).

The syntactic re-ranking component effectively addressed MCN challenges, resolving partial matches and term variations by incorporating structural features. However, processing times of 60 seconds per line in EHR data reveal significant scalability constraints for clinical deployment. Error analysis identified two core challenges: (1) reduced accuracy for context-dependent categories ("qualifier value", "observable entity") requiring broader discourse understanding, and (2) difficulties with discontinuous mentions, exposing limitations of n-gram approaches in capturing clinical semantics. These findings suggest three improvement pathways: context-sensitive disambiguation through neural sequence labeling or hybrid methods, refined normalization filters to reduce false positives, and integrating advanced language models, such as GPT, for enhanced extraction.

In summary, the approach demonstrates clinical promise by automatically extracting standardized concepts from English and German narratives, potentially enhancing semantic interoperability, particularly in resource-constrained settings lacking annotated corpora. The SapBERT-based bi-encoder presents an effective unsupervised MCA solution that minimizes dependence on labeled data. While promising, challenges persist in reducing false positives, improving processing efficiency, and adapting to multilingual contexts. Future work should focus on developing hybrid approaches leveraging LLMs to this task, where a selected investigation is described in Section 2.3.

2.2.3 Contribution

The key contributions include:

- the integration of NER and MCN into a single embedding-based annotation process.
- the unsupervised mapping of clinical text to SNOMED CT using SapBERTs embeddings.

Research Question B: "Can a unified, unsupervised pipeline leverage semantic similarity to jointly perform NER and MCN without labeled data?"
The scarcity of labeled clinical data poses a significant barrier to scalable MCN. Our unsupervised pipeline, leveraging a SapBERT-based bi-encoder, is tryinig to circumvent this limitation by projecting clinical text and SNOMED CT concepts into a shared semantic space, enabling accurate matching of variant expressions (e.g., "heart attack" → "Myocardial Infarction") without manual annotation. This unified approach jointly performs NER and MCN, mitigating error propagation inherent in cascaded architectures. The results demonstrate promising performance across heterogeneous EHR data, highlighting its capacity to handle synonyms, abbreviations, and contextual nuances. While the method achieves scalability for real-time processing, rare or polysemous concepts may benefit from supplementary ontology-guided constraints. The study advances the field by proving that unsupervised semantic matching can support standardizing clinical narratives, thereby reducing reliance on costly labeled datasets.

2.3 Large Language Models for Clinical Text Cleansing Enhance Medical Concept Normalization

2.3.1 Summary

MCN remains challenging due to variations in clinical documentation. While LLMs show promise for text processing, its potential for cleansing clinical text to improve MCN has been underexplored. This study [79] explores the application of LLMs, specifically GPT-4, in processing de-identified clinical narratives from EHRs, which are primarily available

as free text. The research aims to enhance the accuracy of clinical documentation and improve healthcare delivery by focusing on a corpus of anonymized clinical narratives in German.

The study evaluates two primary tasks: text cleansing, which involves automatically rephrasing raw clinical text into a more readable and standardized format, and RAG aimed at enhancing SapBERT-based MCN. We evaluated the framework on 660 manually annotated surface terms extracted from real-world clinical narratives. Identical analyses were conducted on the original narratives and their LLM-cleansed versions to isolate preprocessing effects.

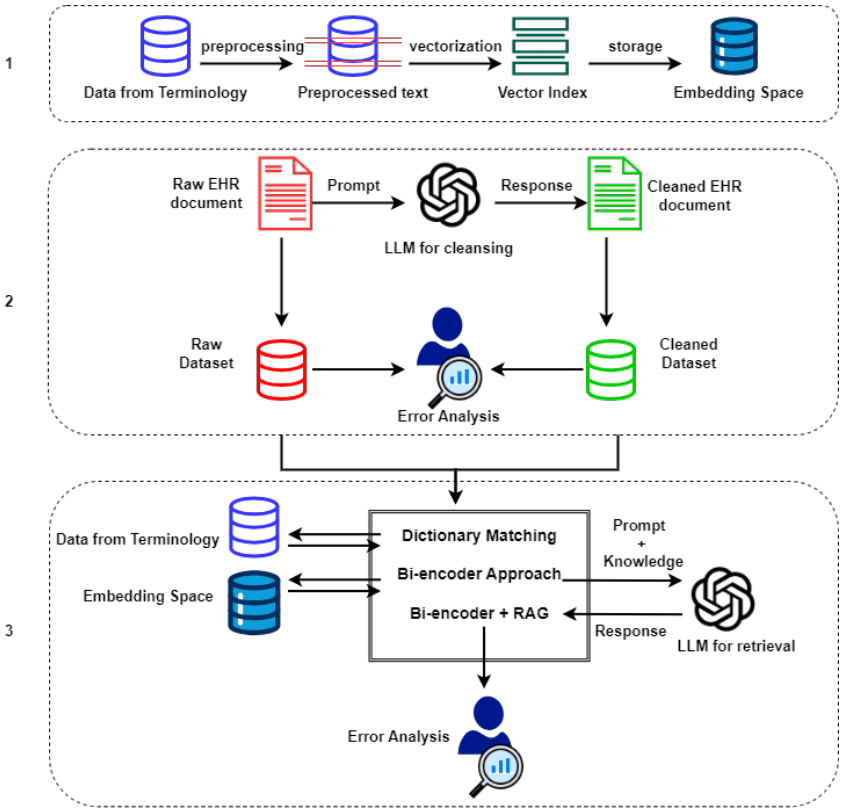

Figure 2.3: Graphical abstract summarizing the methodology: embedding space generation from two German terminologies, clinical narrative cleansing with LLM and error analysis, and MCN via dictionary lookup, bi-encoders, and RAG followed by error analysis [79].

The findings reveal that the application of GPT-4 significantly outperforms classical NLP algorithms for MCN with text cleansing. The study still emphasizes the importance of

a robust terminological foundation, which is crucial for enhancing the LLM supported MCN performance. The research achieved maximum F1 scores of 0.607, 0.735, and 0.754 for the top 1, 5, and 10 matches, respectively. These scores were obtained through a comprehensive pipeline that included document cleansing, bi-encoder-based term matching utilizing an extensive domain dictionary aligned with SNOMED CT, and subsequent re-ranking via RAG.

In conclusion, the study demonstrates that LLMs can significantly enhance the processing of clinical narratives, contributing to improved MCN outcomes. The results suggest further exploration of LLMs in the clinical setting, emphasizing their ability to handle unstructured data effectively. For a visual summary, refer to Figure 2.3 for the graphical abstract.

2.3.2 Discussion

This study addresses a critical challenge of accurate MCN in clinical NLP, particularly for noisy, unstructured clinical narratives that exhibit frequent abbreviations, typographical errors, and documentation variability. Traditional approaches, which predominantly rely on rule-based cleansing or standalone deep learning models, often struggle to generalize across different institutions and languages.

To bridge this gap, this study introduces a LLM-based text-cleansing pipeline that optimizes input data before MCN, enhancing performance. Specifically, the approach leverages the particular LLM, GPT-4 accessed via API calls, in a two-step process. First, to standardize and clarify clinical narratives through text cleansing, and second, to improve term-matching accuracy via RAG. This dual application pipeline significantly enhances the mapping of clinical terms to SNOMED CT codes, with notable improvements observed in German-language clinical texts.

GPT-4-based text cleansing reduced word count by 7.72% and lines by 67.85%, improving clarity but altering 45% of annotated spans (33% beneficial, 8% potentially harmful). Only 0.45% of terms were lost, requiring manual correction. For MCN, cleansing boosted performance: bi-encoder with IT_DE terminology achieved F1 scores of 0.568 (top-1), 0.735 (top-5), and 0.754 (top-10), surpassing dictionary matching by 91.25%. RAG fur-

ther improved top-1 F1 score by 6.87% (to 0.607 for top-1), with IT_DE consistently outperforms UMLS_DE. Expert validation ensured reliability amid AI-driven changes.

The GPT-4-based text cleansing phase proved instrumental in standardizing German clinical narratives, achieving measurable improvements through abbreviation expansion (e.g., "HT" – "Hypertonie"), typo correction, and syntactic normalization. However, challenges such as occasional misinterpretations and hallucinations remain, where the model incorrectly expands abbreviations or modifies medical terminology in unintended ways. For example, GPT-4 misinterpreted "EZ" as "Körpermaße" (body measurements) instead of the correct "Ernährungszustand" (nutritional status), resulting in incorrect SNOMED CT annotations. Similar issues arose with medication-related terms, where incorrect modifications could impact clinical decision-making. The model frequently generated incorrect drug names and chemical compounds, misclassifying substances in ways that could disrupt pharmacovigilance and medication reconciliation. These errors were especially prevalent with ambiguous or infrequent clinical terminology, highlighting the need for rigorous validation before deploying LLMs in real-world medical settings.

Following text cleansing, the second phase – LLM-driven re-ranking within the RAG framework significantly enhanced the MCN process by improving term-matching accuracy of the bi-encoder approach. This step enhances retrieval performance by leveraging LLMs to prioritize the most relevant SNOMED CT concepts from candidate matches. The results indicate that RAG-based re-ranking significantly outperforms traditional dictionary-based and bi-encoder approaches, achieving an F1 score of 0.607 for top-1 match. However, this method's effectiveness remains highly dependent on terminological consistency and the comprehensiveness of language-specific resources. Notably, the German Interface Terminology (IT_DE) provided superior performance compared to the smaller UMLS extract (UMLS_DE), emphasizing the importance of extensive terminological coverage in MCN tasks.

Three main insights emerge for real-world deployment: (1) Terminology specificity is crucial—IT_DE's tailored design for German EHRs outperformed UMLS_DE's broader but shallower coverage. (2) Error profiles vary by clinical domain—pharmacologic terms required three times more manual corrections than anatomic concepts, highlighting the need for specialty-specific validation. (3) Cautious adoption is necessary—while GPT-4 cleansing improved efficiency (67.85% line reduction), its 8% error rate in critical categories

necessitates expert review before EHR integration. These findings position LLM-assisted MCN as a transformative but not yet fully autonomous solution for German clinical data interoperability.

Despite advancements, several challenges remain. Hallucinations persist, manifesting as incorrect terminology modifications during cleansing and misleading term matches during re-ranking. The study's focus on German clinical text limits generalizability, particularly for languages with limited terminological resources, potentially necessitating alternative normalization strategies. Computational overhead presents a barrier to real-time deployment in high-throughput clinical settings. Furthermore, language bias inherent in LLM pretraining data may negatively impact performance for underrepresented languages. While the model demonstrates strong performance on standard SNOMED CT terms (95% coverage), it exhibits limitations in handling rare or institution-specific terminology. Finally, using LLMs with sensitive EHR data necessitates stringent de-identification protocols, presenting an ongoing ethical and practical challenge.

To mitigate these challenges, several potential solutions are proposed. Hybrid approaches, integrating LLMs with rule-based systems or manual verification, may enhance reliability. Domain-specific fine-tuning could improve specificity, thereby reducing hallucinations. Expanding annotated corpora to include non-English languages and low-resource clinical datasets would strengthen generalizability. Optimization of clinical workflows, through techniques such as model distillation or hybrid rule-based/LLMs architectures, could reduce computational costs and enhance scalability. Integration with structured EHR data, such as laboratory results, may further enhance normalization accuracy. Finally, the incorporation of active learning, with clinician feedback loops, could refine the model's ability to address ambiguous or evolving medical terminology.

In conclusion, these studies highlight the potential of LLM-driven MCNs pipelines that incorporate both text cleansing and re-ranking for improved term normalization. While the dual application of LLMs significantly enhances precision and recall, challenges such as hallucinations, terminological consistency, and language-specific limitations must be carefully addressed. Ensuring robust validation mechanisms and optimizing integration strategies will be essential for the safe and effective implementation of LLM-based MCN in clinical practice.

2.3.3 Contribution

The key contributions include:

- LLMs standardize clinical text, improving concept mapping by cleaning errors and normalizing terms via RAG without manual input.
- an unsupervised pipeline (LLM, SapBERT, SNOMED CT retrieval) converts unstructured notes to standardized concepts optimized for local medical vocabularies.

Research Question C: "How can LLMs and RAG mitigate noise (e.g., misspellings, abbreviations) to enhance MCN accuracy?"
Noise in clinical text, such as misspellings, ad-hoc abbreviations, and inconsistent phrasing, compromises the accuracy of MCN. Our LLM-enhanced pipeline addresses this issue through two key innovations: (1) text cleansing via the LLM, GPT-4, and (2) RAG to dynamically ground normalization in authoritative ontologies (e.g., SNOMED CT). The RAG framework retrieves relevant concepts during inference, improving the disambiguation of polysemous terms. This approach achieves a top-5 F1 score of 0.735, demonstrating its efficacy in noisy real-world settings. Beyond normalization, the pipeline establishes a common semantic layer that enhances interoperability across disparate EHR systems by resolving documentation variability. However, the computational overhead of RAG warrants optimization for large-scale deployment. These results position LLM based RAG as a promising paradigm for robust clinical information extraction, with implications for decision support and translational research.

2.4 Gain of Knowledge

2.4.1 Summarized Contribution

In summary, these three studies advance medical text processing by tackling critical challenges, including data efficiency, interpretability, and multilingual applicability. They highlight how advances in classical and modern ML techniques – ranging from unsupervised methods to LLMs – can work together to improve MCN, offering valuable insights for future research and clinical applications.

Key insights include the importance of domain-specific models like SapBERT, which outperform general-purpose models in clinical NLP tasks. Using unsupervised and semi-supervised techniques, such as bi-encoders and LLMs for text cleansing, helps reduce reliance on labeled data, making MCN more efficient. Multilingual capabilities and the integration of standardized medical terminologies such as SNOMED CT help ensure consistency in medical concepts across languages and healthcare systems. Additionally, LLM-based text preprocessing refines clinical narratives, improving concept mapping accuracy. While deep learning models have proven highly effective, classical ML approaches remain valuable, particularly for their computational efficiency and interpretability when handling large-scale datasets, and can be especially suitable for specific data types and tasks. These studies emphasize the need for specialized, efficient, and scalable NLP models to enhance clinical workflows and improve healthcare delivery.

Building on these foundational studies, my subsequent publications further extend this framework, exploring complexities in clinical language and demonstrating how ML techniques address these challenges. They emphasize the role of hybrid approaches that balance contextualized and non-contextualized embeddings for optimal outcomes.

2.4.2 Supportive Investigations

In the first project [80], we explored contextualized and non-contextualized word embeddings for MCN using a k-nearest neighbour (k-NN) approach to map clinical terms to SNOMED CT. The non-contextualized surface term embeddings outperformed contextualized ones in this bi-encoder setting, achieving an F1 score of 0.853 compared to 0.322, showing the sensitivity of the exploited approach. This result suggests that the absence of sufficiently detailed context for many SNOMED CT target concepts limits the performance of contextual embeddings in large terminologies. Despite the promise of transformer-based models such as BERT, they struggled in this case due to the lack of detailed contextual information in SNOMED CT's vast vocabulary. This emphasizes the need for further refinement by enhancing SNOMED CT concepts with real-world clinical context. The error analysis revealed typical issues, such as analogy and granularity errors, underscoring the need for better entity recognition and normalization techniques to handle clinical language variations.

Further research [81] has explored clustering similar diagnosis terms using a string similarity heuristic, as explored in recent work on clustering syntactic variants in large clinical diagnosis lists. This method, which leverages Levenshtein distance [82] and incorporates pair-wise substring expansions, outperformed a deep learning-based approach, achieving a maximum F1 score of 0.71. Such clustering techniques can improve MCN and refine the entity recognition process in EHRs. In the next phase of the research [83], a bi-encoder and cross-encoder re-ranking model was used to map smoking-related terms to SNOMED CT. The bi-encoder alone achieved a Recall@1 performance of 33.49%, but the cross-encoder significantly improved this to 85.10%. This increase highlights the importance of considering the interaction between the input text and candidate concept pairs. While our previous study [84] using a supervised classification approach for smoking status achieved an F1 score of 0.97, the mentioned project, though slightly lower in performance than dataset-specific classifiers, demonstrated its applicability within a bi-encoder re-ranker MCN processing chain.

Visualization is crucial for interpreting MCN results across ML models, ensuring clarity in complex textual data. Common charts, like bar and line graphs, highlight key metrics, but visualization's broader role extends beyond basic comparisons. One such extension was the procedural coding support portal, which automates procedure code predictions from clinicians' surgery notes [85]. This project involved training models such as medBERT.de [86], surgeryBERT.at, and SVMs, to classify procedure codes based on surgery reports. Built using Streamlit [87], the portal allows users to input reports and receive the top five predicted codes. Predictions are explained with LIME, highlighting the textual features behind each code (see Figure 2.4). Bar charts display prediction probabilities, supporting model interpretation. This translation portal emphasizes XAI, ensuring clinicians can easily understand and trust the ML results. Finally, FusionViewer [88] is an interactive framework designed to unify clinical data, including EHRs, digital pathology, and multi-omic studies. It provides oncologists with accessible, structured representations of diverse datasets through interactive timelines and keyword-searchable annotations (see Figure 2.5). Leveraging NLP and ML, FusionViewer integrates textual understanding with visual analytics to offer a comprehensive view of patient health. Although still in development, it aims to have strong potential to support precision medicine and clinical decision-making. Future work will focus on optimizing data integration to streamline oncologists' workflows and enhance patient care.

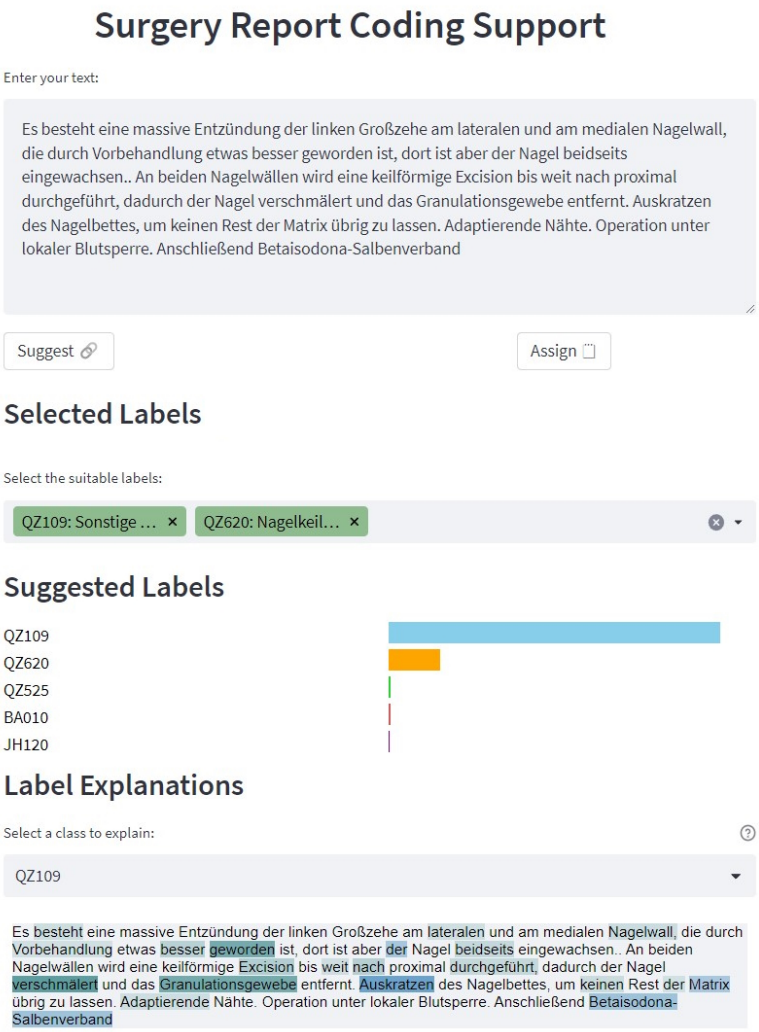

Figure 2.4: Screenshot of the graphical user interface, illustrating the surgery report coding support system with text input and label selection features [85].

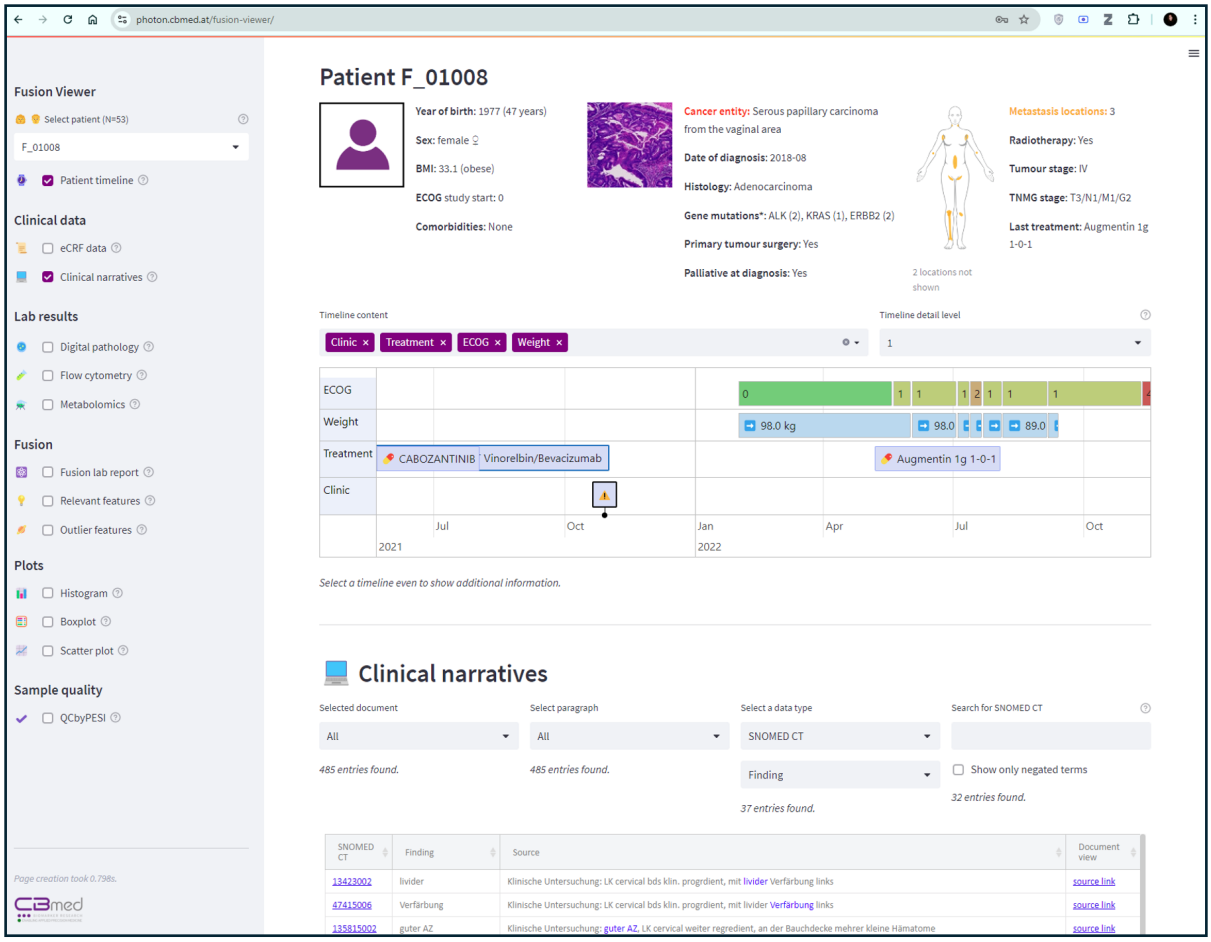

Figure 2.5: FusionViewer patient screen: Sidebar for selecting patient data (clinical, laboratory, fusion, graphical, sample quality). The main view shows the patient profile with a timeline and key navigation features.

Chapter 3

Conclusion and Outlook

This dissertation investigates computational semantics for intelligent digital health applications. Its focus lies on the optimization of the processing of clinical narratives, with the goal to obtain a structured and standardized rendering of their semi-structured and non-standardized representation using existing terminological standards. Methods of processing clinical language by leveraging machine learning (ML), together with language models, were developed and assessed in this context.

The three core publications in this dissertation underscore the transformative impact of these technologies. The first study [77] demonstrated the effectiveness of ML models in identifying oxygen supplementation data within electronic health records (EHRs), with Local Interpretable Model-Agnostic Explanations (LIME) enhancing interpretability. While classical ML models remained competitive, they required lower computational resources. The second study [78] employed an unsupervised SapBERT-based bi-encoder for medical concept annotation (MCA), which reduces reliance on labeled data while improving multilingual interoperability. The third study [79] extended medical concept normalization (MCN) using large language models (LLMs), achieving high F1 scores. Generative Pre-trained Transformer-4 (GPT-4) proved effective in transforming clinical jargon into standardized text, though challenges like hallucinations and linguistic limitations remain.

Further research built on these foundational studies, by examining contextualized and non-contextualized surface term embeddings [80], revealing the need of exemplifications

of SNOMECT CT concept terms in context, which is challenging except a focused value set selection. Bi-encoder and cross-encoder architectures enhanced the accuracy when mapping clinical text to SNOMED CT [83], while interactive platforms such as the procedural coding support portal [85] and FusionViewer [88] demonstrated the potential of integrating diverse data from EHRs, digital pathology, and multi-omics into user-friendly visualization frameworks.

It is expected that in the future, clinical artificial intelligence (AI) will prioritize cross-linguistic applicability, improved computational efficiency, and mitigating challenges such as false positives and hallucinations. Advancements in relation extraction, knowledge graph development, and unified visualization frameworks will further streamline clinical workflows, while real-time, actionable insights will support precision medicine through data-driven decision-making. Multimodal data integration will be central to future AI systems, offering comprehensive patient data views to personalize treatment and enhance outcomes by optimized clinical decision support. As these technologies evolve, explainable AI (XAI) will remain crucial for clinician trust and adoption, ensuring future clinical AI solutions are transparent, user-friendly, and efficient in improving both workflows and patient care.

In conclusion, this dissertation highlights the transformative potential of AI technologies and semantic resources in improving clinical workflows and patient care. While deep learning and LLMs offer adaptability, classical ML remains valuable for its simplicity and efficiency. Integrating unsupervised techniques, XAI, and data visualization establishes a robust framework for clinical data processing and MCN, ensuring interoperability and reusability. Future research will refine these models, enhance computational performance, and expand multimodal data integration to further personalize treatments and improve patient outcomes.

Bibliography

[1] Jiang J., Qi K., Bai G., and Schulman K. Pre-pandemic assessment: a decade of progress in electronic health record adoption among us hospitals. *Health Affairs Scholar*, 1(5):qxad056, 2023.

[2] Kim M. K., Rouphael C., McMichael J., Welch N., and Dasarathy S. Challenges in and opportunities for electronic health record-based data analysis and interpretation. *Gut and Liver*, 18(2):201, 2023.

[3] Lin A. Y., Arabandi S., Beale T., Duncan W. D., Hicks A., Hogan W. R., Jensen M., Koppel R., Martínez-Costa C., Nytrø Ø., and others . Improving the quality and utility of electronic health record data through ontologies. *Standards*, 3(3):316–340, 2023.

[4] Vuokko R., Vakkuri A., and Palojoki S. Systematized nomenclature of medicine–clinical terminology (snomed ct) clinical use cases in the context of electronic health record systems: systematic literature review. *JMIR medical informatics*, 11:e43750, 2023.

[5] Wilkinson M. D., Dumontier M., Aalbersberg I. J., Appleton G., Axton M., Baak A., Blomberg N., Boiten J.-W., Silva Santos da L. B., Bourne P. E., and others . The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3 (1):1–9, 2016.

[6] Negro-Calduch E., Azzopardi-Muscat N., Krishnamurthy R. S., and Novillo-Ortiz D. Technological progress in electronic health record system optimization: Systematic review of systematic literature reviews. *International journal of medical informatics*, 152:104507, 2021.

[7] West V. L., Borland D., and Hammond W. E. Innovative information visualization of electronic health record data: a systematic review. *Journal of the American Medical Informatics Association*, 22(2):330–339, 2015.

[8] Cowie M. R., Blomster J. I., Curtis L. H., Duclaux S., Ford I., Fritz F., Goldman S., Janmohamed S., Kreuzer J., Leenay M., and others . Electronic health records to facilitate clinical research. *Clinical Research in Cardiology*, 106:1–9, 2017.

[9] Kim E., Rubinstein S. M., Nead K. T., Wojcieszynski A. P., Gabriel P. E., and Warner J. L. The evolving use of electronic health records (ehr) for research. *Seminars in Radiation Oncology*, 29(4):354–361, 2019.

[10] Adeniyi A. O., Arowoogun J. O., Chidi R., Okolo C. A., and Babawarun O. The impact of electronic health records on patient care and outcomes: A comprehensive review. *World Journal of Advanced Research and Reviews*, 21(2):1446–1455, 2024.

[11] White A. and Danis M. Enhancing patient-centered communication and collaboration by using the electronic health record in the examination room. *Jama*, 309(22):2327–2328, 2013.

[12] Sreenivasan M. and Chacko A. M. Interoperability issues in ehr systems: Research directions. *Data analytics in biomedical engineering and healthcare*, pages 13–28, 2021.

[13] Sarwar T., Seifollahi S., Chan J., Zhang X., Aksakalli V., Hudson I., Verspoor K., and Cavedon L. The secondary use of electronic health records for data mining: Data characteristics and challenges. *ACM Computing Surveys (CSUR)*, 55(2):1–40, 2022.

[14] Honavar S. G. Electronic medical records–the good, the bad and the ugly, 2020.

[15] Ye J., Hai J., Song J., and Wang Z. The role of artificial intelligence in the application of the integrated electronic health records and patient-generated health data. *medRxiv*, pages 2024–05, 2024.

[16] Zhou L., DeAlmeida D., and Parmanto B. Applying a user-centered approach to building a mobile personal health record app: development and usability study. *JMIR mHealth and uHealth*, 7(7):e13194, 2019.

[17] Goldstein B. A., Navar A. M., Pencina M. J., and Ioannidis J. P. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association: JAMIA*, 24(1):198, 2017.

[18] Mello de B. H., Rigo S. J., Costa da C. A., Rosa Righi da R., Donida B., Bez M. R., and Schunke L. C. Semantic interoperability in health records standards: a systematic literature review. *Health and technology*, 12(2):255–272, 2022.

[19] Keshta I. and Odeh A. Security and privacy of electronic health records: Concerns and challenges. *Egyptian Informatics Journal*, 22(2):177–183, 2021.

[20] 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, 2024.

[21] Tsai C. H., Eghdam A., Davoody N., Wright G., Flowerday S., and Koch S. Effects of electronic health record implementation and barriers to adoption and use: a scoping review and qualitative analysis of the content. *Life*, 10(12):327, 2020.

[22] Hsu W., Taira R. K., El-Saden S., Kangarloo H., and Bui A. A. Context-based electronic health record: toward patient specific healthcare. *IEEE Transactions on information technology in biomedicine*, 16(2):228–234, 2012.

[23] Gianfrancesco M. A., Tamang S., Yazdany J., and Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, 178(11):1544–1547, 2018.

[24] Schulz S., Daumke P., Romacker M., and López-García P. Representing oncology in datasets: standard or custom biomedical terminology? *Informatics in Medicine Unlocked*, 15:100186, 2019.

[25] De Quirós F. G. B., Otero C., and Luna D. Terminology services: standard terminologies to control health vocabulary. *Yearbook of medical informatics*, 27(01): 227–233, 2018.

[26] Chatterjee A., Pahari N., and Prinz A. Hl7 fhir with snomed-ct to achieve semantic and structural interoperability in personal health data: a proof-of-concept study. *Sensors*, 22(10):3756, 2022.

[27] Belenkaya R., Gurley M. J., Golozar A., Dymshyts D., Miller R. T., Williams A. E., Ratwani S., Siapos A., Korsik V., Warner J., and others . Extending the omop common data model and standardized vocabularies to support observational cancer research. *JCO Clinical Cancer Informatics*, 5, 2021.

[28] Haug P. J., Narus S. P., Bledsoe J., and others . Promoting national and international standards to build interoperable clinical applications. In *AMIA Annual Symposium Proceedings*, volume 2018, page 555. American Medical Informatics Association, 2018.

[29] Schulz S., Del-Pinto W., Han L., Kreuzthaler M., Aghaei S., and Nenadic G. Towards principles of ontology-based annotation of clinical narratives. In *Proceedings of the International Conference on Biomedical Ontologies*, volume 2023, 2023.

[30] Wieland-Jorna Y., Kooten van D., Verheij R. A., Man de Y., Francke A. L., and Oosterveld-Vlug M. G. Natural language processing systems for extracting information from electronic health records about activities of daily living. a systematic review. *JAMIA open*, 7(2):ooae044, 2024.

[31] Yang X., Chen A., PourNejatian N., Shin H. C., Smith K. E., Parisien C., Compas C., Martin C., Costa A. B., Flores M. G., and others . A large language model for electronic health records. *NPJ digital medicine*, 5(1):194, 2022.

[32] Juhn Y. and Liu H. Artificial intelligence approaches using natural language processing to advance ehr-based clinical research. *Journal of Allergy and Clinical Immunology*, 145(2):463–469, 2020.

[33] Desai R. J., Wang S. V., Vaduganathan M., Evers T., and Schneeweiss S. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. *JAMA network open*, 3(1):e1918962–e1918962, 2020.

[34] Zhang X., Xiao J., and Gu F. Applying support vector machine to electronic health records for cancer classification. In *2019 Spring Simulation Conference (SpringSim)*, pages 1–9. IEEE, 2019.

[35] DuBrava S., Mardekian J., Sadosky A., Bienen E. J., Parsons B., Hopps M., and Markman J. Using random forest models to identify correlates of a diabetic peripheral

neuropathy diagnosis from electronic health record data. *Pain Medicine*, 18(1):107–115, 2017.

[36] Duan R., Boland M. R., Moore J. H., and Chen Y. Odal: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites. In *BIOCOMPUTING 2019: Proceedings of the Pacific Symposium*, pages 30–41. World Scientific, 2018.

[37] Andry J. F., Silaen F. M., Tannady H., and Saputra K. H. Electronic health record to predict a heart attack used data mining with naïve bayes method. *Int J Inf & Commun Technol ISSN*, 2252(8776):8776, 2021.

[38] Latif J., Xiao C., Tu S., Rehman S. U., Imran A., and Bilal A. Implementation and use of disease diagnosis systems for electronic medical records based on machine learning: A complete review. *IEEE Access*, 8:150489–150513, 2020.

[39] Miotto R., Wang F., Wang S., Jiang X., and Dudley J. T. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

[40] Solares J. R. A., Raimondi F. E. D., Zhu Y., Rahimian F., Canoy D., Tran J., Gomes A. C. P., Payberah A. H., Zottoli M., Nazarzadeh M., and others . Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *Journal of biomedical informatics*, 101:103337, 2020.

[41] Suo Q., Ma F., Yuan Y., Huai M., Zhong W., Zhang A., and Gao J. Personalized disease prediction using a cnn-based similarity learning method. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 811–816. IEEE, 2017.

[42] Rasmy L., Wu Y., Wang N., Geng X., Zheng W. J., Wang F., Wu H., Xu H., and Zhi D. A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous ehr data set. *Journal of biomedical informatics*, 84:11–16, 2018.

[43] Hochreiter S. Long short-term memory. *Neural Computation MIT-Press*, 1997.

[44] Beck M., Pöppel K., Spanring M., Auer A., Prudnikova O., Kopp M., Klambauer G., Brandstetter J., and Hochreiter S. xlstm: Extended long short-term memory. *arXiv preprint arXiv:2405.04517*, 2024.

[45] Guo A., Beheshti R., Khan Y. M., Langabeer J. R., and Foraker R. E. Predicting cardiovascular health trajectories in time-series electronic health records with lstm models. *BMC medical informatics and decision making*, 21:1–10, 2021.

[46] Shi J., Ye M., Chen H., Lu Y., Tan Z., Fan Z., and Zhao J. Enhancing efficiency and capacity of telehealth services with intelligent triage: a bidirectional lstm neural network model employing character embedding. *BMC Medical Informatics and Decision Making*, 23(1):269, 2023.

[47] Vaswani A. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[48] Goodrum H., Roberts K., and Bernstam E. V. Automatic classification of scanned electronic health record documents. *International journal of medical informatics*, 144:104302, 2020.

[49] Mitra A., Rawat B. P. S., McManus D. D., Yu H., and others . Relation classification for bleeding events from electronic health records using deep learning systems: an empirical study. *JMIR medical informatics*, 9(7):e27527, 2021.

[50] Al M. L., Eric A., and Nishant R. Enriching electronic health record with semantic features utilisingpretrained transformers. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 151–161, 2023.

[51] Rasmy L., Xiang Y., Xie Z., Tao C., and Zhi D. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.

[52] Liu F., Shareghi E., Meng Z., Basaldella M., and Collier N. Self-alignment pretraining for biomedical entity representations. In Toutanova K., Rumshisky A., Zettlemoyer L., Hakkani-Tur D., Beltagy I., Bethard S., Cotterell R., Chakraborty T., and Zhou Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages

4228–4238, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.334. URL https://aclanthology.org/2021.naacl-main.334.

[53] Hu Z., Wang L., Lan Y., Xu W., Lim E.-P., Bing L., Xu X., Poria S., and Lee R. K.-W. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2304.01933*, 2023.

[54] 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 (HEALTH)*, 3(1):1–23, 2021.

[55] Ahsan H., McInerney D. J., Kim J., Potter C., Young G., Amir S., and Wallace B. C. Retrieving evidence from ehrs with llms: Possibilities and challenges. *Proceedings of machine learning research*, 248:489, 2024.

[56] AlSaad R., Abd-Alrazaq A., Boughorbel S., Ahmed A., Renault M.-A., Damseh R., and Sheikh J. Multimodal large language models in health care: Applications, challenges, and future outlook. *Journal of Medical Internet Research*, 26:e59505, 2024.

[57] Yao Y., Duan J., Xu K., Cai Y., Sun Z., and Zhang Y. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211, 2024.

[58] Bjöörn A. Employing a transformer language model for information retrieval and document classification: Using openai's generative pre-trained transformer, gpt-2, 2020.

[59] Floridi L. and Chiriatti M. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

[60] Kalyan K. S. A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, page 100048, 2023.

[61] Schopow N., Osterhoff G., and Baur D. Applications of the natural language processing tool chatgpt in clinical practice: comparative study and augmented systematic review. *JMIR Medical Informatics*, 11:e48933, 2023.

[62] Chowdhery A., Narang S., Devlin J., Bosma M., Mishra G., Roberts A., Barham P., Chung H. W., Sutton C., Gehrmann S., and others . Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[63] 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*, page ocae045, 2024.

[64] Mastropaolo A., Scalabrino S., Cooper N., Palacio D. N., Poshyvanyk D., Oliveto R., and Bavota G. Studying the usage of text-to-text transfer transformer to support code-related tasks. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 336–347. IEEE, 2021.

[65] Arslan E. and Harinda E. Innovating sql automation: Evaluating open-source large language models with a dual-stage approach for corporate data solutions. In *2024 9th International Conference on Computer Science and Engineering (UBMK)*, pages 68–73. IEEE, 2024.

[66] Swinckels L., Bennis F. C., Ziesemer K. A., Scheerman J. F., Bijwaard H., Keijzer de A., and Bruers J. J. The use of deep learning and machine learning on longitudinal electronic health records for the early detection and prevention of diseases: scoping review. *Journal of Medical Internet Research*, 26:e48320, 2024.

[67] Yun H. S., Pogrebitskiy D., Marshall I. J., and Wallace B. C. Automatically extracting numerical results from randomized controlled trials with large language models. *arXiv preprint arXiv:2405.01686*, 2024.

[68] Hak F., Guimarães T., and Santos M. Towards effective clinical decision support systems: A systematic review. *PLoS One*, 17(8):e0272846, 2022.

[69] Adlung L., Cohen Y., Mor U., and Elinav E. Machine learning in clinical decision making. *Med*, 2(6):642–665, 2021.

[70] Li J., Sun A., Han J., and Li C. A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering*, 34(1):50–70, 2020.

[71] Song B., Li F., Liu Y., and Zeng X. Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison. *Briefings in Bioinformatics*, 22(6):bbab282, 2021.

[72] Xu D., Gopale M., Zhang J., Brown K., Begoli E., and Bethard S. Unified medical language system resources improve sieve-based generation and bidirectional encoder representations from transformers (bert)–based ranking for concept normalization. *Journal of the American Medical Informatics Association*, 27(10):1510–1519, 2020.

[73] Vashishth S., Newman-Griffis D., Joshi R., Dutt R., and Rosé C. P. Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets. *Journal of biomedical informatics*, 121:103880, 2021.

[74] Xu D., Zhang Z., and Bethard S. A generate-and-rank framework with semantic type regularization for biomedical concept normalization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8452–8464, 2020.

[75] Yuan Z., Zhao Z., Sun H., Li J., Wang F., and Yu S. Coder: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of biomedical informatics*, 126:103983, 2022.

[76] Harnoune A., Rhanoui M., Mikram M., Yousfi S., Elkaimbillah Z., and El Asri B. Bert based clinical knowledge extraction for biomedical knowledge graph construction and analysis. *Computer Methods and Programs in Biomedicine Update*, 1:100042, 2021.

[77] Abdulnazar A., Kugic A., Schulz S., Stadlbauer V., and Kreuzthaler M. O2 supplementation disambiguation in clinical narratives to support retrospective covid-19 studies. *BMC Medical Informatics and Decision Making*, 24(1):29, 2024.

[78] Abdulnazar A., Roller R., Schulz S., and Kreuzthaler M. Unsupervised sapbert-based bi-encoders for medical concept annotation of clinical narratives with snomed ct. *Digital Health*, 10:20552076241288681, 2024.

[79] Abdulnazar A., Roller R., Schulz S., and Kreuzthaler M. Large language models for clinical text cleansing enhance medical concept normalization. *IEEE Access*, 2024.

[80] Abdulnazar A., Kreuzthaler M., Roller R., and Schulz S. Sapbert-based medical concept normalization using snomed ct. In *Caring is Sharing–Exploiting the Value in Data for Health and Innovation*, pages 825–826. IOS Press, 2023.

[81] Schulz S., Abdulnazar A., and Kreuzthaler M. Clustering similar diagnosis terms. In *MIE*, pages 837–838, 2023.

[82] Yujian L. and Bo L. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095, 2007.

[83] Abdulnazar A., Schulz S., and Kreuzthaler M. Smoking status normalization with cross-encoders and snomed ct. In *Proceedings of Medical Informatics Europe (MIE) 2025*, Glasgow, Scotland, 2025. European Federation for Medical Informatics (EFMI). Accepted for publication.

[84] 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 *International Conference on Artificial Intelligence in Medicine*, pages 182–191, 2024.

[85] Veeranki S. P. K., Abdulnazar A., Kramer D., Kreuzthaler M., and Lumenta D. B. Multi-label text classification via secondary use of large clinical real-world data sets. *Scientific Reports*, 14(1):26972, 2024.

[86] 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., and others . Medbert. de: A comprehensive german bert model for the medical domain. *Expert Systems with Applications*, 237:121598, 2024.

[87] Richards T. *Streamlit for Data Science: Create interactive data apps in Python*. Packt Publishing Ltd, 2023.

[88] Abdulnazar A., Kreuzthaler M., Schulz S., Prietl B., and Herbsthofer L. Tumor board visualization: Integrating clinical and laboratory insights. In *Digital Health and Informatics Innovations for Sustainable Health Care Systems*, pages 1750–1751. IOS Press, 2024.

Core Publications

Publication 1
Abdulnazar *et al.*
BMC Medical Informatics and Decision Making (2024) 24:29
https://doi.org/10.1186/s12911-024-02425-2

BMC Medical Informatics and Decision Making

RESEARCH Open Access

O2 supplementation disambiguation in clinical narratives to support retrospective COVID-19 studies

Akhila Abdulnazar[1,2], Amila Kugic[1], Stefan Schulz[1], Vanessa Stadlbauer[2,3] and Markus Kreuzthaler[1*]

Abstract

Background Oxygen saturation, a key indicator of COVID-19 severity, poses challenges, especially in cases of silent hypoxemia. Electronic health records (EHRs) often contain supplemental oxygen information within clinical narratives. Streamlining patient identification based on oxygen levels is crucial for COVID-19 research, underscoring the need for automated classifiers in discharge summaries to ease the manual review burden on physicians.

Method We analysed text lines extracted from anonymised COVID-19 patient discharge summaries in German to perform a binary classification task, differentiating patients who received oxygen supplementation and those who did not. Various machine learning (ML) algorithms, including classical ML to deep learning (DL) models, were compared. Classifier decisions were explained using Local Interpretable Model-agnostic Explanations (LIME), which visualize the model decisions.

Result Classical ML to DL models achieved comparable performance in classification, with an F-measure varying between 0.942 and 0.955, whereas the classical ML approaches were faster. Visualisation of embedding representation of input data reveals notable variations in the encoding patterns between classic and DL encoders. Furthermore, LIME explanations provide insights into the most relevant features at token level that contribute to these observed differences.

Conclusion Despite a general tendency towards deep learning, these use cases show that classical approaches yield comparable results at lower computational cost. Model prediction explanations using LIME in textual and visual layouts provided a qualitative explanation for the model performance.

Keywords Natural language processing, Machine learning, Deep learning, Electronic health records, COVID-19

Background

In January 2020, the World Health Organisation declared a global health emergency based on growing case reports of the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1], leading to the outbreak of Coronavirus disease (COVID-19). The COVID-19 pandemic has created a widespread impact all over the world, with 700 million reported cases and 6 million estimated deaths [2] by late 2023. Up until now, an up-to-date picture of the clinical situation, capable of comparing patient data for a better understanding of all aspects of the

*Correspondence:
Markus Kreuzthaler
markus.kreuzthaler@medunigraz.at
[1] Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, 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

disease, has however been impaired by the lack of access to patient data and their lack of standardization, particularly when locked within narrative EHR (electronic health record) content. The ongoing practice of documenting even crucial facts about critically ill patients as free text is a major barrier to the adoption of novel information extraction methods for health care and research. The manual extraction of specific information, such as diagnoses, symptoms, medications, dates, and patient demographics from clinical narratives is a time-consuming and tiresome process. It would have to be done by clinicians familiar with the domain, who would be urgently needed for healthcare delivery in a pandemic context. This motivates the importance of computerised methods to interpret clinical narratives and to extract structured and meaningful information. Text classification with Natural Language Processing (NLP) has reduced the manual time required for analysing clinical text data. However, it is essential to customise the components to fit the specific use case in advance.

Classical and deep learning approaches

Text Classification. A comprehensive analysis of text classification models, spanning classical to DL approaches, highlights the advantages of DL in automatically generating meaningful representations for text mining. However, it also acknowledges limitations, such as neglecting natural sequential and contextual information [3, 4]. Classical ML approaches such as Support Vector Machines (SVMs) have the advantage that their off-the-shelf implementations can not only be trained much faster when compared to deep neural networks [5] but have also an overall better runtime performance. Disambiguation of clinical abbreviations is another essential information extraction task, due to their abundance in clinical narratives, as demonstrated by Jaber and Martínez [6], who used a one-fits-all classifier based on deep learning (DL) models. Many other studies demonstrated the benefit of classical to deep ML algorithms in various healthcare use cases [7–11].

Machine learning for COVID-19. COVID-19-related information extraction has covered a broad range of methods, from classical ML to DL models. Daher et al. [12] elaborated on the requirement for supplemental oxygen for admitted patients. Several research works predicted the requirements of oxygen and oxygen therapies in COVID-19 patients using ML approaches [13–16]. Prediction of COVID-19 mortality rates used gradient boosting [17], decision trees [18], artificial neural networks [19] and DL models [20]. Additionally, studies on severity score prediction [21] leveraged explainable artificial intelligence (XAI) approaches [22].

Several factors that contribute to advantages of classical ML approaches are (i) data size and complexity because DL models generally require large amounts of data to learn complex hierarchical representations; (ii) intensive computational resources required for DL models, (iii) the tendency of DL towards overfitting, especially when the dataset is small, and finally (iv) problem-specific considerations, (e.g., scalability, noise and outliers, ethical constraints, user requirements, etc.) can influence the performance of different models.

O2 saturation in EHRs

A precise understanding of how oxygenation information is recorded in EHRs is essential in retrospective COVID-19 studies. The details of the fraction of inspired oxygen (FiO2), partial pressure of oxygen (PaO2/PO2) and arterial oxygen saturation (SaO2) require attention. FiO2 is 0.21 in room air and increases with supplemental oxygen [23]. PaO2 is sensitive but lacks specificity for gas exchange. The sigmoid oxygen dissociation curve relates PaO2 and SaO2, representing haemoglobin oxygen saturation. Interpreting clinical data, including FiO2, PaO2, and SaO2, is complicated, so accurately extracting information from narratives requires distinguishing and harmonizing related terms and conflicting results [24]. One of the primary challenges lies in the diversity of medical records and the multitude of abbreviations employed, which are often context-dependent and vary across institutions and even between clinicians. The same abbreviation may carry different meanings in distinct contexts and settings. Information extraction systems therefore need to take this ambiguity into account, as well as the existence of synonyms, variants and typos in clinical texts.

The focus of our work is on the use of unstructured data on oxygen status and supplementation of COVID-19 patients. The supply of the organism with oxygen is of vital importance, particularly in case of respiratory infection. Current practices involve monitoring of PaO2 and SaO2 using pulse oximetry as a common non-invasive tool [25]. A spontaneous fall in oxygen saturation levels, known as "silent hypoxemia", is cardinal because low oxygen levels indicate the severity of the disease and predict poor outcomes [26]. Peripheral oxygen saturation (SpO2) determines whether room air oxygen is no longer sufficient and supplementation of oxygen via masks, nasal cannulas or ventilators is required, which often requires intensive care treatment [27]. SpO2/FiO2 ratio is a reliable tool for hypoxemia screening among patients admitted to the emergency departments, particularly during the SARS-CoV-2 outbreak [23]. In addition, this paper addresses the problem of the lack of structured oxygen status data. The problem is complicated by the fact that

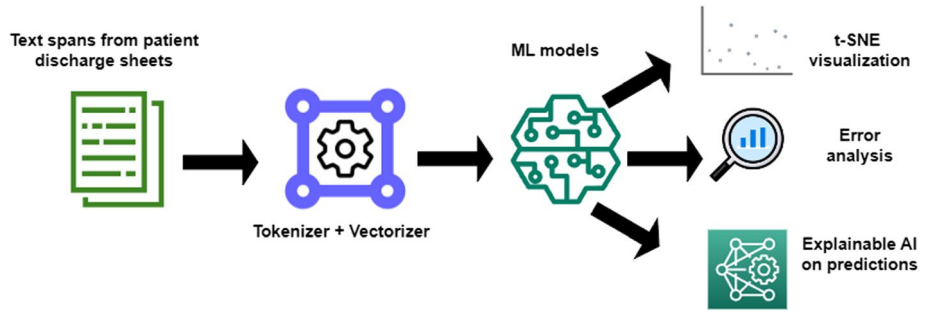

Fig. 1 Overview of the proposed methodology encompassing text lines preprocessing, binary text classification, t-SNE visualization, error analysis, and LIME explanation

one and the same concept, *viz.* oxygen, is on the one hand mentioned as a status variable of the patient and a result of measurement, but on the other hand *supplemental* oxygen is referred to as a treatment administered to the patient. Thus, the word "oxygen" may refer to supplementary oxygen treatment as well as to the measurement of SpO2. Differentiation of the oxygen status was based on the measurements of the supplemental oxygen or indicated features for the supplemental oxygen demand and those without any further information regarding the oxygen requirement. In this investigation, the interpretation of whether the reported oxygen status is with or without the supply of oxygen should be done via an adapted model-based approach, as described in this manuscript. This system should support an expert data curator in the identification of relevant document parts to be processed in the next step. Figure 1 represents the flowchart of the proposed method. In addition to the recognition of mentions of oxygen supplementation, this work adds functionality for data visualisation and a methodology for ML model explainability.

The paper is organised as follows: Materials and methods section describes the data and the different types of classifiers used, Results section compares classifier performances and the computational time needed, along with error analysis and model explanation. The Discussion section compares the results with related work and discusses false positives and false negatives.

Materials and methods

Dataset

Text lines from discharge summaries of patients affected by COVID-19 were collected from the EHR system of KAGes, an Austrian network of public hospitals. Text lines up to a length of 30 characters, denoting potential oxygen status information, are extracted to build the dataset using a regular expression[1]. This expression was created in several iterations, supported by a data science specialist experienced in clinical queries. The binary classification task is formulated as follows: (i) there is evidence that the patient got oxygen supplementation at some time during the hospital stay, vs. (ii) there is no evidence that the patient received oxygen supplementation.

Gold standard creation

The dataset contained 3,844 anonymised text lines. These were annotated by two annotators independently. Both annotators had biomedical backgrounds, were supported by a guideline and passed a series of training sessions. The third annotator with medical expertise validated the annotations so that they could be used as the ground truth. The inter-annotator agreement was high, as evidenced by a Cohen's Kappa [28] of 0.859, which indicates a 94% accuracy.

The annotation was based on specific text features for example, "*l/min O2 über Nasenbrille*" (litre per minute oxygen via nasal cannula), "*Sauerstoffbedarf*" (oxygen requirement), "*unter CPAP*" (under continuous positive airway pressure), "*mit RL*" (with room air), "*mit NIV*" (with non-invasive ventilation), etc. Some typical text lines are shown along with their class assignments in Table 1.

Of the 3,844 anonymised text lines, 1,435 clearly described the use of supplemental oxygen at some point in time and were thus assigned to class "1". The remaining 2,409 text lines were assigned to class "0", of which 45 text lines did not provide any kind of information regarding supplemental oxygen. The dataset was split into training and test data, with a set of constant random state values and a test set size of 20 per cent. The training data consisted of 3,074 spans, with 769 spans in test data.

[1] `SO%|[sS][apP]?[O0o][2²](?!.)|bO2.?Sä|bO2(?!.*?pH)|[Ss]ättigung.`

Table 1 Text spans from the dataset along with their classes and relevant tokens translated. Class "1" means the use of supplemental oxygen at some point of time

Text span	Class	Translation of the relevant tokens
"SpO2 99% **mit 10l O2**"	1	"with 10l O2"
"RR 160/80mmHg, HF 76', Temp. 38,4°C, SpO2 89% **mit RL**"	0	"with room air"
"keine Dyspnoe, kein Fieber, keine Schmerzen, **kein O2 Bedarf**"	0	"no O2 requirement"
"**1L O2/min. über die Nasenbrille** respiratorisch völlig stabil."	1	"1L O2/min. via nasal cannula"
"O2 **2l/min** bei Bed."	1	"2l/min"

Finally, a division into training and testing sets was performed by the 'train_test_split' function from scikit-learn [29]. To ensure robust evaluation, multiple train-test splits using different random state values[2] were done. This process helped mitigate the impact of the initial randomization on model performance and assessed its generalisation ability.

Machine learning approaches

In ML, classification, in general, is a predictive modelling challenge, in which the model has to predict the category of the input data based on fitting of the training dataset. In particular, the classification of text is an elementary NLP task, which is applied wherever input data contain free text. NLP uses different types of ML methods. The following architectures have been applied for the comparative analysis, motivated by a comparison of popular core neural network architectures and their influence on model performance:

Model architectures

Support Vector Machine (SVM) SVM is used for both classification and regression tasks. It uses textual data, which is either represented as a vector or a token in a vector space. SVMs attempt to find a hyperplane that best divides the training data into corresponding classes [30]. We used the Support Vector Classifier (SVC) function from scikit-learn [29].

Random Forest (RF) RF is a classification algorithm based on the principles of decision trees. In RF, the set of attributes is randomly split into many subsets, each of which is used to construct decision trees with a few layers. These decision trees collectively form the 'forest'. The overall performance is then determined based on the outputs of each tree. This randomness in attribute selection and tree construction helps reduce overfitting and enhances the diversity of the trees in the forest. RF is therefore considered a robust and accurate machine learning algorithm [31]. We used the RandomForestClassifier function from scikit-learn.

Long Short-Term Memory (LSTM) LSTM [32] is a type of recurrent neural network. LSTM networks contain feed-forward networks along with corresponding feedback connections. This makes it distinguishable from other neural networks, as it processes sequences of data points. A single LSTM unit is known as a cell, which consists of an input, output and forget gate. These gates control the flow of data in and out of the cell, along with remembering essential information at random time intervals [33]. We used the Keras [34] library for implementing the model layers. Our model architecture consisted of (i) an embedding layer for word representation, (ii) an LSTM layer incorporating dropout for regularisation, and (iii) a dense layer with a sigmoid activation function to produce binary classification output.

Bidirectional Long Short-Term Memory (Bi-LSTM) Bi-LSTM networks consist of two LSTM networks, in which one feeds the data in a forward direction, while the other feeds the data in a backward direction [35]. We used Keras [34] library for implementing the model layers. Our architecture consisted of the following components: (i) an embedding layer for word representation, (ii) a Bi-LSTM layer incorporating dropout for regularisation, and (iii) a dense layer with a sigmoid activation function to produce binary classification output.

Convolutional Neural Network (CNN) CNN adaptively learns the different hierarchies of features through back propagation using different layers such as convolution layers, pooling layers and fully connected layers [36]. Even though convolutional networks were initially developed by the neural network image processing community where it excelled in recognising objects in predefined classes, it has recently shown excellent outcomes in NLP tasks, especially in sentence classification into predefined

[2] [509, 906, 331, 172, 729, 250, 762, 629, 926, 392]

categories. CNN and extended CNN architectures [37] have been successfully applied to text classification tasks of different granularity [38]. They also capture the neighbourhood relation via the window size of the CNN filter. We used the Keras [34] library for implementing the model layers. Our model architecture includes the following components: (i) an embedding layer for word representation, (ii) a 1-Dimensional convolutional layer with multiple filters and rectified linear unit (ReLU) activation, (iii) a global max pooling layer to capture relevant features and (iv) a dense layer with a sigmoid activation for binary classification.

Text preprocessing and representation

TF-IDF (Term Frequency-Inverse Document Frequency) vectorisation converts the text data into numerical features in SVM and RF. Text data preprocessing, without the removal of stop words, includes tokenisation and sequence padding. For LSTM, Bi-LSTM and CNN models, we tokenised the text using the Keras [34] tokenizer with a specified maximum word count. To ensure that we capture the full context without unnecessary truncation, we selected a maximum token length of 30. This choice is well-justified, as it allows us to handle all sequences within our dataset, thereby capturing the most of the context and information from each text entry, without disclosing any patient-specific information.

In order to enhance the representational capacity and capture intricate patterns in our data across various machine learning models, our classical ML models with TF-IDF vectorization has a dimension within a range of 2450 to 2520 for the applied random state values, and we choose 300-dimensional input vectors consistently in all our DL models, following current practice [39]. Unlike some deep learning models that have fixed dimensions, TF-IDF vectors adapt their dimensionality based on the dataset's linguistic diversity. TF-IDF captures word significance across documents, while fixed dimensions in deep learning aim for computational efficiency and concise representations.

Hyperparameter tuning

For each model, hyperparameter tuning is performed using grid search [40] and five-fold cross-validation. The goal is to find the optimal set of hyperparameters for each model. For SVM, various combinations of hyperparameters, including 'C' (regularization parameter), 'kernel' (kernel function), and 'gamma' (kernel coefficient) were tried. For RF, combinations of 'n_estimators' (the number of trees in the forest), 'max_depth' (the maximum depth of the trees), and 'max_features' (the number of features to consider when splitting nodes) were used. For LSTM and Bi-LSTM we combined the hyperparameters units, dropout and recurrent dropout. Finally, for CNN, hyperparameters, such as filters and kernel size were selected using grid search. In LSTM, Bi-LSTM and CNN models cross entropy was used as the loss function, adamax as the optimizer, early stopping as the stop criteria and run with an epoch value of ten. Model architectures with the best hyperparameters are summarised in Table 2.

Model assessment and selection

For each iteration through the random state values, we trained the classifiers with the best hyperparameters identified during the tuning phase. The performance was evaluated using precision, recall, and F1-score. We calculated the mean and standard deviation of these performance metrics across the different iterations to assess the overall model performance. The model with the highest F1 score was selected for further analysis.

Visualisation

To gain insights into the distribution of the data in a lower-dimensional space, we applied t-distributed Stochastic Neighbour Embedding (t-SNE) [41] to the TF-IDF vectors of the test data in SVM and RF. For the DL models, the word embeddings learned by the best model are extracted and visualised. The resulting 2D scatter plot visualises the data points based on their predicted labels ('y_test') and serves as an additional tool for understanding the model's behaviour. t-SNE is a technique commonly used to explore intricate patterns and relationships within complex datasets by projecting them into a lower-dimensional space. This reveals hidden insights not apparent in the original data. t-SNE visualisations also offer an intuitive way to comprehend model performance, decision boundaries, and data separability. These are particularly popular for visualising text data due to their ability to capture complex relationships in high-dimensional data, making t-SNE a preferred choice to linear techniques, like Principal Component Analysis (PCA). However, one must be aware of the limitations of t-SNE, such as sensitivity to the perplexity parameter and difficulty in interpreting distances in the reduced space.

Model explanation using LIME

After determining the most effective model, we perform feature relevance analysis to understand which terms or features have the greatest impact on the classification result. This analysis provides valuable insights into the key phrases or structures from which the model creates its predictions. To this end, we use a method called LIME [42], suited for predictions of complex black-box models. LIME starts by creating variations of the input text, involving actions like removing, replacing, or

Table 2 Parameters used in different machine learning models after grid search optimization

Classifier	Parameters - Values
SVM	vectorisor - TF-IDF vectorisor (in a range of 2450 to 2520 dimensions)
	kernel - rbf (Radial Basis Function) kernel
	regularisation parameter - C value of 10
	cross-validation - 5 fold
RF	vectorisor - TF-IDF vectorisor (in a range of 2450 to 2520 dimensions)
	maximum features- square root of the total number of features
	number of decision trees - 100
	cross-validation - 5 fold
LSTM & Bi-LSTM	Embedding layer - 300 dimensional
	LSTM / Bi-LSTM layer - 128 nodes
	dropout and recurrent dropoutlayer -probability of 0.2
	dense output layer - 1 node
	activation layer - sigmoid
	cross-validation - 5 fold
	loss - binary cross entropy
	optimizer - adamax
CNN	Embedding layer - 300 dimensional
	1D convolutional layer with:
	- filters - 256
	- window size - 5
	- activation layer - relu
	dropout layer - probability of 0.5
	dense output layer - 1 node
	activation layer - sigmoid
	cross-validation - 5 fold
	loss - binary cross entropy
	optimizer - adamax

rearranging tokens randomly. It then passes these variations through the model and records the resulting predictions. LIME selects a subset of token features from both the original input text and the variations, focusing on those that significantly influence prediction. A linear SVM is then built using these selected features, which helps estimate the model's behaviour regarding specific features. The feature importance weights calculated by this process clarify how the model's output class is determined, highlighting the most influential tokens for predicted class probabilities. Higher weights signify stronger contributions, while lower weights indicate less influence. The application of LIME on machine learning models assists non-experts in comprehending the internal processes of a model and tracking decision details related to predictions.

Results

Classifier results

The model is optimised using the training data, and the chosen hyperparameters via grid search are then implemented in the model and analysed in the test performance. Performance metrics for each model were calculated in terms of precision (P), recall (R) and F1-score (F1) as shown in Table 3, with the different classifiers producing comparable good results. We opt for the F1-score instead of ROC and AUC, because of its better handling of imbalanced data and its alignment with the substantial clinical impact of both false positives and false negatives.

Furthermore, classifier models were profiled with their computational speed assistance by identifying performance bottlenecks and enhancing the underlying hardware or software infrastructure to acquire faster execution times. The computational speed for each sample is estimated using the Python module "time"[3] to measure the model prediction time. We calculated the performance of classifier models using an AMD Ryzen7 5700U with Radeon Graphics processor with a clock frequency of 1.8 GHz and 8 GB of RAM. The experiments were

[3] https://docs.python.org/3/library/time.html

Table 3 Performance metrics for SVM, RF, LSTM, Bi-LSTM, and CNN models on the test data and their average prediction time per sample

Classifier	Metric	Mean ± Std Error	95% Confidence Intervals	Average Prediction Time (seconds)
SVM	P	0.955 ± 0.002	[0.951 – 0.959]	0.258
	R	0.955 ± 0.002	[0.951 – 0.959]	
	F1	0.955 ± 0.002	[0.951 – 0.959]	
RF	P	0.942 ± 0.003	[0.936 – 0.948]	0.057
	R	0.942 ± 0.003	[0.936 – 0.948]	
	F1	0.942 ± 0.003	[0.936 – 0.948]	
LSTM	P	0.948 ± 0.002	[0.944 – 0.952]	0.501
	R	0.948 ± 0.002	[0.944 – 0.952]	
	F1	0.948 ± 0.002	[0.944 – 0.952]	
Bi-LSTM	P	0.944 ± 0.003	[0.938 – 0.950]	0.502
	R	0.946 ± 0.003	[0.938 – 0.950]	
	F1	0.944 ± 0.003	[0.938 – 0.950]	
CNN	P	0.954 ± 0.002	[0.950 – 0.958]	0.130
	R	0.954 ± 0.002	[0.950 – 0.958]	
	F1	0.954 ± 0.002	[0.950 – 0.958]	

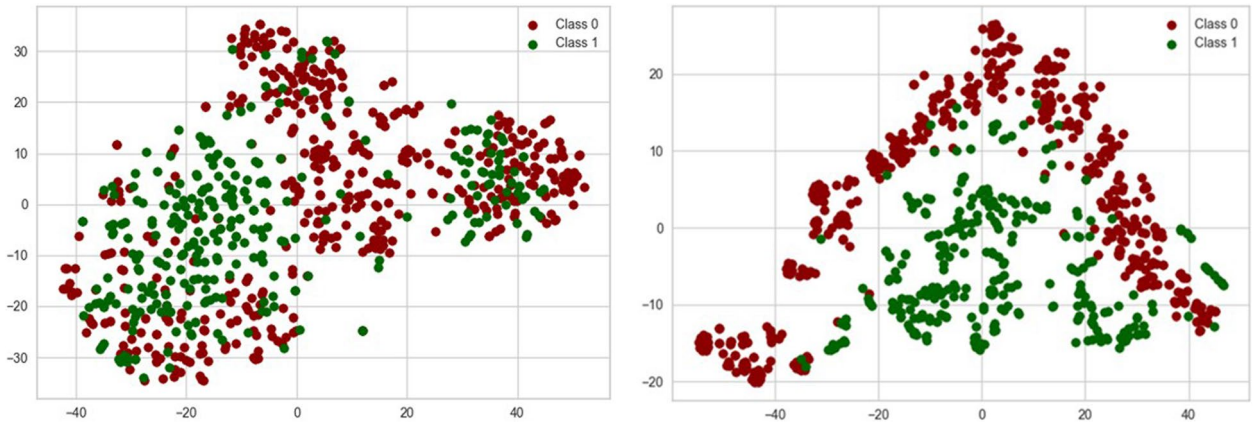

Fig. 2 Visualization of vector representations of test data using t-SNE. (i) Static TF-IDF weighted vectors with no clear separability among the classes and (ii) Dynamic embedding representation showing a better separability among the classes

conducted using Python 3.8.16 running on Windows 11. Table 3 lists the mean prediction time per sample for each of the best models.

t-SNE visualisation

The input representation for the SVM is a high dimensional vector based on token occurrence, which remains static during training and can lead to reduced separability in the dimension-reduced visualisation. While the embeddings in the CNN model adapt to the downstream task, optimising their representation for the domain-specific task, enabling separability in the visualization. The grouping into corresponding class clusters is therefore recognisable in the embedding case for CNN but less clear for the SVM representation. Figure 2 plots the visualisation of test data using the t-SNE method for the SVM and the CNN model. In summary, the t-SNE visualisation shows that the dynamic embedding approach of the CNN results in a better separability of data clusters compared to the static representation of SVM, highlighting the adaptability of neural networks in domain-specific tasks.

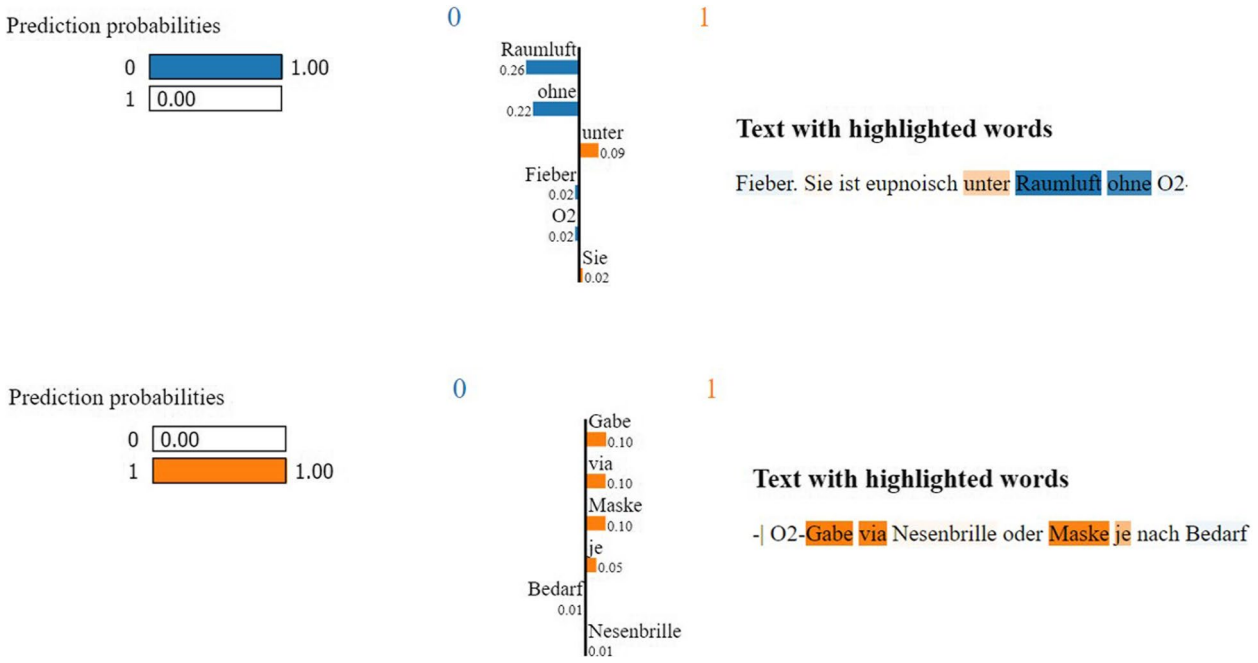

Fig. 3 LIME providing insights into model predictions by highlighting the key tokens influencing the classification decision

Table 4 Top influential tokens in the dataset for SVM and CNN models using LIME

Ranking	Class 0		Class 1	
	SVM	CNN	SVM	CNN
1	raumluft	kein/keine	l/L/Litre	fiO2/FiO2
2	akuter	ohne	fiO2/FiO2	l/L/Litre
3	Kein/kein/keinem	via	Flüssigsauerstofftherapie	Mit
4	ohne	raumluft	pflichtig	Gabe
5	auszugehen	AF	Zufuhr	O2
6	nicht	mmol	Gabe	Flow
7	Pulsoxy	nicht	inadäquaten	Sauerstoff
8	Aufsättigung	K	mehr	für
9	niedriger	zufuhr	Darunter	Brille
10	ausgeprägter	seit	Mit	wurde

LIME explanation

LIME explanations are generated to give insight into the model's prediction. Figure 3 illustrates the LIME explanation for the SVM model prediction for both class "0" and class "1" on specific text lines. LIME identified the most influential tokens contributing to the model prediction. The weights of each of the influential tokens are sorted based on their class predictions, and the sum of the weights for each class is calculated to reach the predicted class [43]. Tokens in input text are highlighted based on their probabilities of falling into a class. These explanations contribute significantly to understanding the decision-making process [44].

In our comprehensive analysis of top tokens using LIME for both SVM and CNN models, we gained granular insights into the differentiating features that determine classification performance. This not only enhances our understanding of model predictions but also provides interpretability, shedding light on the key factors influencing the decision-making process within these complex models. The most important tokens for this dataset according to our analysis were "raumluft", "kein", "nicht",

Table 5 Common text segments identified as false positives (FP) and false negatives (FN) during error analysis of the test data

FP/FN	Text span
FP	"O2 Sättigungswert vom 1.12.2019 mit Normalwerten"
	"O2 saturation value from 1.12.2019 with normal values"
FP	"respiratorischem Infekt und keinem erhöhten O2 Bedarf"
	"respiratory infection and no increased O2 requirement"
FN	"SO2-Bedarf"
	"SO2 requirement"
FN	"Laut Pflegebericht: im PH trotz Sauerstoffgabe Sättigung von 65% und"
	"According to the care report: saturation of 65% in the PH despite oxygen administration and"

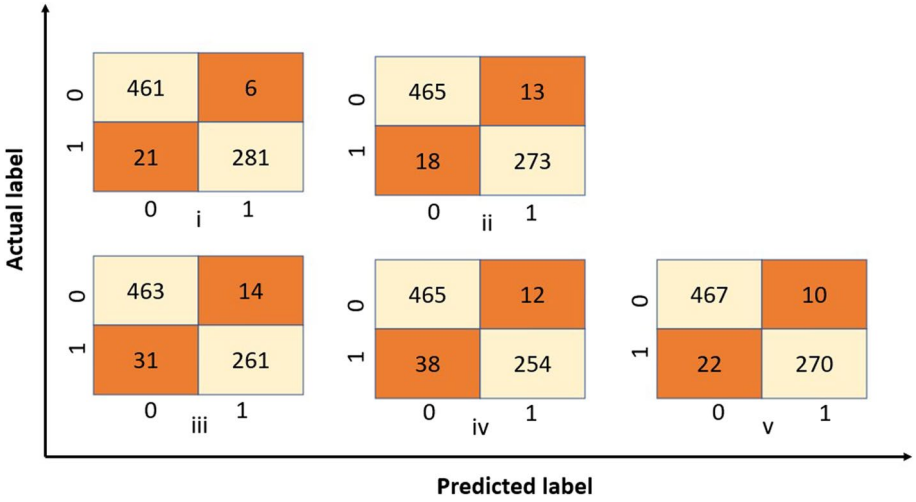

Fig. 4 Confusion matrices depicting the performance of (i) SVM, (ii) RF, (iii) LSTM, (iv) Bi-LSTM, and (v) CNN models

"ohne" for class 0 (no supplemental oxygen) and "FiO2", "mit", "gabe", numbers followed by "L" or "l" determining the litres for with supplemental oxygen (class 1), cf. Table 4 for the top 10 tokens per class.

Error analysis

Confusion matrices are designed to give the predicted values in a count format, which distinguishes between correct and incorrect predictions. The true positive and true negative values provide a clear picture of the correct predictions within the network, while the false positives and false negatives are the topics of interest for error analysis.

Analysing the false positives, i.e. the number of incorrectly assigned text lines to have received oxygen supplementation and false negatives, i.e. the number of incorrectly assigned text lines to not have received oxygen supplementation, it was of interest that for all random state values, there were overlapping texts in these categories within all models, i.e., for a random state value of 729 there were 7 and 8 overlapping false positive and false negatively classified texts in all models, cf. Table 5. Figure 4 illustrated the values obtained for the confusion matrices for different models at the random state value of 729.

Discussion

In all languages, but particularly in languages other than English, the access to comprehensive clinical narrative datasets for public use poses a challenge. Legal restrictions, privacy policies, and stringent data protection regulations limit their availability and hinder their publication [45]. Unfortunately, this constrained environment has impeded our ability to test our models on diverse datasets, thereby limiting our capacity to confirm the generalizability of our findings. Consequently, faced with these limitations, we opted to create our dataset for the experiment. The process of manual annotation proved to be a particularly arduous task, underscoring the challenges associated with compiling and annotating clinical narrative data under such restrictions.

Since the dataset is imbalanced with a substantially higher number of text snippets in class "0" compared to class "1", accuracy is not a suitable metric for evaluating model performance. Hence, precision, recall and F1-score are more informative as they provide a better measure of the model's potential to detect the minority class. Table 3 highlights that the performance of classical ML models overlaps with DL models. Despite the emergence of DL models, there are several applications where classical ML such as SVM outperformed DL approaches [46]. Even image classifiers related to the COVID-19 context have observed this phenomenon [47, 48].

In addition to the model's predictive performance, the mean prediction time per sample was also assessed. Out of this assessment, the classical ML model (RF) is the fastest network for this text classification task. In comparison with Saadatmand et al. [13] and Yamanaka et al. [14], who used certain features such as demographics, symptoms, patient background, etc. for determining the requirement of oxygen therapy, our experiments were especially focused on clinical narratives for oxygen status. In contrast to Muto et al. [16], which relies on decision support from clinicians, our methodology leverages an explainable AI module to understand the model decisions.

Even Fig. 2 does not show clear and distinct clusters for CNN as one class appears as an inverted V shape, while the other class is spread inside, which suggests that the classes might not be easily separable in the embedded space. This reveals the possibility of (i) overlap between classes, (ii) high intrinsic dimensionality that may not be captured by t-SNE, and (iii) complex non-linear relationships within the data.

Conclusion

In this paper, text lines extracted from German-language discharge summaries of COVID-19 patients were used to detect patients who received supplementary oxygen therapy, which constitutes important information for building cohorts for retrospective COVID-19 clinical studies. The classification task had to distinguish the mention of oxygen related to oxygen measurement from the mention of oxygen in the context of oxygen supplementation.

Of the applied classification methods using classical machine learning to deep learning models, the performance of all of them (SVM, RF, LSTM, Bi-LSTM, and CNN) was similar. When comparing their computational efficiency, the RF model stood out, being the fastest classifier for this task, as well as in terms of training efforts. LIME aided in analysing and explaining the model predictions and played a crucial role in understanding the model performance. The pandemic highlighted the need for computerised classifications for the effective management of patient information in hospitals and for clinicians.

In future work, we aim to expand our research by acquiring additional datasets, thereby enhancing the robustness and generalizability of our model. We also plan to investigate its performance across diverse languages, ensuring its applicability and effectiveness in a broader linguistic context.

Abbreviations

COVID-19	Coronavirus disease 2019
SARS-CoV-2	Severe acute respiratory syndrome coronavirus
FiO2	Fraction of inspired oxygen
PaO2/PO2	Partial pressure of oxygen
SaO2	Arterial oxygen saturation
SpO2	Peripheral oxygen saturation
NLP	Natural Language Processing
SVM	Support vector machine
SVC	Support vector classifier
LSTM	Long short-term memory
Bi-LSTM	Bidirectional LSTM
CNN	Convolutional neural network
RF	Random Forest
EHRs	Electronic health records
NER	Named entity recognition
NEN	Named entity normalization
ANN	Artificial neural networks
RNN	Recurrent neural network
LIME	Local interpretable model-agnostic explanations
XAI	Explainable artificial intelligence
DL	Deep learning
ML	Machine learning
t-SNE	t-distributed Stochastic Neighbour Embedding
TF-IDF	Term frequency-inverse document frequency
PCA	Principal component analysis

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-024-02425-2.

Additional file 1.

Authors' contributions

AA and MK designed the project and the processing workflow with feedback from SS. AA, AK and SS annotated the dataset. VS initiated the retrospective clinical study. MK triggered the problem motivation and AA is responsible for the core implementation. All authors read and approved the final version of the manuscript.

Funding

The project was in part conducted at the Center for Biomarker Research in Medicine (CBmed), a COMET K1 centre funded by the Austrian Research Promotion Agency (Project 3.23).

Availability of data and materials

The datasets generated and analysed during this study are not publicly available in accordance to the local ethics approval. However, they can be made accessible from the corresponding author in consultation with the institutional review board of the Medical University of Graz on reasonable request.

Declarations

Ethics approval and consent to participate

This study was approved by the institutional review board of the Medical University of Graz (30-496 ex 17/18) and (32-431 ex 19/20). Informed consent was waived due to the retrospective nature of the initial clinical study by the institutional review board of the Medical University of Graz (32-431 ex 19/20) and was registered at clinicaltrials.gov (NCT04420637). Only anonymised data has been used, and all methods were carried according to the approved ethics regulations mentioned before.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 1 June 2023 Accepted: 15 January 2024
Published online: 31 January 2024

References

1. Velavan TP, Meyer CG. The COVID-19 epidemic. Trop Med Int Health. 2020;25(3):278.
2. WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int. Accessed 11 Dec 2023.
3. Li Q, Peng H, Li J, Xia C, Yang R, Sun L, et al. A survey on text classification: From traditional to deep learning. ACM Trans Intell Syst Technol (TIST). 2022;13(2):1–41.
4. Dogra V, Verma S, Chatterjee P, Shafi J, Choi J, Ijaz MF, et al. A complete process of text classification system using state-of-the-art NLP models. Comput Intell Neurosci. 2022;2022:26.
5. Chen D, Liu S, Kingsbury P, Sohn S, Storlie CB, Habermann EB, et al. Deep learning and alternative learning strategies for retrospective real-world clinical data. NPJ Digit Med. 2019;2(1):43.
6. Jaber A, Martínez P. Disambiguating Clinical Abbreviations Using a One-Fits-All Classifier Based on Deep Learning Techniques. Methods Inf Med. 2022;61:e28–34.
7. Idris S, Badruddin N. Classification of Cognitive Frailty in Elderly People from Blood Samples using Machine Learning. In: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). 2021. p. 1–4.
8. Mittas N, Chatzopoulou F, Kyritsis KA, Papagiannopoulos CI, Theodoroula NF, Papazoglou AS, et al. A Risk-Stratification Machine Learning Framework for the Prediction of Coronary Artery Disease Severity: Insights From the GESS Trial. Front Cardiovasc Med. 2022;8.
9. Yang B, Dai G, Yang Y, Tang D, Li Q, Lin D, et al. Automatic text classification for label imputation of medical diagnosis notes based on random forest. In: Health Information Science: 7th International Conference, HIS 2018, Cairns, QLD, Australia, October 5–7, 2018, Proceedings 7. Springer; 2018. p. 87–97.
10. Ong CJ, Orfanoudaki A, Zhang R, Caprasse FPM, Hutch MR, Ma L, et al. Machine learning and natural language processing methods to identify ischemic stroke, acuity and location from radiology reports. PLoS ONE. 2020;15(6):e0234908.
11. Sun B, Wei HL. Machine Learning for Medical and Healthcare Data Analysis and Modelling: Case Studies and Performance Comparisons of Different Methods. In: 2022 27th International Conference on Automation and Computing (ICAC). 2022. p. 1–6.
12. Daher A, Balfanz P, Aetou M, Hartmann B, Müller-Wieland D, Müller T, et al. Clinical course of COVID-19 patients needing supplemental oxygen outside the intensive care unit. Sci Rep. 2021;11(1):1–7.
13. Saadatmand S, Salimifard K, Mohammadi R, Marzban M, Naghibzadeh-Tahami A. Predicting the necessity of oxygen therapy in the early stage of COVID-19 using machine learning. Med Biol Eng Comput. 2022;60(4):957–68.
14. Yamanaka S, Morikawa K, Azuma H, Yamanaka M, Shimada Y, Wada T, et al. Machine-learning approaches for predicting the need of oxygen therapy in early-stage COVID-19 in Japan: multicenter retrospective observational study. Front Med. 2022;9:846525.
15. Chung J, Kim D, Choi J, Yune S, Song K, Kim S, et al. Prediction of oxygen requirement in patients with COVID-19 using a pre-trained chest radiograph xAI model: efficient development of auditable risk prediction models via a fine-tuning approach. Sci Rep. 2022;12(1):21164.
16. Muto R, Fukuta S, Watanabe T, Shindo Y, Kanemitsu Y, Kajikawa S, et al. Predicting oxygen requirements in patients with coronavirus disease 2019 using an artificial intelligence-clinician model based on local non-image data. Front Med. 2022;9:1042067.
17. Kar S, Chawla R, Haranath SP, Ramasubban S, Ramakrishnan N, Vaishya R, et al. Multivariable mortality risk prediction using machine learning for COVID-19 patients at admission (AICOVID). Sci Rep. 2021;11(1):1–11.
18. Sánchez-Montañés M, Rodríguez-Belenguer P, Serrano-López AJ, Soria-Olivas E, Alakhdar-Mohmara Y. Machine learning for mortality analysis in patients with COVID-19. Int J Environ Res Pub Health. 2020;17(22):8386.
19. Becerra-Sánchez A, Rodarte-Rodríguez A, Escalante-García N, Olvera-González JE, la Rosa-Vargas JID, Zepeda-Valles G, et al. Mortality analysis of patients with COVID-19 in Mexico based on risk factors applying machine learning techniques. Diagnostics. 2022;12(6):1396.
20. Li X, Ge P, Zhu J, Li H, Graham J, Singer A, et al. Deep learning prediction of likelihood of ICU admission and mortality in COVID-19 patients using clinical variables. PeerJ. 2020;8:e10337.
21. Marcos M, Belhassen-García M, Sánchez-Puente A, Sampedro-Gomez J, Azibeiro R, Dorado-Díaz PI, et al. Development of a severity of disease score and classification model by machine learning for hospitalized COVID-19 patients. PLoS ONE. 2021;16(4):e0240200.
22. Gabbay F, Bar-Lev S, Montano O, Hadad N. A LIME-Based Explainable Machine Learning Model for Predicting the Severity Level of COVID-19 Diagnosed Patients. Appl Sci. 2021;11(21):10417.
23. Catoire P, Tellier E, de La Rivière C, Beauvieux MC, Valdenaire G, Galinski M, et al. Assessment of the SpO2/FiO2 ratio as a tool for hypoxemia screening in the emergency department. Am J Emerg Med. 2021;44:116–20.
24. Piraino T, Madden M, Roberts KJ, Lamberti J, Ginier E, Strickland SL. AARC clinical practice guideline: management of adult patients with oxygen in the acute care setting. Respir Care. 2022;67(1):115–28.
25. Hafen BB, Sharma S. Oxygen Saturation. StatPearls Publishing; 2023. http://www.ncbi.nlm.nih.gov/books/NBK525974/. Accessed 11 Dec 2023.
26. Wilkerson RG, Adler JD, Shah NG, Brown R. Silent hypoxia: a harbinger of clinical deterioration in patients with COVID-19. Am J Emerg Med. 2020;38(10):2243-e5.
27. Velavan TP, Meyer CG. Mild versus severe COVID-19: laboratory markers. Int J Infect Dis. 2020;95:304–7.
28. McHugh ML. Interrater reliability: the kappa statistic. Biochemia Med. 2012;22:276–82.
29. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.
30. Noble WS. What is a support vector machine? Nat Biotechnol. 2006;24(12):1565–7.
31. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
32. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.
33. Greff K, Srivastava RK, Koutník J, Steunebrink BR, Schmidhuber J. LSTM: A search space odyssey. IEEE Trans Neural Netw Learn Syst. 2016;28(10):2222–32.
34. Chollet F, et al. Keras. GitHub. 2015. https://github.com/fchollet/keras. Accessed 11 Dec 2023.
35. Li C, Zhan G, Li Z. News text classification based on improved Bi-LSTM-CNN. In: 2018 9th International conference on information technology in medicine and education (ITME). IEEE; 2018. p. 890–3.
36. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging. 2018;9(4):611–29.
37. Mullenbach J, Wiegreffe S, Duke J, Sun J, Eisenstein J. Explainable prediction of medical codes from clinical text. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for

Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers); 2018. p. 1101–11.
38. Kim Y. Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics; 2014.
39. Wang C, Nulty P, Lillis D. A comparative study on word embeddings in deep learning for text classification. In: Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval; 2020. p. 37–46.
40. Liashchynskyi P, Liashchynskyi P. Grid search, random search, genetic algorithm: a big comparison for NAS. 2019. arXiv preprint arXiv:191206059.
41. van der Maaten L, Hinton GE. Visualizing Data using t-SNE. J Mach Learn Res. 2008;9:2579–605.
42. Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. p. 1135–44.
43. Visani G, Bagli E, Chesani F, Poluzzi A, Capuzzo D. Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models. J Oper Res Soc. 2020;73:91–101.
44. Sathyan A, Weinberg AI, Cohen K. Interpretable AI for bio-medical applications. Complex Eng Syst (Alhambra, Calif). 2022;2(4):18.
45. Frei J, Kramer F. GERNERMED: An open German medical NER model. Softw Impacts. 2022;11:100212.
46. Basu A, Walters C, Shepherd M. Support vector machines for text categorization. In: Proceedings of the 36th Annual Hawaii International Conference on System Sciences, 2003. IEEE; 2003. p. 7–pp.
47. Huda NLI, Islam MA, Goni MO, Begum N. Covid-19 Classification Using HOG-SVM and Deep Learning Models. In: 2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET). 2022. p. 1–5.
48. Dairi A, Harrou F, Sun Y. Deep Generative Learning-Based 1-SVM Detectors for Unsupervised COVID-19 Infection Detection Using Blood Tests. IEEE Trans Instrum Meas. 2022;71:1–11.

Publisher's Note

Publication 2

Original Research Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–12

DOI: 10.1177/20552076241288681
journals.sagepub.com/home/dhj

Sage

Unsupervised SapBERT-based bi-encoders for medical concept annotation of clinical narratives with SNOMED CT

Akhila Abdulnazar[1,2], Roland Roller[3], Stefan Schulz[1] and Markus Kreuzthaler[1]

Abstract

Objective: Clinical narratives provide comprehensive patient information. Achieving interoperability involves mapping relevant details to standardized medical vocabularies. Typically, natural language processing divides this task into named entity recognition (NER) and medical concept normalization (MCN). State-of-the-art results require supervised setups with abundant training data. However, the limited availability of annotated data due to sensitivity and time constraints poses challenges. This study addressed the need for unsupervised medical concept annotation (MCA) to overcome these limitations and support the creation of annotated datasets.

Method: We use an unsupervised SapBERT-based bi-encoder model to analyze n-grams from narrative text and measure their similarity to SNOMED CT concepts. At the end, we apply a syntactical re-ranker. For evaluation, we use the semantic tags of SNOMED CT candidates to assess the NER phase and their concept IDs to assess the MCN phase. The approach is evaluated with both English and German narratives.

Result: Without training data, our unsupervised approach achieves an F1 score of 0.765 in English and 0.557 in German for MCN. Evaluation at the semantic tag level reveals that "disorder" has the highest F1 scores, 0.871 and 0.648 on English and German datasets. Furthermore, the MCA approach on the semantic tag "disorder" shows F1 scores of 0.839 and 0.696 in English and 0.685 and 0.437 in German for NER and MCN, respectively.

Conclusion: This unsupervised approach demonstrates potential for initial annotation (pre-labeling) in manual annotation tasks. While promising for certain semantic tags, challenges remain, including false positives, contextual errors, and variability of clinical language, requiring further fine-tuning.

Keywords

Named entity recognition, medical concept normalization, SNOMED CT, natural language processing, interoperability

Submission date: 13 February 2024; Acceptance date: 3 September 2024

Introduction

Electronic health records (EHRs) store extensive health data, including patient details on diseases, risks, procedures, and medications.[1] Most of this information, crafted by healthcare professionals under time constraints, is in narrative form, often dense, filled with abbreviations, and disregarding grammar rules. This emphasizes the need to map expressions to standardized codes for effective communication.[2,3] To address this need, named entity recognition (NER) and medical concept normalization (MCN),[4] also known as entity linking, a subfield of natural language processing (NLP),[5] plays an important role. As healthcare organizations increasingly adapt EHR systems, the demand for clinical terminology in real-life clinical applications is

[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]German Research Center for Artificial Intelligence (DFKI), Berlin, Germany

Corresponding author:
Markus Kreuzthaler, Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Graz, Austria. markus.kreuzthaler@medunigraz.at

increasing rapidly. In our study, we prioritize standardization of EHR data using clinical terminology.

Conventional biomedical NER methods can be broadly classified into dictionary-based, semantic, and statistical approaches.[6] In recent years, state-of-the-art (SOTA) approaches heavily rely on deep learning (DL) algorithms,[7–9,11,10] and most prominently transformer models, such as bidirectional encoder representations from transformers (BERT)[12] and its variations.[6] However, to solve medical NER, BERT requires training data, which is often very limited due to the sensitive nature of the text.

From a technological point of view, the same applies to MCN. BERT models outperform, for instance, other SOTA architectures.[13,14] They also excelled in handling multilingual data and capturing contextual information, surpassing the previous SOTA for the normalization of biomedical concepts.[15–18,4] A pairwise learning-to-rank with a vector space model,[15] enhanced performance of BERT, BioBERT,[19] and ClinicalBERT[20,21] models across different datasets, surpassing previous methods. In the 2019 n2c2/UMass Lowell task on MCN, various methods were tested, such as dictionary matching, DL, retrieval, and rank techniques[22–24] using similarity metrics such as the cosine distance. The most accurate approach used a DL structure with a pre-trained SciBERT layer.[25] Deep neural network models for generating sentence embeddings as semantic representations, also enhanced cross-lingual biomedical concept normalization.[26] Another method, utilizing target concept guidance in MCN within noisy user-generated texts,[27] effectively integrates target concept information and domain lexicon knowledge to enhance model performance.

Self-alignment pre-training for BERT (SapBERT), a scheme that self-aligns the representation space of text elements, exhibited better performance compared to seven other BERT models.[28] Fine-tuning SapBERT set a new standard in DL for recognition of multilingual entities,[29] and cross-lingual normalization.[30] xMEN[31] excels in cross-lingual MCN with unsupervised candidate generation and supervised cross-encoders surpassing previous benchmarks. The findings of Lin et al.[26] demonstrated that the SapBERT model achieved the highest performance on both English and cross-lingual datasets.

Combining NER and MCN is a challenging task.[32] Several existing methodologies also employ a combined strategy, leveraging the strengths of both NER and MCN to achieve more comprehensive and accurate results. MetaMap,[33] maps text to UMLS Metathesaurus but faces challenges with spelling mistakes and ambiguous concepts. Bio-YODIE[34] improves extraction speed and disambiguation but requires annotated data. SemEHR[35] builds on Bio-YODIE but relies on manual rules for enhancement. cTAKES[36] utilizes existing technologies but needs plugins to handle certain challenges. ScispaCy[37] is a supervised NER model with limited linking capabilities. CLAMP[38] is a comprehensive clinical NLP tool, while BioPortal[39] offers annotation for various ontologies but may face data protection issues due to its external interface. MedCAT[40] is a flexible concept extraction tool using any terminology based vocabulary. It boasts a user-friendly interface for customization and model training, making it versatile for clinical and research tasks. However, it requires annotated data for optimal performance. Despite the progress in transformer technologies, challenges persist in solving NER and MCN. Key issues include:

Partial matches. Partial matches may occur due to spelling variations or errors. If "femoral neck fracture" misspells "fracture" as "fractur," this can lead to partial matches because the terms are similar but not exactly the same.

Ambiguity. Clinical terms often have different ways of being expressed, resulting in ambiguity. "FNF," an acronym for "femoral neck fracture," would also be found as a synonym for "finger-nose-finger" (a neurological test) in a comprehensive dictionary.[41]

Contextual information. The term "fracture" can be assigned with a concept ID as well as based on context, as "fracture of the neck of the femur" can be assigned with another concept ID without the context, see Table 1.

Non-contiguous mentions. Dealing with non-contiguous mentions and variations in token order means recognizing these different expressions as referring to the same medical concept. The term "femoral neck fracture" may be rearranged or expressed differently, such as "fracture of the neck of the femur."

Incomplete terminology. Incomplete clinical terminology systems often lack synonyms or short forms. This means that alternative terms such as "hip fracture" may not be recognized as partial matches.

Medical data often contains sensitive information, which makes it difficult to share. Even in the current era of large language models such as ChatGPT, their application in medical settings poses ethical and privacy issues.[42] Concerns include patient privacy breaches, unclear responsibility in case of harm, and the need for clear rules to protect users.[43] For these reasons, it is crucial to weigh the implications and explore privacy-focused alternatives. Low-resource languages often suffer from a lack of publicly available datasets due to various factors such as small corpus sizes, different formats suited for specific tasks, and limited accessibility.[44]

Clinical gold standards refer to a benchmark available under reasonable conditions.[45] However, the number of publicly available gold standards is limited, necessitating unsupervised approaches when labeled data is unavailable.[46] Different unsupervised NER approaches were reported, utilizing adversarial training[47] and contextualized word representations.[48] Nath et al.[49] focused on unsupervised specialized word embeddings and NER for clinical coding. Within unsupervised approaches for MCN, Yan et al.[50] utilized multi-instance learning for linking Chinese medical symptoms to ICD-10 classifications, surpassing the baseline by 1.72%. Tahmasebi et al.[51]

Table 1. Examples of named entities with SNOMED CT codes and preferred terms related to a text mention from a clinical narrative: "suspected fracture of the neck of the right femur," showing multiple possibilities of concept mapping for a single input text.

Text	Semantic tag	Code	Preferred term
suspected	qualifier value	415684004	Suspected
fracture of the neck of the femur	disorder	5913000	Fracture of neck of femur
right	qualifier value	24028007	Right
suspected	qualifier value	415684004	Suspected
fracture	morphologic abnormality	72704001	Fracture
of the neck of the right femur	body structure	773710001	Structure of neck of right femur

demonstrated effective unsupervised anatomical phrase normalization using word embeddings in SNOMED CT. Karadeniz et al.[52] achieved precision scores of 65.9% and 68.7% for bacteria biotope entities and adverse drug reactions, respectively, using unsupervised entity linking methods with word embeddings and syntactic re-ranking.

The first general and complete unsupervised solution for NER with entity detection and classification used a noun phrase chunker with inverse document frequency for boundary detection and distributional semantics for terminology code assignment.[53] The overall classification shows good results, considering that only 39% and 19% of the entities could be found according to the datasets used. Another unsupervised framework for recognizing and linking medical entities from Chinese online medical text, namely unMERL[54] uses a combination of offline linguistic resources and online detection approaches to improve the recognition and linking performance. The results show that unMERL consistently outperforms current approaches and has good generalizability.

To address missing entities compared to the other unsupervised approaches,[53,54] we employed n-gram-based entity detection and leveraged (SapBERT), a SOTA pre-trained model for biomedical entity linking, to vectorize entities for similarity matching utilizing FAISS. The matching process yields two key pieces of information: (i) semantic tags for assessing the NER phase and (ii) concept IDs crucial for evaluating the MCN phase within the narrative under scrutiny. We hypothesize that this result is useful to semi-automatically support the manual medical concept annotation (MCA) task, essential for training supervised methods. To the best of the authors' knowledge, this study is the only one to integrate entity recognition and normalization using an unsupervised method that relies solely on data from knowledge bases and can be adapted to different languages, providing maximum coverage of semantic understanding.

Material and methods

This section outlines our methodology and materials. We employ the NLTK n-gram generator[8] to extract linguistic patterns, specifically for entity mention detection. SNOMED CT,[55] described in Section "Terminologies," is used as the reference terminology for mapping clinical entities. Details on the datasets and the proposed framework are discussed in Sections "Datasets" and "Proposed approach."

Terminologies

The UMLS Methathesaurus is a large dataset unifying about 150 biomedical terminologies, such as MeSH, SNOMED CT, and RxNORM, and links concepts of 200 different vocabularies.[56] In this work, we are particularly interested in the subset SNOMED CT, a standardized, multilingual clinical terminology that includes more than 350,000 entities.[57,55,58] It facilitates the comprehension and exchange of health information among diverse systems through the use of codes and expressions.[57]

Datasets

For our experiments, we use the 2019 n2c2/UMass Lowell shared task on MCN dataset,[24] consisting of 100 discharge summaries from U.S. hospitals. In these texts, 10,919 mentions of medical problems (diagnoses), treatments, and tests were manually annotated using UMLS.[59] In this work, we consider only SNOMED CT annotations due to their global acceptance, comprehensive scope, compatibility with FHIR, widespread adoption, and broad coverage, ensuring standardized and comprehensive healthcare data representation.[58] Within the dataset, we considered the top 10 most frequent semantic tags ("procedure," "disorder," "qualifier

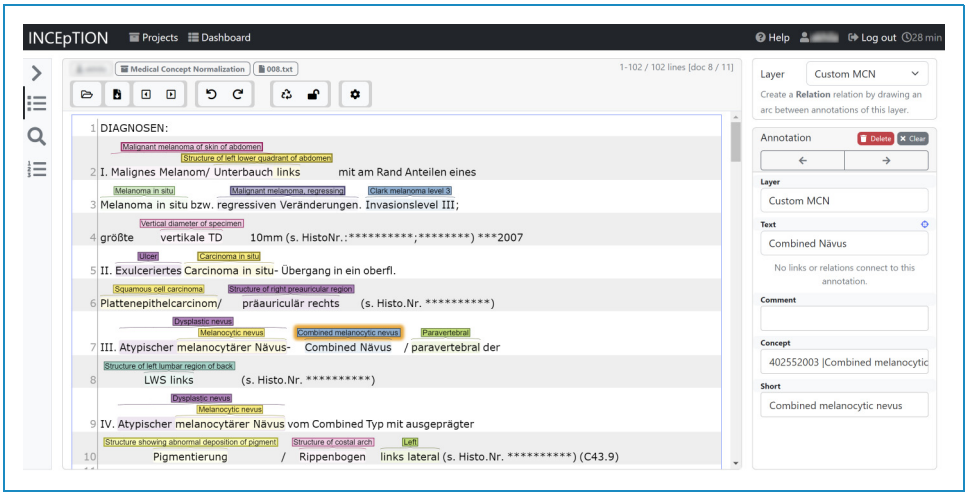

Figure 1. Manually annotated clinical narrative in German using INCEpTION.

value," "finding," "substance," "body structure," "morphologic abnormality," "observable entity," "physical object," and "regime/therapy"), representing 95% of all SNOMED CT concepts within the dataset. Focusing on SNOMED CT candidates of the n2c2 dataset results in 6232 training mentions and 6528 test mentions.

To maintain consistency with the annotation standards used in existing datasets, we created a new German dataset in addition to the English dataset, for evaluation. The data is from an Austrian network of public hospitals and contains de-identified narratives from 10 EHRs. The texts were manually annotated using INCEpTION,[60] (see Figure 1) following the English annotation guidelines from n2c2.[59] The annotated set of discharge summaries resulted in 600 SNOMED CT normalized mentions that have 97% of the mentions within the aforementioned semantic tags.

Proposed approach

Our work targets unsupervised MCA that combines n-gram decomposition with embedding-based similarity matching, as shown in Figure 2. Given an entity mention, classical normalization approaches would rely on the terms and their synonyms, as mentioned in the terminology, to find a corresponding entry. In this study, we rely on vector representations of those concepts modeled as embeddings. Therefore, given a candidate mention in the form of an embedding, the best vectorized SNOMED CT term needs to be found. Both steps, the vectorization of SNOMED CT and the vector search, are described in the following Section "Embedding space." The detailed overview of our approach is provided in Section "Framework."

Embedding space. SAPBERT serves as a pre-training framework designed to align synonyms of the same biomedical concept into clusters, with a focus on biomedical texts. It offers versatility for both pre-training on the UMLS Metathesaurus and fine-tuning on task-specific datasets. In this work, a SNOMED CT embedding space in English and German[61] is created separately. To generate embeddings, we employed two pre-trained language models within the SAPBERT framework: (i) *PubMedBERT-based SAPBERT (UMLS 2020AA—English).* This model, based on SAPBERT trained with UMLS 2020AA (English only) and utilizing PubMedBERT[62] as a base model, was evaluated on the n2c2 dataset. (ii) Cross-lingual SAPBERT (UMLS 2020AB —all languages). The cross-lingual SAPBERT model, trained with UMLS 2020AB (all languages), employs XML-RoBERTa (large)[63] as the underlying model,[28] was evaluated on the German dataset. The models selected for our article were guided by relevant literature, including the work of Lin et al.,[26] and were further validated through an evaluation outlined in Appendix 2. Our evaluation demonstrated in consistency with the experiments from Lin et al.[26] that SAPBERT performed best for MCN. We leveraged FAISS, an open-source library to perform fast similarity searches in high-dimensional vector spaces,[64] using either cosine similarity or L2 (Euclidean) distance. All SNOMED CT terms (English and German) are represented as 768-dimensional embeddings and are FAISS-indexed to create the corresponding embedding space. In this work, we use cosine similarity for vector similarity matching.

Framework. In the following, we describe the unsupervised MCA framework, as shown in Figure 2.

Block A. Creating embedding spaces for SNOMED CT. *Preprocessing*: Each SNOMED CT term undergoes (i) lower casing, removal of diacritics (extra marks on letters, such as accents or tildes) and stop words except for negations. This ensures a consistent and clean representation of SNOMED CT terms for better analysis. (ii) Vectorization using SAPBERT.[65] *FAISS indexing*: The term vectors are indexed with FAISS for efficient search

and retrieval and stored as an embedding space, ensuring quick and easy access during analysis.

Block B. Preparation of clinical data for testing. *Documents*: Creating a secure storage repository to facilitate organized access to necessary documents for subsequent computational stages. The data in this repository is carefully de-identified to safeguard patient confidentiality.

Block C. N-gram generation, entity recognition and normalization. *N-gram generation*: (i) The input document is read in line by line. (ii) A sliding window approach is utilized to generate token n-grams from the text. This approach allows for the extraction of both single words and short phrases, providing comprehensive coverage of entities within the document. (iii) N-grams of varying lengths (from 1 to 5) are considered to encompass different types of entities. *Preprocessing*: Each generated n-gram undergoes a standardized preprocessing procedure, as discussed in "Block A." This preprocessing ensures consistency in the representation of textual data and prepares it for subsequent analysis and matching steps. *Bi-encoder matching*: (i) The preprocessed n-grams are subjected to FAISS similarity matching against SNOMED CT concepts within the embedding space. (ii) Through similarity matching, n-grams are mapped to their closest corresponding concepts in SNOMED CT, providing the semantic tags and concept IDs. This mapping enables the detection and normalization of spelling variants, errors, and non-adjacent mentions within the text. *Thresholding*: (i) A threshold limit, set at 0.9, is established for the similarity scores obtained from FAISS matching. (ii) This threshold value is derived from overall similarity scores between synonyms and SNOMED CT terms, it filters out ambiguous mentions. Only scores surpassing the threshold are considered, enhancing precision and reliability in entity recognition and normalization.

Block D. Selection and evaluation of the best n-grams. *Syntactical re-ranker*: (i) The input sentence (lines) is tokenized. (ii) For each token, identify all n-grams that contain it. (iii) Sort these n-grams based on whether the token is present or not. (iv) Identify the SNOMED CT candidates with the highest syntactical similarity score. (v) The syntactical similarity score is calculated using the partial ratio of the Levenshtein distance[66] from the terms. (vi) If more than one n-gram has identical scores, the longest n-gram mention is chosen. These introduced steps address the issue of overlapping entity candidates resulting from the window-based approach. *Normalized entities*: For the given input sentence, the best n-grams with their corresponding SNOMED CT terms are obtained. *Evaluation*: (i) To evaluate our approach, we examine if an entity could be found, and if yes, if the extracted entity matches the mentions exactly (*exact* match) or just parts of it (*partial* match). (ii) Given a detected entity, we explore if the correct SNOMED CT term could be linked correctly. (iii) Precision, recall, and the F1 score are used for evaluation, utilizing the test datasets.

We utilize n-grams due to their ability to capture contextual information surrounding entities and accommodate variations in token order within reference terminologies. This approach minimizes errors arising from partial matches and ensures robust entity identification across diverse text structures. By aligning textual mentions (preprocessed n-grams) with semantically similar concepts in SNOMED CT, this step enhances the accuracy and completeness of entity normalization, even amidst lexical variations and structural complexities in the text. Prioritizing mentions with high syntactic similarity ensures precision and accuracy in entity identification. While embeddings effectively capture semantic similarities and handle synonyms, relying solely on semantic similarity may introduce ambiguity and noise, especially in domains with complex terminology. Emphasizing syntactic similarity aims to prioritize entities exhibiting structural and contextual consistency with reference terms in SNOMED CT. This ensures a reliable mapping between textual mentions and corresponding concepts, minimizing the risk of misinterpretation or incorrect normalization. Therefore, incorporating syntactic similarity as a key criterion complements the strengths of embeddings, enhancing overall precision and reliability in entity identification and normalization within biomedical text analysis.

Baseline approach. Our methodology for the mapping of clinical concepts is compared with a baseline approach, using an n-gram generator to extract information from clinical texts, followed by a basic dictionary matching method to identify matches to SNOMED CT. This straightforward baseline serves as a benchmark to assess the effectiveness of our proposed methodology.

Results

Medical concept normalization (MCN)

We first evaluate our SapBERT-based bi-encoder for MCN using SNOMED CT annotations in the n2c2 dataset. In this experiment, the surface terms of interest in the narrative under investigation are already known, therefore concentrating solely on the normalization approach. This allows a focused analysis of our mapping method's performance. The MCN column of Tables 2 and 3 shows the precision (P), recall (R), and F1 score (F1) of this approach in different datasets. The F1 score of 0.765 for MCN on the n2c2 dataset indicates a good performance, and it is important to note that this was achieved without using any training data. While the F1 score of 0.557 on the German dataset raises the necessity for thorough error analysis. Additionally, it is noteworthy that among the two tables, "disorder" exhibits a higher F1 score at the semantic tag level.

Medical concept annotation (MCA)

Following the initial evaluation, we proceed to analyze the MCA method. This integrated approach combines MCN with

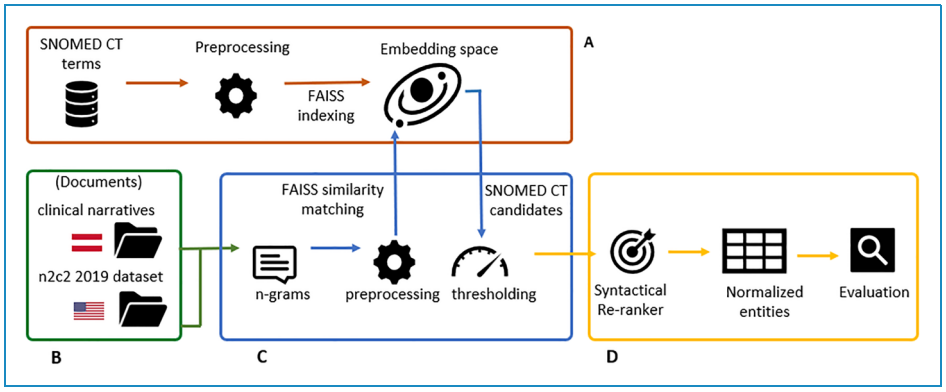

Figure 2. Illustration of the proposed framework with (A) creating an embedding space of SNOMED CT, (B) preparation of clinical data for testing, (C) n-gram generation, entity recognition and normalization, and (D) selection and evaluation of the best n-grams.

Table 2. Performance of MCN and MCA on semantic tag levels using n2c2 dataset.

	MCN			MCA: NER			MCA: MCN		
Top 10 semantic tags	P	R	F1	P	R	F1	P	R	F1
Procedure	0.612	0.554	0.572	0.732	0.454	0.560	0.409	0.359	0.362
Disorder	0.900	0.860	0.871	0.957	0.746	0.839	0.745	0.680	0.696
Qualifier value	0.914	0.854	0.871	0.262	0.848	0.400	0.192	0.240	0.204
Finding	0.782	0.748	0.759	0.563	0.667	0.611	0.410	0.403	0.400
Substance	0.871	0.849	0.855	0.717	0.814	0.762	0.640	0.596	0.602
Body structure	0.783	0.709	0.720	0.255	0.575	0.353	0.163	0.190	0.164
Morphologic abnormality	0.825	0.770	0.786	0.747	0.750	0.748	0.578	0.557	0.551
Observable entity	0.764	0.703	0.726	0.136	0.485	0.213	0.107	0.108	0.101
Physical object	0.711	0.655	0.675	0.476	0.618	0.538	0.273	0.303	0.279
Regime/therapy	0.600	0.517	0.542	1.0	0.400	0.571	0.585	0.385	0.430
Macro	0.776	0.722	0.748	0.585	0.634	0.608	0.410	0.382	0.396
Weighted	0.776	0.755	0.765	0.507	0.684	0.582	0.354	0.351	0.353

MCN: medical concept normalization; MCA: medical concept annotation; NER: named entity recognition; P: precision; R: recall; F1: F1 score.

n-gram decomposition for a comprehensive analysis of our methodology. The mentions proposed by the n-grams are utilized within the MCN, potentially deviating from the test data mentions, therefore resulting in either exact, partial, or no matches. The results at the semantic tag (MCA: NER) and concept ID level (MCA: MCN) are presented in Tables 2 and 3.

Similar to the MCN method, the analysis of the MCA approach reveals that "disorders" consistently exhibit higher performance in both evaluated cases. Performance in the semantic tag "regime/therapy," achieving perfect precision (1.0) and a fair F1 score, showcasing its effectiveness in this particular semantic category. Comparing the results in the MCN and MCA: MCN columns of Tables 2 and 3, shows a uniform decline in MCA:MCN performance irrespective of the semantic tags, which highlight the need for better entity recognition systems.

To address computational requirements and processing times in real-world applications, we adopted a pragmatic

Table 3. Performance of MCN and MCA on semantic tag levels using German EHRs dataset.

	MCN			MCA: NER			MCA: MCN		
Top 10 semantic tags	P	R	F1	P	R	F1	P	R	F1
Procedure	0.614	0.504	0.530	0.492	0.473	0.482	0.234	0.230	0.224
Disorder	0.693	0.634	0.648	0.706	0.664	0.685	0.463	0.439	0.437
Qualifier value	0.680	0.557	0.596	0.146	0.598	0.235	0.106	0.108	0.096
Finding	0.520	0.485	0.498	0.340	0.350	0.344	0.159	0.173	0.164
Substance	0.328	0.219	0.247	0.437	0.484	0.459	0.133	0.144	0.134
Body structure	0.635	0.595	0.599	0.545	0.571	0.558	0.325	0.258	0.240
Morphologic abnormality	0.686	0.600	0.629	0.474	0.514	0.493	0.282	0.309	0.291
Observable entity	0.448	0.448	0.448	0.200	0.379	0.262	0.089	0.096	0.091
Physical object	0.857	0.857	0.857	0.122	0.714	0.208	0.116	0.116	0.116
Regime/therapy	0.700	0.600	0.633	0.250	0.375	0.300	0.206	0.235	0.216
Macro	0.616	0.545	0.578	0.371	0.512	0.430	0.210	0.211	0210
Weighted	0.599	0.521	0.557	0.352	0.531	0.424	0.198	0.196	0.197

MCN: medical concept normalization; MCA: medical concept annotation; NER: named entity recognition; P: precision; R: recall; F1: F1 score.

approach, by extracting random samples of 100 lines of different lengths from the documents under consideration. By calculating the processing time for each line and deriving an average, we obtained a representative measure of the time required to process individual lines. Our findings indicate an approximate processing time of 60 s per line, which has to be optimized when applying this method for large-scale document processing. The main reason is the decomposition of the line under scrutiny into varying n-gram lengths, each of them a possible candidate that has to be processed.

Error analysis

The significant performance gap between English and German datasets prompted an investigation into their linguistic and structural differences. English's analytic nature contrasts with German's synthetic structure, impacting sentence comprehension due to differences in word order. German's complex morphology poses challenges for tasks such as part-of-speech tagging, and divergent vocabulary and idiomatic expressions require tailored approaches. Additionally, variations in naming conventions and syntax offer insights into language model processing.

Upon closer examination of Tables 2 and 3, distinct groups of errors were identified, including contextual errors, granularity errors, analogy errors, similarity errors, wrong IDs, nugatory IDs, acronym errors, and spelling errors. Contextual, granularity, and analogy errors emerged as the most prevalent categories. Contextual errors primarily manifested as non-contiguous mentions, stemming from incomplete span coverage or ambiguous spans. To address these errors, efforts should focus on refining entity recognition algorithms for improved span delineation accuracy. Granularity and analogy errors, stemming from challenges in providing a singular "correct" normalization to a mention, were also significant contributors to performance degradation. In contrast, less frequently occurring errors included spelling and acronym errors. A detailed overview of different types of errors, along with examples, is provided in Appendix 1.

The performance of MCA was compared with the baseline approach of dictionary matching, see Table 4. MCA generally achieved higher precision, recall, and F1 scores for both NER and MCN compared to the baseline dictionary matching in the n2c2 dataset. However, in the German EHRs dataset, MCA showed lower P, R, and F1 scores compared to the baseline. A detailed analysis revealed that MCA outperformed the baseline in identifying exact mentions in both German and English data but exhibited a higher incidence of false positives. Addressing false positives, particularly for semantic tags such as "qualifier value" and "observable entity," is crucial

Table 4. Comparison of NER and MCN of MCA with the baseline dictionary matching approach using n-grams on the n2c2 (en) and German EHR (de) datasets.

		NER			MCN		
Dataset	Approach	P	R	F1	P	R	F1
n2c2	Dictionary matching	0.483	0.530	0.506	0.290	0.270	0.280
	MCA	0.507	0.684	0.582	0.354	0.351	0.353
German EHRs	Dictionary matching	0.521	0.359	0.528	0.256	0.232	0.252
	MCA	0.352	0.531	0.424	0.198	0.196	0.197

MCN: medical concept normalization; MCA: medical concept annotation; NER: named entity recognition; P: precision; R: recall; F1: F1 score.

for improving overall precision. Moreover, reducing false negatives is essential to ensure a comprehensive capture of all relevant mentions, highlighting the importance of ongoing refinement for comprehensive MCN.

Discussion

In this work, we focus exclusively on SNOMED CT and employ an unsupervised approach. Our model achieves an F1 score of 0.765 in MCN, as shown in Table 2, which indicates that it has no prior knowledge from the training set. This score can be compared to the "unseen concepts" category in the work mentioned by Xu et al.,[30] which attained an accuracy of 0.691. Their findings indicate that future research on MCN should more effectively address previously unconsidered concepts, which was another motivating factor for this study. In addition to the unsupervised approach, our experiments showed promising results reaching an F1 score of 0.872, especially when leveraging additional training data within the embedding space for MCN, as shown in Appendix 2—SNOMED CT*. This outperformed the top-performing teams in the 2019 n2c2 challenge and other BERT models. This result also competes with the supervised method as investigated by Xu et al.,[30] considering only SNOMED CT concepts.

In contrast to Zhang and Chen[10] and Chen et al.,[11] we refrained from employing advanced preprocessing steps before NER, such as abbreviation expansion or numeral replacement, which reduce the complexity and computational overhead associated with preprocessing. Even though, MCA resulted in a higher incidence of false positives, with approximately 40% to 50% of detected entities contributing to these errors. A closer examination revealed that nearly 70% of identified mentions as "qualifier value" and "observable entity" were false positives. These stemmed from entities not present in the gold standard data, matched within the terminology. Addressing this abundance is pivotal for precision improvement. Re-evaluating "qualifier values" inclusion may enhance precision. Additionally, refining the extraction process is crucial to minimize false positives and enhance precision. While MCA outperformed the baseline in terms of false negatives, particularly in the German dataset, there remains room for improvement in this aspect. Reducing false negatives is also essential to ensure the comprehensive capture of all relevant mentions.

The consistently high performance of the "disorder" underscores method reliability in MCN. MCA demonstrates adaptability, showing significant F1 score improvements for "disorder" and the lowest false positive entry rates in both datasets, highlighting its efficacy across diverse medical contexts. This performance of "disorder" was competing with the supervised approach by Leaman et al.[67]

Variations in performance across datasets imply dataset-specific challenges, warranting further exploration for optimization. A detailed analysis comparing MCA across the dictionary-matching baseline approach in the n2c2 dataset and German EHRs underscores the need for a better NER approach and reduced false positives. This fully unsupervised method can serve as a starter for pre-annotations in languages lacking publicly available datasets, such as clinical narratives, significantly reducing manual annotation time.[68] Unlike other unsupervised methods,[51,69] our approach focuses on all semantic tags found in EHR narratives, potentially improving overall algorithm performance. However, it is important to acknowledge that adapting our method to new languages or terminologies may require language-specific preprocessing and domain-specific knowledge integration. Overall, our study lays the groundwork for exploring the practical applications of unsupervised MCA on real-world clinical narratives, potentially enhancing efficiency and accuracy in medical data annotation. Our approach also offers valuable insights into computational demands by estimating processing times at the line level, facilitating the understanding and targeted optimizations for enhanced system performance and resource allocation. Future research could explore techniques for automatic adaptation and scaling to diverse linguistic and

medical contexts, taking into account further validation and fine-tuning to ensure seamless integration and address challenges such as false positives, contextual errors, and the idiosyncratic nature of clinical language.

System limitation

The MCA method exhibits a significant drawback in generating numerous false positives across both datasets, undermining overall recall and precision. Semantic ambiguity, where a word or phrase holds multiple interpretations, poses a complex challenge in clinical NLP. Efforts to mitigate this issue, such as employing rule-based filters, result in a performance drop. Acronyms further contribute to semantic ambiguity, complicating the analysis of mentions within their original context.

In contrast to systems that restrict semantic tags to predefined categories, our approach adopts a generalized approach, accommodating diverse medical domains. However, this flexibility may challenge precision and coverage. Dataset limitations, particularly biased tag distribution, can skew the model. The standardized content of clinical terminology systems remains a challenge, exacerbated by the scarcity of publicly available training data. Our unsupervised MCA method effectively addresses this challenge by semantic types that can lead to misclassification and decreased performance in medical concept recognition, thus impacting practical applicability in clinical settings. Nevertheless, the currently high processing time of 60 s per document line must be considered when developing optimized versions of the algorithm in the future.

Conclusion and outlook

The alignment between language expressions in clinical sociolects and standardized content of clinical terminology systems remains a challenge, exacerbated by the scarcity of publicly available training data. Our unsupervised MCA method addresses this challenge effectively, particularly in the absence of training data for supervised machine learning approaches.

Our proposed method demonstrates suitability for identifying and annotating text mentions in clinical narratives using codes from terminology systems. It holds promise as an initial annotation step to support manual annotation tasks in the future. The achieved F1 score performance of 0.765 for MCN sets a baseline, to be further explored with advanced language model techniques such as ChatGPT in future investigations.

Recognizing the importance of addressing fragmented mentions, we intend to incorporate techniques, such as context-based modeling or neural sequence labeling, in future iterations. These enhancements aim to improve the coverage and accuracy of entity recognition, thereby enhancing overall effectiveness. Future improvements for our methodology include enhanced normalization filters, improved entity recognition, and reduced false positive rates to further support coding in the context of clinical narrative data.

Acknowledgements: The authors have no specific acknowledgments to declare for this research.

Contributorship: AA and MK designed the project and the processing workflow with feedback from RR and SS. AA and SS annotated the dataset. MK triggered the problem motivation and AA is responsible for the core implementation. All authors read and approved the final version of the manuscript.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: This study was approved by the Institutional Review Board (IRB) of the Medical University of Graz (30-496 ex 17/18).

Funding: The authors received no financial support for the research, authorship, and/or publication of this article.

Guarantor: Markus Kreuzthaler

Informed consent: Informed consent was waived because the data being studied was de-identified, as approved by the IRB of the Medical University of Graz (30-496 ex 17/18).

ORCID iD: Markus Kreuzthaler https://orcid.org/0000-0001-9824-9004

Note

8 https://tedboy.github.io/nlps/generated/generated/nltk.ngrams.html

References

1. Cowie MR, Blomster JI, Curtis LH, et al. Electronic health records to facilitate clinical research. *Clin Res Cardiol* 2017; 106: 1–9.
2. Schulz S, Daumke P, Romacker M, et al. Representing oncology in datasets: Standard or custom biomedical terminology? *Inf Med Unlocked* 2019; 15: 100186.
3. Kreuzthaler M, Brochhausen M, Zayas C, et al. Linguistic and ontological challenges of multiple domains contributing to transformed health ecosystems. *Front Med* 2023; 10.
4. Sung M, Jeong M, Choi Y, et al. Bern2: An advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics* 2022; 38: 4837–4839.
5. Allen KS, Hood DR, Cummins J, et al. Natural language processing-driven state machines to extract social factors from unstructured clinical documentation. *JAMIA Open* 2023; 6: ooad024.
6. Zhao S, Su C, Lu Z, et al. Recent advances in biomedical literature mining. *Brief Bioinf* 2021; 22: bbaa057.
7. Pattisapu N, Anand V, Patil S, et al. Distant supervision for medical concept normalization. *J Biomed Inform* 2020; 109: 103522.

8. Silva JF, Almeida JR and Matos S. Extraction of family history information from clinical notes: Deep learning and heuristics approach. *JMIR Med Inform* 2020; 8: e22898.
9. Howard J and Ruder S. Universal language model fine-tuning for text classification. *arXiv preprint arXiv: 180106146*, 2018.
10. Zhang Z and Chen ALP. Biomedical named entity recognition with the combined feature attention and fully-shared multi-task learning. *BMC Bioinf* 2022; 23: 458.
11. Chen L, Varoquaux G and Suchanek FM. A lightweight neural model for biomedical entity linking. *Proc AAAI Conf Artif Intell* 2021; 35: 12657–12665.
12. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process Syst* 2017; 30: 6000–6010.
13. Miftahutdinov Z, Kadurin A, Kudrin R, et al. Medical concept normalization in clinical trials with drug and disease representation learning. *Bioinformatics* 2021; 37: 3856–3864.
14. Kalyan KS and Sangeetha S. Bertmcn: Mapping colloquial phrases to standard medical concepts using BERT and highway network. *Artif Intell Med* 2021; 112: 102008.
15. Ji Z, Wei Q and Xu H. Bert-based ranking for biomedical entity normalization. *AMIA Jt Summits Transl Sci Proc* 2020; 2020: 269–277.
16. Cho H, Choi D and Lee H. Re-ranking system with Bert for biomedical concept normalization. *IEEE Access* 2021; 9: 121253.
17. Wajsbürt P, Sarfati A and Tannier X. Medical concept normalization in French using multilingual terminologies and contextual embeddings. *J Biomed Inform* 2021; 114: 103684.
18. Sung M, Jeon H, Lee J, et al. Biomedical entity representations with synonym marginalization. In: *Proceedings of the 58th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, 2020-07, Online, pp.3641–3650.
19. Lee J, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020; 36: 1234–1240.
20. Si Y, et al. Enhancing clinical concept extraction with contextual embeddings. *J Am Med Inform Assoc* 2019; 26: 1297–1304.
21. Huang K, Altosaar J and Ranganath R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint* arXiv:1904.05342, 2019.
22. Xu D, Gopale M, Zhang J, et al. Unified medical language system resources improve sieve-based generation and bidirectional encoder representations from transformers (BERT)-based ranking for concept normalization. *J Am Med Inform Assoc* 2020; 27: 1510–1519.
23. Silva JF, Antunes R, Almeida JR, et al. Clinical concept normalization on medical records using word embeddings and heuristics. *Stud Health Technol Inform* 2020 Jun 16; 270: 93–97.
24. Chen L, Fu W, Gu Y, et al. Clinical concept normalization with a hybrid natural language processing system combining multilevel matching and machine learning ranking. *J Am Med Inform Assoc* 2020; 27: 1576–1584.
25. Luo YF, Henry S, Wang Y, et al. The 2019 n2c2/UMass Lowell shared task on clinical concept normalization. *J Am Med Inform Assoc* 2020; 27: 1529-e1.
26. Lin Y-C, Hoffmann P and Rahm E. Enhancing cross-lingual biomedical concept normalization using deep neural network pretrained language models. *SN Comput Sci* 2022; 3: 387.
27. Kalyan KS and Sangeetha S. Target concept guided medical concept normalization in noisy user-generated texts. In: *Proceedings of deep learning inside out (DeeLIO): The first workshop on knowledge extraction and integration for deep learning architectures*. Association for Computational Linguistics, 2020 Nov, Online, pp.64–73.
28. Liu F, Shareghi E, Meng Z, et al. 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*. Association for Computational Linguistics, 2021, Online, pp.4228–4238.
29. Schwarz M, Chapman K and Häussler B. Multilingual medical entity recognition and cross-lingual zero-shot linking with facebook ai similarity search. *ceur-wsorg*, 2022.
30. Xu D and Miller T. A simple neural vector space model for medical concept normalization using concept embeddings. *J Biomed Inform* 2022; 130: 104080.
31. Borchert F, Llorca I, Roller R, et al. xmen: a modular toolkit for cross-lingual medical entity normalization. *arXiv preprint arXiv:231011275*, 2023.
32. Weston L, Tshitoyan V, Dagdelen J, et al. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *J Chem Inf Model* 2019; 59: 3692–3702.
33. Aronson AR and Lang FM. An overview of metamap: Historical perspective and recent advances. *J Am Med Inform Assoc* 2010; 17: 229–236.
34. Gorrell G, Song X and Roberts A. Bio-yodie: a named entity linking system for biomedical text. *arXiv preprint arXiv:181104860*, 2018.
35. Wu H, Toti G, Morley KI, et al. Semehr: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *J Am Med Inform Assoc* 2018; 25: 530–537.
36. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010; 17: 507–513.
37. Neumann M, King D, Beltagy I, et al. Scispacy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:190207669*, 2019.
38. Soysal E, Wang J, Jiang M, et al. CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines. *J Am Med Inform Assoc* 2018; 25: 331–336.
39. Whetzel PL, Noy NF, Shah NH, et al. Bioportal: Enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic Acids Res* 2011; 39: W541–W545.
40. Kraljevic Z, Searle T, Shek A, et al. Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit. *Artif Intell Med* 2021; 117: 102083.
41. Schwarz CM, Hoffmann M, Smolle C, et al. Structure, content, unsafe abbreviations, and completeness of discharge summaries: a retrospective analysis in a university hospital in Austria. *J Eval Clin Pract* 2021; 27: 1243–1251.

42. Wang C, Liu S, Yang H, et al. Ethical considerations of using ChatGPT in health care. *J Med Internet Res* 2023; 25: e48009.
43. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Commun Med* 2023; 3: 141.
44. Mati DN, Hamiti M, Susuri A, et al. Building dictionaries for low resource languages: challenges of unsupervised learning. *Ann Emerging Technol Comput (AETiC)* 2021; 5: 52–58.
45. Cardoso JR, Pereira LM, Iversen MD, et al. What is gold standard and what is ground truth? *Dental Press J Orthod* 2014; 19: 27–30.
46. Liu K and El-Gohary N. Unsupervised named entity normalization for supporting information fusion for big bridge data analytics. In: *Advanced computing strategies for engineering: 25th EG-ICE international workshop 2018*, Lausanne, Switzerland: Springer, 10–13 June 2018, Proceedings, part II 25, pp.130–149.
47. Peng Q, et al. Unsupervised cross-domain named entity recognition using entity-aware adversarial training. *Neural Netw* 2021; 138: 68–77.
48. Yan H, et al. Unsupervised cross-lingual model transfer for named entity recognition with contextualized word representations. *PLoS ONE* 2021; 16: e0257230.
49. Nath N, Lee S-H and Lee I. Application of specialized word embeddings and named entity and attribute recognition to the problem of unsupervised automated clinical coding. *Comput Biol Med* 2023; 165: 107422.
50. Yan C, et al. Enhancing unsupervised medical entity linking with multi-instance learning. *BMC Med Inform Decis Mak* 2021; 21: 1–10.
51. Tahmasebi AM, Zhu H, Mankovich G, et al. Automatic normalization of anatomical phrases in radiology reports using unsupervised learning. *J Digit Imaging* 2019; 32: 6–18.
52. Karadeniz I and Özgür A. Linking entities through an ontology using word embeddings and syntactic re-ranking. *BMC Bioinf* 2019; 20: 1–12.
53. Zhang S and Elhadad N. Unsupervised biomedical named entity recognition: experiments with clinical and biological texts. *J Biomed Inform* 2013; 46: 1088–1098.
54. Xu J, et al. Unsupervised medical entity recognition and linking in Chinese online medical text. *J Healthc Eng* 2018: 130–149.
55. Chang E and Mostafa J. The use of SNOMED CT, 2013-2020: a literature review. *J Am Med Inform Assoc* 2021; 28: 2017–2026.
56. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004; 32: D267–D270.
57. SNOMED International. SNOMED CT starter guide. International release (US English), 2023. https://confluence.ihtsdotools.org/pages/viewpage.action?pageId=26837109.
58. Schulz S, Del-Pinto W, Han L, et al. Towards principles of ontology-based annotation of clinical narratives. In: *Proceedings of the international conference on biomedical ontologies*, 2023, August 28th-September 1st, 2023, Brasilia, Brazil.
59. Luo YF, Sun W and Rumshisky A. MCN: a comprehensive corpus for medical concept normalization. *J Biomed Inform* 2019; 92: 103132.
60. Klie JC. INCEpTION: Interactive machine-assisted annotation. In: *DESIRES*, 2018-08, Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, Santa Fe, New Mexico: Association for Computational Linguistics, pp.5–9.
61. Nik DH, Kasác Z, Goda Z, et al. Building an experimental German user interface terminology linked to SNOMED CT. *Stud Health Technol Inform* 2019 Aug 21; 264: 153–157.
62. Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing, 2020. arXiv:2007.15779.
63. Conneau A, Khandelwal K, Goyal N, et al. Unsupervised cross-lingual representation learning at scale. *CoRR*, 2019; abs/1911.02116. http://arxiv.org/abs/1911.02116. 1911.02116.
64. Johnson J, Douze M and Jégou H. Billion-scale similarity search with GPUs. *IEEE Trans Big Data* 2019; 7: 535–547.
65. Abdulnazar A, Kreuzthaler M, Roller R, et al. SapBERT-based medical concept normalization using SNOMED CT. *Stud Health Technol Inform* 2023; 302: 825–826.
66. Rao GA, Srinivas G, Rao KV, et al. A partial ratio and ratio based fuzzy-wuzzy procedure for characteristic mining of mathematical formulas from documents. *ICTACT J Soft Comput* 2018; 8: 1728–1732.
67. Leaman R, Khare R and Lu Z. Challenges in clinical natural language processing for automated disorder normalization. *J Biomed Inform* 2015; 57: 28–37.
68. Kholghi M, Sitbon L, Zuccon G, et al. Active learning reduces annotation time for clinical concept extraction. *Int J Med Inform* 2017; 106: 25–31.
69. Zhang Y, Ma X and Song G. Chinese medical concept normalization by using text and comorbidity network embedding. In: *2018 IEEE international conference on data mining (ICDM)*, 2018-11, IEEE Xplore, pp.777–786.

Appendix 1. Types of errors encountered along with their occurrence rate for both embedding-based MCN and MCA on the n2c2 dataset

The following types of errors were observed during error analysis for both the embedding-based MCN on the test mentions and MCA on the test documents.

Contextual errors. Fails to capture the meaning of the word/n-gram when it is dependent on the surrounding context.

Embeddings-based MCN error rate: 32.6%.

MCA error rate: 42.6% (en), 46.5% (de).

Sentence from gold standard: "There were diffuse ST segment and T-wave lightgrayabnormalities, which were nonspecific."

Candidate term: "abnormalities."

Gold standard target: 55930002, "ECG ST segment changes."

Output: 263654008, "Abnormal."

Granularity errors. Fails to distinguish between different levels of detail or granularity.

Embeddings-based MCN error rate: 14.3%.

MCA error rate: 12.7% (en), 12.8% (de).

Sentence from gold standard: "She has also had some discomfort in her lightgrayleft lower abdomen and notes diarrhoea every 4–5 days."

Candidate term: "left lower abdomen."

Gold standard target: 68505006, "Left lower quadrant of abdomen."

Output: 1017212007, "Left abdominal lumbar region."

Analogy errors. Fails to understand and generate correct analogies.

Embeddings-based MCN error rate: 24.6%.

MCA error rate: 12.8% (en), 8.5% (de).

Sentence from gold standard: "The patient had been taking his usual medications and using his lightgraynasal oxygen at home."

Candidate term: "nasal oxygen."

Gold standard target: 371907003, "Oxygen administration by nasal cannula."

Output: 71786000, "Intranasal oxygen therapy."

Similarity errors. Fails to accurately capture the semantic similarity or relatedness between words or phrases.

Embeddings-based MCN error rate: 8.4%.

MCA error rate: 3.8% (en), 6.7% (de).

Sentence from gold standard: "lightgrayEstratab."

Candidate term: "Estratab."

Gold standard target: 126099009, "Esterified estrogen."

Output: 446265008 ,"Estrilda."

Wrong IDs. Fails in assigning correct label or identification to a given input.

Embeddings-based MCN error rate: 7.5%.

MCA error rate: 16.0% (en), 11.1% (de).

Sentence from gold standard: "He was admitted to the Short Stay Unit, given lightgrayAncef and Gentamicin per the team for antibiotic prophylaxis and observed overnight"

Candidate term: "Ancef."

Gold standard target: 387470007, "Cefazolin."

Output: 81123006, "Interleukin-5."

Nugatory IDs. Assigning non-existing IDs

Embeddings-based MCN error rate: 7.3%.

MCA error rate: 7.4% (en), 4.1% (de).

Sentence from gold standard: "The patient is a 78-year-old female who has had osteoarthritis and noted the sudden onset of lightgrayleft knee pain in 09/89. "

Candidate term: "left knee pain."

Gold standard target: 468251000124107, "Not Valid ID."

Output: 287047008, "Pain in left leg."

Acronym errors. Fails to correctly interpret or expand an acronym within the given context.

Embeddings-based MCN error rate: 5.1%.

MCA error rate: 4.2% (en), 9.6% (de).

Sentence from gold standard: "Cholecystectomy in 1994, colonoscopy 2004, status post tonsillectomy, status post appendectomy, status post lightgrayORIF of left wrist, status post left ear surgery."

Candidate term: "ORIF."

Table 5. Evaluation results of MCN on the n2c2 and German EHRs dataset using the SNOMED CT embedding space.

Dataset	Model	P	R	F1
n2c2	SapBERT	0.776	0.755	0.765
	Coder-eng	0.736	0.701	0.706
German EHRs	SapBERT-XLMR-large	0.599	0.521	0.557
	Coder-all	0.436	0.435	0.430

MCN: medical concept normalization; SNOMED CT: systematized nomenclature of medicine–clinical terms; EHR: electronic health record; P: precision; R: recall; F1: F1 score; SapBERT: self-alignment pre-training for bidirectional encoder representations from transformers.

Table 6. Evaluation results of MCN on the n2c2 dataset using the SNOMED CT embedding space enriched with n2c2 training data–SNOMED CT*.

Dataset	Model	P	R	F1
n2c2 dataset	SapBERT	0.889	0.857	0.872

MCN: medical concept normalization; SNOMED CT: systematized nomenclature of medicine–clinical terms; P: precision; R: recall; F1: F1 score; SapBERT: self-alignment pre-training for bidirectional encoder representations from transformers.

Gold standard target: 133863002, "open reduction with internal fixation."

Output: 413042008, "Immature reticulocyte fraction."

Spelling errors. Mistake or deviation from the correct spelling of an input word.

Embeddings-based MCN error rate: 1.1%.

MCA error rate: 3.0% (en), 5.8% (de).

Sentence from gold standard: "diabetes mellitus with lightgraydiabetic renopathy, renovascular occlusive disease, with thrombosis of the right renal artery, hypertension, probably renal vascular, hypertensive cardiac disease with history of congestive heart failure."

Candidate term: "diabetic renopathy."

Gold standard target: 4855003,"Diabetic retinopathy."

Output: 127013003, "Diabetic nephropathy."

Appendix 2. Medical concept normalization

Based on the results of Tables 5 and 6, SapBERT models outperformed the Coder models, and therefore we evaluated the MCN also using SNOMED CT embedding space enriched with n2c2 training data—SNOMED CT*.

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 September 2024, accepted 25 September 2024, date of publication 2 October 2024, date of current version 18 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3472500

RESEARCH ARTICLE

Large Language Models for Clinical Text Cleansing Enhance Medical Concept Normalization

AKHILA ABDULNAZAR[1,2], ROLAND ROLLER[3], STEFAN SCHULZ[1], AND MARKUS KREUZTHALER[1]

[1]Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, 8036 Graz, Austria
[2]CBmed GmbH–Center for Biomarker Research in Medicine, 8010 Graz, Austria
[3]German Research Center for Artificial Intelligence (DFKI), 10559 Berlin, Germany

Corresponding author: Markus Kreuzthaler (markus.kreuzthaler@medunigraz.at)

ABSTRACT Most clinical information is only available as free text. Large language models (LLMs) are increasingly applied to clinical data to streamline communication, enhance the accuracy of clinical documentation, and ultimately improve healthcare delivery. This study focuses on a corpus of anonymized clinical narratives in German. On the one hand it evaluates the use of ChatGPT for text cleansing, i.e., the automatic rephrasing of raw text into a more readable and standardized form, and on the other hand for retrieval-augmented generation (RAG). In both tasks, the final goal was medical concept normalization (MCN), i.e., the annotation of text segments with codes from a controlled vocabulary using natural language processing. We found that ChatGPT (GPT-4) significantly improves precision and recall compared to simple dictionary matching. For all scenarios, the importance of the underlying terminological basis was also demonstrated. Maximum F1 scores of 0.607, 0.735 and 0.754 (i.e, for top 1, 5 and 10 matches) were achieved through a pipeline including document cleansing, bi-encoder-based term matching based on a large domain dictionary linked to SNOMED CT, and finally re-ranking using RAG.

INDEX TERMS ChatGPT, medical concept normalization, retrieval augmented generation, text cleansing.

I. INTRODUCTION

Electronic health records include both structured data and unstructured narrative content. Structured data are easy to analyze, particularly when coded by standardized semantic identifiers. In contrast, clinical narratives exhibit the whole range of phenomena that emerge when humans take notes in a hurry. To bridge the semantic gap between unstructured data and semantically explicit, coded information, natural language processing (NLP) methods such as named entity recognition and medical concept normalization (MCN) have been used [1]. Traditionally, the two tasks have been tackled separately, with most approaches focusing on either entity recognition [2], [3] or normalization independently [4], [5]. However, optimizing this process relies heavily on annotated data [6]. Current trends in large language models (LLMs) have shown strong performance across various NLP tasks without requiring extensive parameter tuning or training [7]. This suggests their potential and versatility for better few-shot and transfer learning abilities, indicating that they may ultimately serve as a comprehensive framework for various NLP tasks [7], [8], [9], [10], [11]. In this article, we explore how the potential of LLMs can be harnessed to enhance performance in MCN.

MCN links words and phrases (entity mentions) to standardized and language-independent codes in controlled vocabularies and ontologies [12]. MCN is crucial for information extraction and heavily relies on the coverage of language resources, particularly terminology systems, often further specialized as controlled vocabularies, thesauri, statistical classifications and ontologies. English has a big advantage over all other languages because it is by far the language best

The associate editor coordinating the review of this manuscript and approving it for publication was Ines Domingues.

covered by these systems. Although achieving acceptable MCN results is still challenging for English-language clinical texts, it is largely more difficult for clinical texts in other languages.

Over the past decade, deep learning and language models have revolutionized NLP, enabling machines to understand and generate human language with unprecedented accuracy. These advancements began with the development of pre-trained word embeddings from large non-annotated text corpora, which have shown their usefulness for MCN, unsurprisingly with a strong bias towards English texts [13], [14]. ELMo introduced contextual word embeddings, advancing the cutting-edge for several major NLP benchmarks [15]. The Generative Pre-trained Transformer (GPT) further minimized task-specific parameters by allowing simple fine-tuning for downstream tasks [16]. Unlike earlier models such as ELMo and GPT, which used unidirectional language models, BERT introduced masked language models for pre-training bidirectional representations, significantly improving performance, as evidenced for eleven NLP tasks [17], [18]. One of the currently most popular LLM, ChatGPT [19], incorporates generative techniques to produce contextually relevant text. Not specifically trained on medical data, ChatGPT has showcased its versatility in various research and healthcare applications [20], including diagnosis support, treatment optimization, and medical question-answering. However, the proprietary nature of many LLMs, particularly GPT and their opaque, ''black box'' character has raised concerns about transparency, accountability, and potential biases regarding their deployment in the context of health care [21].

This paper reports on the combination of BERT and generative models for MCN in German texts, supported by two German medical terminology resources linked to the clinical terminology standard SNOMED CT, an ontology with more than 350,000 units of meaning (concepts) [22], [23], [24]. For MCN, we use a bi-encoder model specifically pre-trained to understand biomedical terminology [25], [26]. On the one hand, ChatGPT's, GPT-4 [27] architecture is used in a preprocessing step to make raw clinical narratives more uniform and interpretable, thus not only easier for human understanding but also for MCN. On the other hand, GPT-4 is used to optimize the selection of candidates for concept matching using retrieval-augmented generation (RAG).

Our investigation is structured as follows. First, we report on how we created a German-language annotated corpus for MCN, using SNOMED CT as the annotation vocabulary, adhering to the annotation guidelines of the n2c2 (National NLP Clinical Challenge) normalization task [28]. Then we expose how GPT-4 was prompted to perform text cleansing. This resulted in two datasets: (i) raw data and (ii) cleansed data. Simple dictionary matching and bi-encoder-based matching were then employed for MCN. Additionally, the RAG capability of GPT-4 was implemented to re-rank the best match from the mapped list.

II. RELATED WORK

A. MEDICAL CONCEPT NORMALIZATION (MCN)

Conventional MCN includes dictionary lookup, deep learning, retrieval, and ranking methods [29], [30], [31], [32]. Deep learning models such as convolutional neural networks and recurrent neural networks with pre-trained word embeddings had shown significant improvements in MCN accuracy, surpassing the previous state-of-the-art [33]. One method involves encoding terminology labels and synonyms into a vector space that uses text and graph embeddings to represent text sequences as vectors. Using a semantic proximity measure, e.g. cosine similarity, the nearest terminology item can be found – and, in consequence, the most appropriate terminology concept – for a given input, resulting in improved classification accuracy across benchmark datasets [34]. BERT models have demonstrated superior performance for MCN compared to other architectures [35], as they excel in managing multilingual data and better capture contextual information [1], [18], [36], [37], [38]. In the 2019 n2c2/UMass Lowell task on MCN, the most accurate approach involved a deep learning architecture with a pre-trained SciBERT layer [28]. Self-alignment pretraining for biomedical entity representation (SapBERT), is a pretraining scheme designed for learning representations of biomedical entities. It outperforms existing models in MCN tasks and achieves cutting-edge results across various datasets without the need to fine-tune the labeled data of the task [25]. Fine-tuning SapBERT established a new benchmark for MCN, including cross-lingual normalization [39], [40], a task also tackled by the modular xMEN [41] system, which uses unsupervised candidate generation and supervised cross-encoders for re-ranking, surpassing previous state-of-the-art performance on diverse benchmarks.

These approaches offer advantages such as improved classification accuracy, scalability to millions of target concepts, and efficient accommodation of growing lexicon sizes [34]. However, they require manual mapping of training data, show difficulty in mapping unseen concepts, and require the retraining of models whenever new content is added. Limitations also arise from the dynamic and fragmented nature of clinical language, a genre replete with spelling errors, jargon expressions, and shorthand expressions, which require constant contextual corrections and disambiguation [42]. Kartchner et al. [43] suggested that LLM-based normalization can enhance the performance of existing models and improve the quality and accuracy of LLM-generated text.

B. LLMS IN THE CLINICAL DOMAIN

Great expectations are associated with the integration of LLM into clinical data management workflows. These models have shown excellent language understanding skills across many domains and performed well in tasks such as summarizing [44]. Agrawal et al. [45] described how ChatGPT optimizes content retrieval and saves time for healthcare

professionals. The provision of concise patient summaries facilitated rapid access to essential information [46]. ChatGPT also demonstrated the potential to generate diagnostic reports or recommendations based on past clinical data, aiding in identifying patterns and connections not immediately apparent to clinicians [47]. Together with genetic information and biomarkers, ChatGPT was shown to offer tailored treatment recommendations and predict individual responses to therapies [48]. In telemedicine, it has facilitated virtual patient-physician interactions, assisted in triaging, and provided remote guidance for home care [49]. LLMs have also shown remarkable potential in biomedical applications, particularly in named entity recognition, by leveraging strategic prompting and integration of external resources. While BERT excels in precision, GPT surpasses in recall and F-score, making it more comprehensive in identifying relevant entities [11], [50]. Nevertheless, seamless integration of LLMs into existing clinical workflows and systems is crucial for their effective use in healthcare.

Ethical issues, privacy concerns, and technical limitations have constantly been discussed [21], particularly in the current context where the leading LLMs are proprietary, and the performance of open models that can be run on premises has lagged. There is a broad consensus that integrating LLMs in healthcare poses risks, including bias in training data, incorrect content, lack of explanations, and reduced need for human expertise. Privacy breaches, legal disputes, interpretability challenges, and misinformation are also concerns [51]. Mitigation requires accurate benchmarking, the use of explainable AI methodologies, and rigorous certification before deployment as medical products.

1) TEXT CLEANSING

Language models have been shown to help cleanse the typical hastily written clinical jargon by correcting spelling, standardizing terms, and improving text clarity [52], [53], across document types such as discharge summaries, radiology reports and other clinical narratives [54] and reaching a satisfactory quality level [55]. Even highly elliptical texts overloaded with short forms can be organized in a coherent manner [56]. For instance, a physician can instruct a language model to include specific elements and to briefly explain some ideas, allowing it to rapidly generate a formal discharge summary. Preliminary studies suggest that ChatGPT could improve the quality of discharge summaries [57], potentially reducing the risk of miscommunication and improving patient care [58]. However, while these generated summaries may appear well-structured, their accuracy and reliability must be rigorously evaluated to ensure they meet clinical standards.

2) RETRIEVAL-AUGMENTED GENERATION (RAG)

RAG is a common practice to address the limitations of models that may not contain all necessary information or have become outdated. This technique combines retrieval with text generation models to incorporate additional information dynamically at runtime [59], [60]. The retrieval component fetches relevant information from a database in response to a query, while the generative language model uses this information to craft contextually relevant responses. Thus, the model's output generation capabilities are enhanced without requiring costly re-training cycles.

Advanced RAG additionally incorporates sophisticated pre-retrieval and post-retrieval processes. One critical post-retrieval aspect in advanced RAG is "Re-Rank", which reorders the retrieved documents by relevance [61]. It employs algorithms that adjust the ordering based on criteria such as document diversity or relevance to the query. Re-ranking aims to present the most pertinent information to the LLM, thereby improving the quality and relevance of the generated responses [62]. The use of re-ranking in ChatGPT's RAG approach can lead to more accurate and contextually relevant responses, as it allows the model to consider the latest information from knowledge bases and adjust its responses accordingly. This can be particularly beneficial in the clinical domain, where the accuracy and relevance of responses are critical for effective communication and decision-making. A recent study has explored LLMs for few-shot information extraction tasks and introduced a novel paradigm to enhance their effectiveness [63]. Using prompting strategies and an adaptive filter-then-rerank approach, the system achieved notable improvements (averaging 2.4% F1 gain) over existing methodologies, showcasing the potential of LLMs to tackle challenging information extraction tasks.

Besides this broad scope of LLM applications presented since 2022, to the best of our knowledge, no work has prompted a conversational LLM with the specific goal of optimizing narratives for MCN. Following the suggestions by [43], our goal is to leverage LLMs to enhance MCN. By combining the text cleansing capability of ChatGPT with RAG, our objective is to improve the accuracy and efficiency of MCN.

III. METHODOLOGY AND DATA

That pre-processing of clinical narratives using an LLM improves the precision of the MCN in various settings is a central hypothesis of our work, investigating a scenario where SapBERT is used for similarity search. To validate this hypothesis, we created an annotated corpus for MCN using SNOMED CT as an annotation vocabulary. Then, we used GPT-4 [27] for cleansing the narratives and as a re-ranker for the concepts retrieved by SapBERT. The results were compared to a simple dictionary lookup baseline. Since the proposed methodology does not include any pretraining, it should be considered as an unsupervised way of MCN using the advantages of LLMs. The diagrammatic representation of the study is shown in Figure 1.

A. DATASET CREATION

We created a corpus of clinical texts in German, representative of the clinical information system of KAGes, an Austrian network of public hospitals. Ten discharge

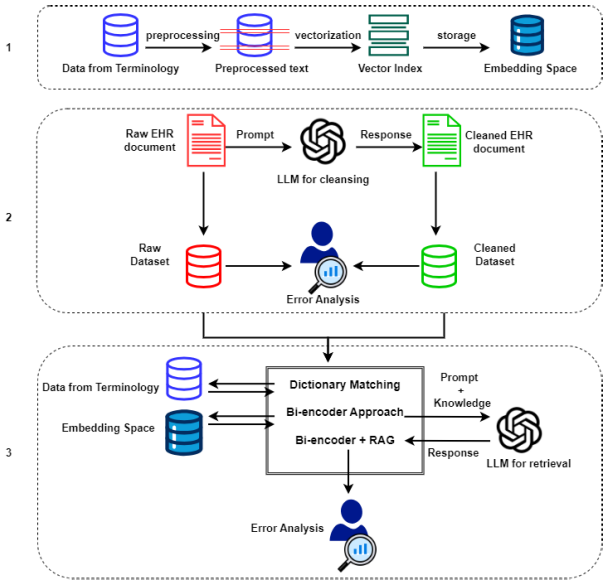

FIGURE 1. Diagrammatic representation of the proposed methodology, (1) generation of the embedding space base from two German custom terminologies, (2) cleansing of the clinical narratives using GPT-4 and its error analysis, (3) MCN using different approaches such as dictionary lookup and bi-encoders. Re-ranking the retrieved concepts by the bi-encoder approach using GPT-4 and error analysis.

summaries from different clinical departments had first been manually anonymized. Then, their content was moderately alienated so that it could be safely assumed not to denote any particular patient. This was done by the third author, a medical expert, in a way that these synthetic documents appeared maximally authentic both in content and structure, following an approach described in [64]. The assignment of SNOMED CT [23] codes was carried out collaboratively by two experts, one experienced in clinical practice and the other in biomedical sciences, following the n2c2 Annotation Guidelines [28]. The output consisted of 660 annotated surface terms.

B. PROMPT FOR CLEANSING

Following the initial phase of corpus construction, a data cleansing procedure was initiated to refine the quality of corpus content. The ten narratives underwent cleansing using two types of GPT-4 prompts. A first general prompt aimed at standardization by expanding acronyms and correcting spelling errors to ensure clarity and consistency. This design was inspired by practical guidelines such as those provided by Google Gemini [65], which emphasize the importance of clarity in the prompts. A second, more specific prompt included these aims and additionally aimed at the enrichment of branded drug names by their corresponding substance names, which can then be aligned with SNOMED CT codes (SNOMED CT does not contain any brand names). This approach was informed by best practices that advocate for detailed and specific instructions to improve accuracy and relevance. Upon cleansing, terms aligned with the annotated corpus were extracted from the narratives. This process yielded two distinct datasets: one from the raw narratives and the other one from the cleansed ones. Each dataset was designed to meet specific analytical objectives; the second prompt was preferred for its detailed instructions and the potential to improve the overall completeness of the dataset. The prompts are as follows, accompanied by an example:

System: You are an expert in the clinical domain.

Prompt: Standardize and transform the given German clinical narrative into standardized language without abbreviations and spelling errors in German. Any abbreviations should be expanded into the corresponding long form, any existent spelling errors should be corrected. The corresponding substance should be added in round brackets after the given drug name for any drug or pharmaceutical name found in the clinical narrative. All input information should be considered and be represented in the output. No additional explanations should be added other than the transformed clinical narrative text. The clinical narrative is found below:

Example:

Raw text snippet[1]*: "Anamn. besteht eine dilat. CMP, eine incipiente KHK sowie eine intermitt. VHFA."*

Cleansed text snippet: "In der Anamnese besteht eine dilatative Kardiomyopathie, eine beginnende koronare Herzkrankheit sowie ein intermittierendes Vorhofflimmern."

C. TERMINOLOGY BASE

To standardize clinical terms, MCN links information to codes from terminology systems, which act as semantic identifiers to domain terms in one or more languages. The Metathesaurus of the Unified Medical Language System (UMLS) [66] links lexical and conceptual information between approximately 200 different biomedical terminology systems, encompassing MeSH, SNOMED CT, ICD-10, MedDRA, and RxNORM. UMLS was created and is maintained by the U.S. National Library of Medicine, which explains its strong focus on English. To a minor extent, it includes lexical content from other languages, such as Spanish or German. Due to the terminology cross-links, each concept can be enhanced by (quasi-)synonymous terms from different terminologies, including non-English ones.

SNOMED CT translation to German is still in an early stage. Therefore, we used two custom German term collections linked to SNOMED CT codes, in order to build two embedding spaces. The first custom terminology, UMLS_DE, was created on the fly by extracting all German terminological units from the UMLS Metathesaurus for which a connection to some SNOMED CT code could be established via a common identifier (CUI). Thus, approximately 79,000 medical terms for 41,000 SNOMED CT concepts were harvested. The second custom terminology, IT_DE, consists of a German Interface Terminology built semi-automatically [67] during the past ten years. The

[1]Raw: "H/O dil CMP, beginning CHF and intermitt AFib"; Cleansed: "History of dilatative cardiomyopathy, beginning congestive heart failure and intermittent atrial fibrillation."

extract used for this study links approximately 2.5 million manually harvested and automatically combined clinical terms to 278,000 SNOMED CT concepts. Both custom terminologies can be described as term collections annotated with SNOMED CT codes without claiming to be SNOMED CT translations in a proper sense.

D. PRE-PROCESSING

A series of terminology pre-processing steps were done to enhance the readiness of the data, including the removal of special characters, conversion to lowercase, and vectorization using a cross-lingual SapBERT trained with UMLS 2020AB (all languages) and XML-RoBERTa (large) [68] as the underlying model. Subsequently, the processed data is indexed via FAISS, an open-source library to perform fast similarity searches in high-dimensional vector spaces [69], facilitating efficient retrieval and utilization in subsequent analyses and applications. The entries in both custom terminologies (UMLS_DE and IT_DE) are then represented as 768-dimensional embeddings and FAISS-indexed and create the corresponding embedding space. The cosine similarity function is used for vector similarity matching.

E. MCN–MEDICAL CONCEPT NORMALIZATION

The following presents different normalization techniques used in this work, starting with a baseline method of dictionary mapping, followed by a bi-encoder approach, and further improving it with an RAG approach.

1) BASELINE APPROACH

A baseline for MCN was created using simple dictionary matching to identify correspondences with SNOMED CT codes by matching the text in the clinical corpus against the custom terminologies UMLS_DE and IT_DE. The process included pre-processing the clinical text as detailed in Section (III-D) and ensuring exact matches to identify the correspondences.

2) BI-ENCODER APPROACH

The input clinical terms underwent the pre-processing steps detailed in Section (III-D), followed by using cross-lingual SapBERT to convert the clinical terms into vectors. Then, these pre-processed and vectorized terms were submitted to a cosine similarity search within the embedding space using FAISS. A threshold of 0.9 had been set for similarity matching to ensure precision. The score was chosen based on the average similarity scores observed between known synonyms and the preferred terms in SNOMED CT, aiming to reduce ambiguity in the terms identified. This process retrieved the top ten candidate concepts from the embedding space.

3) RETRIEVAL-AUGMENTED GENERATION (RAG) APPROACH

In this approach, the ten candidate concepts retrieved by the bi-encoder approach (in Section (III-E2)), were presented to the GPT-4 model, along with their input terms and contexts. A prompt was created that instructs the model to re-rank these candidates based on contextual relevance. We used two prompts in experiments across four examples, selecting the most effective prompt based on its ranking performance. The following refers to the prompt used for this process.

System: As an expert ranker of related words tailored to specific contexts, your role is pivotal in identifying the most pertinent terms within a given context.

Prompt: Upon receiving an input text along with its context and a predefined list of 10 terms, your task is to re-rank these terms based on their contextual significance, prioritizing the most suitable choices at the top. Importantly, ensure that the re-ranked list includes only the terms provided in the input list, without filtering or adding any additional terms. Guideline: Re-rank the terms based on their contextual relevance, optimizing their alignment with the provided context. Output Always present the re-ranked list without any additional explanations in the format [term: id].

TABLE 1. Summary of ChatGPT improvements at narrative and surface term level.

Evaluation metric	Surface terms
Average number of words changed per narrative	7.72%
Average number of lines changed per narrative	67.85%
Missing annotated spans	0.45%
Changed annotated spans after cleansing	45%
Changes with potential good influence:	
Well-formed	31.0%
Synonym	2.0%
Changes with potential neutral influence:	
Synonym	3.0%
Misspell	1.0%
Changes with potential bad influence:	
Wrong	3.0%
Misspell	1.0%
Incomplete	1.0%
Hypernym	2.5%
Hyponym	0.5%

IV. RESULTS

Table 1 details the evaluation of the cleansing approach, including systematic analysis and medical content evaluation by an expert. The raw and cleansed narratives were thoroughly examined, resulting in a 7.72% reduction in word count and a 67.85% decrease in line numbers. A detailed comparison of surface terms between the raw and cleansed texts revealed that 0.45% (3 out of 660 annotated terms) were missing after cleansing, necessitating the addition of surface terms from the raw narratives to ensure completeness. Furthermore, after cleansing, 45% of the annotated spans were found to be changed. Changes were categorized according to their potential influence: well-formed and beneficial synonyms, suggesting a good impact (33.0%). Neutral changes include neutral synonyms and misspellings (4.0%). Potentially negative changes (8.0%) include wrong entries, misspellings, incomplete information, hypernyms,

TABLE 2. Concept normalization using different approaches, referred to under the column 'Method' based on SNOMED CT codes, using different terminologies as in column 'Source', (i) a custom terminology for German extracted from the UMLS Metathesaurus (UMLS_DE) and (ii) the German Interface Terminology for SNOMED CT (IT_DE). Column 'Data' shows the different types of data used, (i.e., raw and cleansed). The performance of these approaches is evaluated for precision (P), recall (R) and F1 score for top one, five and ten matches.

Method	Source	Data	P@1	R@1	F1@1	P@5	R@5	F1@5	P@10	R@10	F1@10
Dictionary matching	UMLS_DE	raw text	0.137	0.115	0.122						
		cleansed text (GPT-4)	0.195	0.173	0.178						
	IT_DE	raw text	0.330	0.284	0.296						
		cleansed text (GPT-4)	0.339	0.282	0.297						
Bi-encoder	UMLS_DE	raw text	0.252	0.232	0.232	0.343	0.323	0.325	0.366	0.341	0.346
		cleansed text (GPT-4)	0.306	0.285	0.286	0.440	0.435	0.432	0.458	0.453	0.450
	IT_DE	raw text	0.593	0.523	0.542	0.673	0.629	0.641	0.702	0.656	0.669
		cleansed text (GPT-4)	**0.618**	**0.552**	**0.568**	0.764	0.724	0.735	**0.782**	**0.745**	**0.754**
RAG	UMLS_DE	raw text	0.253	0.244	0.241	0.351	0.332	0.333			
		cleansed text (GPT-4)	0.321	0.300	0.298	0.444	0.443	0.436			
	IT_DE	raw text	0.615	0.558	0.572	0.686	0.641	0.653			
		cleansed text (GPT-4)	**0.646**	**0.595**	**0.607**	**0.768**	**0.724**	**0.735**			

and hyponyms. These findings highlight the importance of the cleansing process in reducing narrative length and improving the clarity of terms. However, they also reveal potential risks associated with content loss or alteration. To address this, we ensure transparency and reliability through manual evaluation by a domain expert. AI in clinical settings should assist, not replace, human decision-making, with rigorous validation and testing to catch errors and ensure reliability.

Subsequently, the datasets underwent MCN using three distinct approaches: dictionary matching, the bi-encoder method, and the RAG for re-ranking. Table 2 details the performance of both raw and cleaned corpora across the two custom terminologies (i.e, (UMLS_DE) and (IT_DE)) and approaches. Cleansing text with GPT-4 often improves precision and recall, leading to higher F1 scores when compared to using raw text in all three approaches. Moving from dictionary matching to the bi-encoder approach showed a significant 91.25% increase in the F1 score, indicating a substantial performance improvement using the text cleaned by GPT-4 and the IT_DE terminology. Notably, the unsupervised bi-encoder method using GPT-4-cleaned text achieved an F1 score of 0.568 for top 1 matches, 0.735 for top 5 matches and 0.754 for top 10 matches. Furthermore, transitioning from the bi-encoder to the RAG resulted in a smaller but notable 6.87% gain in F1 score for the top 1 matches, improving from 0.568 to 0.607, demonstrating continued enhancement in performance within the same setting for data and terminology. The re-ranked terms have been cross-checked using an algorithm to ensure that no changes were made to the list of terms other than their order, thereby ensuring that no hallucinations occurred. The IT_DE terminology consistently shows higher performance metrics compared to UMLS_DE. This observation, combined with the performance improvements, highlights the efficacy of these methods in enhancing the efficiency in MCN systems.

A. ERROR ANALYSIS

We performed a qualitative error analysis on the top ten candidates retrieved by the best-performing scenario, i.e. bi-encoder on the cleansed corpus using the custom terminology IT_DE. Typical errors are described in the following. We distinguish between document-cleansing errors and concept-matching errors. Out of the 660 terms, for the top 10 matches, 25.6% (169/660) were wrong matches, of which 35.5% (60/169) were due to document cleansing.

1) DOCUMENT CLEANSING ERRORS

These errors affected the text in a way that the meaning of the cleansed text was altered. E.g., the meaning (in English) of "Abdomen: soft abdominal wall, no tenderness, intense bowel sounds, no flank pain" can be unambiguously obtained from the following passage in the German raw text: "Abd. BD weich, kein DS, DG's rege, NL frei.". Cleansing transformed this to "Abdomen: Bauchdecke weich, kein Druckschmerz, Darmgeräusche rege, Leber nicht vergrößert", where "NL frei" ("no flank pain", annotated in the gold standard with "300447004 Kidney non-tender (situation)"), was erroneously rephrased to "Leber nicht vergrößert" ("liver not enlarged").

A similar phenomenon characterises the next example, which means in English:"The cardiac activity is rhythmical, with normal sinus rhythm" and appears in raw text as "HA rhythmisch, nc.". A cardiologist understands from the context that "nc." is the abbreviation of "normocard", which justifies the annotation of the whole passage with "64730000 Normal sinus rhythm (finding)". Cleansing transformed this passage to "Herzaktivität ist rhythmisch, normal korrigiert", which ended up being annotated by the system with "263699000 Cardiac activity (observable entity)". This is not wrong, like in the previous example, but not specific enough. The (non-existing) expansion of "nc." to "normal korrigiert" was fortunately not annotated by the

system, but clearly shows the tendency toward hallucination when the language model encounters unknown expressions in uncommon contexts. When analyzing the document cleansing errors, 42% were due to incorrect expansion of short forms, while 20% were caused by adding incorrect substances to drug names.

2) MAPPING ERRORS

These were frequent errors, independent of the cleansing process. So had ''Mäßig inhomogene Parenchymstruktur der Leber wie bei Steatose/ LPS.'' (which means ''Moderately heterogeneous parenchymal structure of the liver as in steatosis / Liver Parenchymal Steatosis'') the gold standard annotation ''197321007 Steatosis of liver (disorder)'', because the annotator concluded from the context that ''steatosis'' here means ''steatosis of the liver''. However, the system mapped it to the parent concept ''1187537008 Steatosis (disorder), along with ''127879008 Structure of parenchyma of liver (body structure)''. This reveals a fundamental problem of compositional terminologies, *viz.* that different sequences of pre-coordinated concepts can be derived from the same atomic elements.

V. DISCUSSION

A. IMPACT OF DOCUMENT CLEANSING

Similar to the study by Ayre et al. [53], our cleansing process resulted in substantial reductions in word and line counts, cf. Table 1. This suggests that redundant, unnecessary, or irrelevant content was removed, which was likely to enhance text clarity and conciseness. Despite variations in the extent of reduction among different narratives, there was a consistent trend toward a more concise and structured information presentation. However, in some cases, the word number increased, suggesting restructuring for clarity improvement. A typical case is the expansion of acronyms. Across all methods and for both terminology sources, a notable performance improvement occurred whenever text was cleansed by GPT-4. This suggests that this task successfully approximated the medical terms to more common, normalized terms corresponding to the terminology collections used, leading to improved normalization results.

B. EFFECTIVENESS OF DICTIONARY MATCHING

Dictionary matching, in isolation, achieved the lowest performance, especially when applied to raw text. This underscores the limitations of a simple matching approach, particularly with clinical texts known for their special jargon that is often not covered by standard terminologies.

C. BI-ENCODER METHOD PERFORMANCE

This method consistently outperformed the baseline across all metrics and for both terminology sources, underscoring its MCN effectiveness. Furthermore, the result demonstrates the significant advantage when applied to the (large) interface terminology (IT_DE) over the much smaller UMLS Metathesaurus extract (UMLS_DE). This finding strongly supports the critical importance of good terminology coverage in MCN.

D. IMPACT OF GPT-4 re-ranker

The GPT-4 re-ranker, in addition to the bi-encoder method, further enhanced performance, particularly on already cleansed text. As per the study by Ma et al. [63], LLMs as re-rankers in challenging information extraction tasks exhibited an F1 gain around 2.4%. This observation underscores the effective refinement of candidate concepts generated by the bi-encoder, leading to more precise normalization outcomes. To the best of our knowledge, LLMs as re-rankers have not been previously utilized for MCN tasks. Our investigation thus demonstrates a notable improvement, achieving a 6.87% F1 gain compared to the bi-encoder approach alone. A previous attempt to perform MCN directly by prompting GPT-4 for finding the right SNOMED code for a given expression had been immediately abandoned due to its propensity to the suggestion of completely hallucinated codes, indicating that GPT-4 lacks sufficient medical terminology knowledge. This highlights the necessity for a framework, such as bi-encoders, that first retrieves the correct concepts from the terminology and then improves accuracy by re-ranking.

E. THE IMPORTANCE OF TERMINOLOGIES

The result provides interesting insight when comparing UMLS_DE with IT_DE, two very different term collections with links to SNOMED CT codes. The former is of limited size (79,000 biomedical terms for 41,000 concepts) and only linked to SNOMED CT indirectly, whereas the latter is huge (2.5 million biomedical terms for 278,000 concepts) and provides good coverage of the numerous varieties of clinical jargon. This explains the almost doubling of the F-values across experiments. However, the high terminological coverage also explains that IT_DE benefitted comparatively less from document cleansing. It is, however, not surprising that document cleansing is the more beneficial, the scarcer is the coverage by existing terminologies. This is an important message for other languages, for which UMLS terminology extracts similar to our UMLS_DE could easily be obtained, but for which resources like IT_DE do not exist.

F. THE INTERPRETATION OF THE BEST-PERFORMING SCENARIO

The combination of IT_DE with text cleansing, bi-encoder and re-ranking yields an optimal F1@1 value of 0.607. This is all the more remarkable when compared to results of an earlier human text annotation exercise with SNOMED CT, where only low inter-annotator agreement values (Krippendorff's Alpha approx. 40%) had been achieved on a mix of clinical texts [70], and which had been interpreted as a result of SNOMED CT's huge size and of the unclear meaning of semantically related concepts.

G. PATIENT SAFETY

While challenges such as dealing with medical terminology, correctly resolving brand names, and preventing hallucinations remain, the proposed method faces risks such as loss or alteration of content at a relatively low level, given the black-box nature of LLMs. Their potential impact on patient safety must be assessed individually for each implementation scenario, which is, however, beyond the scope of this study. This means, particularly, the adaptation of the weighting between false positives and false negatives to the respective scenario. For instance, false positives may be more acceptable in cohort-building use cases for patient recruitment than in decision support scenarios, where a hallucinated drug or disease code can put the patient at risk. In contrast, when LLMs are used to improve content retrieval in medical records, a high recall is to be aimed at and false positives are more readily tolerated.

H. DATA PROTECTION

For our study, we chose GPT-4 was as the currently best-performing language model. However, the use of AI in healthcare requires strict compliance with data protection and security regulations, which makes GPT-4's proprietary nature a severe obstacle. Open source models such as Llama 3 should be preferred, but they have so far had significantly lower performance, especially for languages other than English. So it remains to wait for open source models with better performance. The deployment of commercial LLMs on trusted environments, such as Azure Cloud, could be an alternative. The extent to which this actually ensures the protection of highly sensitive patient data will be the subject of future discussions, as will all the trends that are currently observed in the extremely dynamic landscape of language models.

VI. CONCLUSION

Our study addressed medical concept normalization (MCN) of clinical narratives, i.e., the automated annotation of clinically relevant text passages with codes from the terminology SNOMED CT. The results support the transformative benefits of integrating LLMs such as ChatGPT into a concept normalization workflow. More precisely, ChatGPT significantly enhanced the performance of MCN, particularly by applying text cleansing and retrieval augmented generation (RAG) to clinical narratives in German language. Our investigation reveals a notable 6.87% increase in F1 score when using a bi-encoder with a re-ranker, compared to traditional bi-encoder approaches, underscoring ChatGPT's superiority in enhancing MCN tasks.

The benefits of document cleansing are especially pronounced in scenarios with limited terminological coverage. The impressive ability of ChatGPT to transform raw text into text with more standardized medical terms is an important message, particularly when using MCN for languages with limited support for terminological resources. Nevertheless, our experiments on German texts reveal the importance of a good terminological basis, even in times of LLMs, when comparing an extract of UMLS Metathesaurus linked to SNOMED codes with a German-specific interface terminology where F1 values roughly doubled.

Looking ahead, adopting LLMs holds promise for streamlining healthcare workflows, reducing documentation errors, and ultimately improving patient care outcomes. Future research should focus on refining these models, expanding their integration into diverse healthcare settings, and evaluating their impact in the real world on clinical practice.

Continuous attention must be paid to the settings (e.g., cloud platforms) in which commercial LLMs can be safely used for routine patient data. Likewise, the black-box nature of LLMs with the unpredictability of hallucinations requires careful monitoring and mitigation, e.g., by using additional resources for checking the plausibility of medical term changes proposed by the LLM.

Finally, clinical corpora in languages other than English, annotated with codes from standard terminologies at high granularity, are urgently needed as a source of additional ground truth data covering the full spectrum of clinical specialties. This would solve the problems of studies with a limited sample size, such as ours. Nevertheless, we can claim that this study has indicated directions for future research on larger corpora, such as those being created in the ongoing German GeMTeX project [71], which marks an important new step in resource creation for medical concept normalization.

ACKNOWLEDGMENT

This work was approved by the IRB of the Medical University of Graz (30-496 ex 17/18).

REFERENCES

[1] M. Sung, M. Jeong, Y. Choi, D. Kim, J. Lee, and J. Kang, "BERN2: An advanced neural biomedical named entity recognition and normalization tool," *Bioinformatics*, vol. 38, no. 20, pp. 4837–4839, Oct. 2022.

[2] A. Dash, S. Darshana, D. K. Yadav, and V. Gupta, "A clinical named entity recognition model using pretrained word embedding and deep neural networks," *Decis. Anal. J.*, vol. 10, Mar. 2024, Art. no. 100426.

[3] M. Y. Landolsi, L. Ben Romdhane, and L. Hlaoua, "Hybrid medical named entity recognition using document structure and surrounding context," *J. Supercomput.*, vol. 80, no. 4, pp. 5011–5041, Mar. 2024.

[4] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu, "Clinical information extraction applications: A literature review," *J. Biomed. Informat.*, vol. 77, pp. 34–49, Jan. 2018.

[5] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.

[6] B. Santana, R. Campos, E. Amorim, A. Jorge, P. Silvano, and S. Nunes, "A survey on narrative extraction from textual data," *Artif. Intell. Rev.*, vol. 56, no. 8, pp. 8393–8435, Aug. 2023.

[7] X. Yang, A. Chen, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, A. B. Costa, M. G. Flores, and Y. Zhang, "A large language model for electronic health records," *NPJ Digit. Med.*, vol. 5, no. 1, p. 194, 2022.

[8] C. Peng, X. Yang, A. Chen, K. E. Smith, N. PourNejatian, A. B. Costa, C. Martin, M. G. Flores, Y. Zhang, T. Magoc, G. Lipori, D. A. Mitchell, N. S. Ospina, M. M. Ahmed, W. R. Hogan, E. A. Shenkman, Y. Guo, J. Bian, and Y. Wu, "A study of generative large language model for medical research and healthcare," *npj Digit. Med.*, vol. 6, no. 1, p. 210, Nov. 2023.

[9] M. Javaid, A. Haleem, and R. P. Singh, "ChatGPT for healthcare services: An emerging stage for an innovative perspective," *BenchCouncil Trans. Benchmarks, Standards Evaluations*, vol. 3, no. 1, Feb. 2023, Art. no. 100105.

[10] T. Dave, S. A. Athaluri, and S. Singh, "ChatGPT in medicine: An overview of its applications, advantages, limitations, future prospects, and ethical considerations," *Frontiers Artif. Intell.*, vol. 6, May 2023, Art. no. 1169595.

[11] Á. García-Barragán, A. G. Calatayud, O. Solarte-Pabón, M. Provencio, E. Menasalvas, and V. Robles, "GPT for medical entity recognition in Spanish," *Multimedia Tools Appl.*, pp. 1–20, Apr. 2024.

[12] S. Schulz, P. Daumke, M. Romacker, and P. López-García, "Representing oncology in datasets: Standard or custom biomedical terminology?" *Informat. Med. Unlocked*, vol. 15, Jan. 2019, Art. no. 100186.

[13] H. Li, Q. Chen, B. Tang, X. Wang, H. Xu, B. Wang, and D. Huang, "CNN-based ranking for biomedical entity normalization," *BMC Bioinf.*, vol. 18, no. 11, pp. 79–86, Oct. 2017.

[14] Y. Luo, G. Song, P. Li, and Z. Qi, "Multi-task medical concept normalization using multi-view convolutional neural network," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–8.

[15] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, 2018, pp. 2227–2237.

[16] A. Radford. (2018). *Improving Language Understanding With Unsupervised Learning*. [Online]. Available: https://openai.com/research/language-unsupervised

[17] J. Devlin, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, vol. 1, 2019, pp. 1–16.

[18] Z. Ji, Q. Wei, and H. Xu, "BERT-based ranking for biomedical entity normalization," in *Proc. AMIA Summits Transl. Sci.*, 2020, p. 269.

[19] E. Ullah, A. Parwani, M. M. Baig, and R. Singh, "Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology—A recent scoping review," *Diagnostic Pathol.*, vol. 19, no. 1, p. 43, Feb. 2024.

[20] R. K. Sinha, A. D. Roy, N. Kumar, and H. Mondal, "Applicability of ChatGPT in assisting to solve higher order problems in pathology," *Cureus*, vol. 15, no. 2, Feb. 2023.

[21] A. Alsadhan, F. Al-Anezi, A. Almohanna, N. Alnaim, H. Alzahrani, R. Shinawi, H. Aboalsamh, A. Bakhshwain, M. Alenazy, W. Arif, S. Alyousef, S. Alhamidi, A. Alghamdi, N. AlShrayfi, N. B. Rubaian, T. Alanzi, A. AlSahli, R. Alturki, and N. Herzallah, "The opportunities and challenges of adopting ChatGPT in medical research," *Frontiers Med.*, vol. 10, Dec. 2023, Art. no. 1259640.

[22] SNOMED International. *SNOMED CT Starter Guide*. Accessed: Apr. 18, 2024. [Online]. Available: https://confluence.ihtsdotools.org/display/DOCSTART

[23] E. Chang and J. Mostafa, "The use of SNOMED CT, 2013–2020: A literature review," *J. Amer. Med. Inform. Assoc.*, vol. 28, no. 9, pp. 2017–2026, Aug. 2021.

[24] S. Schulz, W. Del-Pinto, L. Han, M. Kreuzthaler, S. Aghaei, and G. Nenadic, "Towards principles of ontology-based annotation of clinical narratives," in *Proc. ICBO*, 2023, pp. 1–12. [Online]. Available: https://ceur-ws.org/Vol-3603/

[25] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, and N. Collier, "Self-alignment pretraining for biomedical entity representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2021, pp. 4228–4238.

[26] A. Abdulnazar, M. Kreuzthaler, R. Roller, and S. Schulz, "SapBERT-based medical concept normalization using SNOMED CT," in *Caring is Sharing-Exploiting the Value in Data for Health and Innovation*. IOS Press, 2023.

[27] E. Waisberg, J. Ong, M. Masalkhi, S. A. Kamran, N. Zaman, P. Sarker, A. G. Lee, and A. Tavakkoli, "GPT-4: A new era of artificial intelligence in medicine," *Irish J. Med. Sci.*, vol. 192, no. 6, pp. 3197–3200, Dec. 2023.

[28] Y.-F. Luo, S. Henry, Y. Wang, F. Shen, O. Uzuner, and A. Rumshisky, "The 2019 n2c2/UMass lowell shared task on clinical concept normalization," *J. Amer. Med. Inform. Assoc.*, vol. 27, no. 10, p. 1529, Oct. 2020.

[29] D. Xu, M. Gopale, J. Zhang, K. Brown, E. Begoli, and S. Bethard, "Unified medical language system resources improve sieve-based generation and bidirectional encoder representations from transformers (BERT)-based ranking for concept normalization," *J. Amer. Med. Inform. Assoc.*, vol. 27, no. 10, pp. 1510–1519, Oct. 2020.

[30] L. Pape-Haugaard, "Clinical concept normalization on medical records using word embeddings and heuristics," *Stud. Health Technol. Inform.*, vol. 270, pp. 93–99, Jan. 2020.

[31] L. Chen, W. Fu, Y. Gu, Z. Sun, H. Li, E. Li, L. Jiang, Y. Gao, and Y. Huang, "Clinical concept normalization with a hybrid natural language processing system combining multilevel matching and machine learning ranking," *J. Amer. Med. Inform. Assoc.*, vol. 27, no. 10, pp. 1576–1584, Oct. 2020.

[32] K. S. Kalyan and S. Sangeetha, "Target concept guided medical concept normalization in noisy user-generated texts," in *Proc. Deep Learn. Inside Out (DeeLIO), 1st Workshop Knowl. Extraction Integr. Deep Learn. Architectures*, 2020, pp. 64–73.

[33] K. Lee, S. A. Hasan, O. Farri, A. Choudhary, and A. Agrawal, "Medical concept normalization for online user-generated texts," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, Aug. 2017, pp. 462–469.

[34] N. Pattisapu, S. Patil, G. Palshikar, and V. Varma, "Medical concept normalization by encoding target knowledge," in *Proc. Mach. Learn. Health Workshop*, 2020, pp. 246–259.

[35] Z. Miftahutdinov, A. Kadurin, R. Kudrin, and E. Tutubalina, "Medical concept normalization in clinical trials with drug and disease representation learning," *Bioinformatics*, vol. 37, no. 21, pp. 3856–3864, Nov. 2021.

[36] H. Cho, D. Choi, and H. Lee, "Re-ranking system with BERT for biomedical concept normalization," *IEEE Access*, vol. 9, pp. 121253–121262, 2021.

[37] P. Wajsbürt, A. Sarfati, and X. Tannier, "Medical concept normalization in French using multilingual terminologies and contextual embeddings," *J. Biomed. Informat.*, vol. 114, Feb. 2021, Art. no. 103684.

[38] M. Sung, H. Jeon, J. Lee, and J. Kang, "Biomedical entity representations with synonym marginalization," 2020, *arXiv:2005.00239*.

[39] D. Xu and T. Miller, "A simple neural vector space model for medical concept normalization using concept embeddings," *J. Biomed. Informat.*, vol. 130, Jun. 2022, Art. no. 104080.

[40] M. Schwarz, K. Chapman, and B. Häussler, "Multilingual medical entity recognition and cross-lingual zero-shot linking with Facebook AI similarity search," in *Proc. IberLEF@ SEPLN*, 2022, pp. 1–11.

[41] F. Borchert, I. Llorca, R. Roller, B. Arnrich, and M.-P. Schapranow, "XMEN: A modular toolkit for cross-lingual medical entity normalization," 2023, *arXiv:2310.11275*.

[42] D. Newman-Griffis, G. Divita, B. Desmet, A. Zirikly, C. P. Rosé, and E. Fosler-Lussier, "Ambiguity in medical concept normalization: An analysis of types and coverage in electronic health record datasets," *J. Amer. Med. Inform. Assoc.*, vol. 28, no. 3, pp. 516–532, Mar. 2021.

[43] D. Kartchner, J. Deng, S. Lohiya, T. Kopparthi, P. Bathala, D. Domingo-Fernández, and C. Mitchell, "A comprehensive evaluation of biomedical entity linking models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2023, pp. 14462–14478.

[44] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Med.*, vol. 29, no. 8, pp. 1930–1940, 2023.

[45] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, and D. Sontag, "Large language models are few-shot clinical information extractors," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 1–26.

[46] Y. Liu, S. Ju, and J. Wang, "Exploring the potential of ChatGPT in medical dialogue summarization: A study on consistency with human preferences," *BMC Med. Informat. Decis. Making*, vol. 24, no. 1, p. 75, Mar. 2024.

[47] Z. Zhou, "Evaluation of ChatGPT's capabilities in medical report generation," *Cureus*, vol. 15, no. 4, Apr. 2023, Art. no. e37589.

[48] N. Shrestha, Z. Shen, B. Zaidat, A. H. Duey, J. E. Tang, W. Ahmed, T. Hoang, M. R. Mejia, R. Rajjoub, J. S. Markowitz, J. S. Kim, and S. K. Cho, "Performance of ChatGPT on NASS clinical guidelines for the diagnosis and treatment of low back pain: A comparison study," *Spine*, vol. 49, no. 9, pp. 640–651, May 2024.

[49] R. K. Garg, V. L. Urs, A. A. Agrawal, S. K. Chaudhary, V. Paliwal, and S. K. Kar, "Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: A systematic review," *Health Promotion Perspect.*, vol. 13, no. 3, pp. 183–191, Sep. 2023.

[50] S. J. Jung, H. Kim, and K. S. Jang, "LLM based biological named entity recognition from scientific literature," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Feb. 2024, pp. 433–435.

[51] M. Sallam, "ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns," *Healthcare*, vol. 11, no. 6, p. 887, Mar. 2023.
[52] S. Biswas, "ChatGPT and the future of medical writing," *Radiology*, vol. 307, no. 2, Apr. 2023, Art. no. e223312.
[53] J. Ayre, O. Mac, K. McCaffery, B. R. McKay, M. Liu, Y. Shi, A. Rezwan, and A. G. Dunn, "New frontiers in health literacy: Using ChatGPT to simplify health information for people in the community," *J. Gen. Internal Med.*, vol. 39, no. 4, pp. 573–577, Mar. 2024.
[54] T. Brown, "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.
[55] S. R. Ali, T. D. Dobbs, H. A. Hutchings, and I. S. Whitaker, "Using ChatGPT to write patient clinic letters," *Lancet Digit. Health*, vol. 5, no. 4, pp. e179–e181, Apr. 2023.
[56] M. Cascella, J. Montomoli, V. Bellini, and E. Bignami, "Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios," *J. Med. Syst.*, vol. 47, no. 1, p. 33, Mar. 2023.
[57] S. B. Patel and K. Lam, "ChatGPT: The future of discharge summaries?" *Lancet Digit. Health*, vol. 5, no. 3, pp. 107–108, Mar. 2023.
[58] H. L. Walker, S. Ghani, C. Kuemmerli, C. A. Nebiker, B. P. Müller, D. A. Raptis, and S. M. Staubli, "Reliability of medical information provided by ChatGPT: Assessment against clinical guidelines and patient information quality instrument," *J. Med. Internet Res.*, vol. 25, Jun. 2023, Art. no. e47479.
[59] E. Y. Song, S. Kim, H. Lee, J. Kim, and J. Thorne, "Re3val: Reinforced and re-ranked generative retrieval," in *Proc. EACL*, 2024, pp. 393–409.
[60] Y. Guo, W. Qiu, G. Leroy, S. Wang, and T. Cohen, "Retrieval augmentation of large language models for lay language generation," *J. Biomed. Informat.*, vol. 149, Jan. 2024, Art. no. 104580.
[61] M. Eibich, S. Nagpal, and A. Fred-Ojala, "ARAGOG: Advanced RAG output grading," 2024, *arXiv:2404.01037*.
[62] O. Ram, Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown, and Y. Shoham, "In-context retrieval-augmented language models," *Trans. Assoc. Comput. Linguistics*, vol. 11, pp. 1316–1331, Nov. 2023.
[63] Y. Ma, Y. Cao, Y. Hong, and A. Sun, "Large language model is not a good few-shot information extractor, but a good reranker for hard samples!" in *Proc. Findings Assoc. Comput. Linguistics*, 2023, pp. 10572–10601.
[64] L. Modersohn, S. Schulz, C. Lohr, and U. Hahn, "GRASCCO—The first publicly shareable, multiply-alienated German clinical text corpus," *Stud. Health Technol. Inform.*, vol. 296, pp. 72–76, Jan. 2022.
[65] *Gemini for Google Workspace Prompting Guide*. Accessed: Jun. 26, 2024. [Online]. Available: https://services.google.com/fh/files/misc/gemini-for-google-workspace-prompting-guide-101.pdf
[66] O. Bodenreider, "The unified medical language system (UMLS): Integrating biomedical terminology," *Nucleic Acids Res.*, vol. 32, pp. 267–270, Jan. 2004.
[67] D. H. Nik, Z. Kasác, Z. Goda, A. Semlitsch, and S. Schulz, "Building an experimental German user interface terminology linked to SNOMED CT," in *MEDINFO 2019: Health and Wellbeing e-Networks for All*, 2019.
[68] A. Conneau, "Unsupervised cross-lingual representation learning at scale," 2019, *arXiv:1911.02116*.
[69] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, Jul. 2021.
[70] J. A. Miñarro-Giménez, R. Cornet, M. C. Jaulent, H. Dewenter, S. Thun, K. R. Gøeg, D. Karlsson, and S. Schulz, "Quantitative analysis of manual annotation of clinical text samples," *Int. J. Med. Informat.*, vol. 123, pp. 37–48, Mar. 2019.
[71] F. A. Meineke, L. Modersohn, M. Loeffler, and M. Boeker, "Announcement of the German medical text corpus project (GeMTeX)," *Stud. Health Technol. Inform.*, vol. 302, pp. 835–836, May 2023.

AKHILA ABDULNAZAR received the bachelor's degree in electronics and biomedical engineering from Cochin University of Science and Technology, India, and the master's degree in applied electronics and instrumentation from Kerala Technological University, India. She is currently pursuing the Ph.D. degree with the Medical University of Graz, Austria, specializing in standardizing clinical text data using computational semantics, NLP, and LLMs to enhance healthcare interoperability. Prior to her current roles, she was a Software Engineer with Siemens and a Research Intern with Robert Bosch, specializing in AI/ML applications. Her research interests include data visualization for complex biomedical data; empowering healthcare professionals through NLP, LLM, and image classifiers; and real-time signal classifiers.

ROLAND ROLLER received the degree in computer science and computational linguistics from the University of Trier and Saarland University, and the Ph.D. degree in information extraction from biomedical literature from the University of Sheffield, in 2015. He is a Senior Researcher and the Project Manager with the Speech and Language Technology Group, German Research Center for Artificial Intelligence (DFKI). Currently, he is working on topics related to information extraction, clinical decision support, anonymization, and chatbots.

STEFAN SCHULZ is a Full Professor of medical informatics with the Medical University of Graz. He trained as a Medical Doctor with a doctorate in theoretical medicine, he transitioned to medical informatics after two years of clinical practice. An expert in biomedical terminologies and ontologies, he contributes to standards development with SNOMED International. He also holds a part-time position with German NLP Company Averbis. His research group has been pivotal in several EU projects and has hosted international events. With over 250 peer-reviewed publications, he is a highly published researcher and has received multiple awards for his contributions to the field. His research interests include biomedical terminologies, ontologies, electronic health records, and medical language.

MARKUS KREUZTHALER is an Assistant Professor with the Institute for Medical Informatics, Statistics, and Documentation, Medical University of Graz, specializing in computational semantics for health. He focuses on extracting and representing relevant information from clinical texts using natural language processing, particularly machine learning methods. His current national and international research projects involve creating extended structured and standardized clinical patient profiles to support secondary use cases scenarios from clinical real-world data, utilizing international standards, such as SNOMED CT.