Natural language processing in healthcare: applications and value

Buchem, M.M. van

Chapter 1

Introduction

The healthcare sector is currently facing several critical challenges that threaten its accessibility and affordability. Among these challenges are personnel shortages, limited resources, and an ever-increasing demand for healthcare services. In response to these challenges, the healthcare sector has increasingly turned to technological innovations to find solutions. The most significant of these technological advancements has been the introduction of Electronic Health Records (EHRs). The shift from paper-based practices to EHRs in many hospitals aimed to streamline administrative processes and improve the efficiency of healthcare delivery. While the introduction of EHRs has transformed the landscape of healthcare data management, generating vast amounts of electronic, mostly unstructured healthcare data, it has not realized the anticipated reduction in administrative burdens[1]. This situation has paved the way for exploring further technological solutions to leverage the vast amount of data for enhancing healthcare services. Artificial Intelligence (AI), particularly through its subfield of Natural Language Processing (NLP), offers a promising avenue. In this thesis, applications and value of NLP are studied for healthcare.

1.1 Natural language processing

NLP combines linguistics and computer science, aiming to enable computers to process ('understand') natural language. NLP consists of many different subfields, with a wide variety of tasks and techniques. At the core of NLP are several key stages: data preprocessing, feature extraction, and modeling.

Data preprocessing involves several tasks aimed at enhancing the quality of the text and normalizing the text to improve the efficiency of subsequent processing steps. Common preprocessing techniques are shown in Figure 1. Which preprocessing techniques to apply depend on the NLP model that is used. For classical machine learning models such as logistic regression, support vector machines, and decision trees, more preprocessing steps are needed to reduce the dimensionality and sparsity of the data. More recent models are able to deal with the complexity of natural language and may only require tokenization.

Tokenization	Spelling correction	Stemming	Lemmatization	Stop word removal
Segmenting the text into smaller parts, such as letters, words, or sentences.	Corrects spelling mistakes	A text normalization technique that simplifies words to their root by removing prefixes, suffixes, and pluralizations.	Similar to stemming but uses vocabulary and morphological analysis of words to bring them back to their lemma.	Removes common words that often do not carry any meaning.
Example sentence				
I felt happier after I went running yesterdy.	'I', 'felt', 'happier', 'after', 'I', 'went', 'running', 'yesterdy', '.'	'I', 'felt', 'happier', 'after', 'I', 'went', 'running', 'yesterday', '.'	'I', 'felt', 'happier', 'after', 'I', 'went', 'running', 'yesterday', '.'	'I', 'feel', 'happy', 'after', 'I', 'go', 'run', 'yesterday', '.'
↓	↓	↓	↓	↓
'I', 'felt', 'happier', 'after', 'I', 'went', 'running', 'yesterdy', '.'	'I', 'felt', 'happier', 'after', 'I', 'went', 'running', 'yesterday', '.'	'I', 'fel', 'happ', 'after', 'I', 'went', 'run', 'yesterday', '.'	'I', 'feel', 'happy', 'after', 'I', 'go', 'run', 'yesterday', '.'	'I', 'feel', 'happy', 'I', 'run', 'yesterday', '.'

Figure 1: A visualization of common techniques for preprocessing in natural language processing[2].

Feature extraction is needed to transform the preprocessed data into a numerical format that can be analyzed by an algorithm. There are many different feature extraction techniques, with different levels of granularity (see Figure 2). Again, the model that is used establishes which feature extraction technique is suitable. As with preprocessing, this relates to the complexity the model can handle.

Bag-of-Words

BoW is one of the simplest techniques, where each document is represented as a bag of its words. This leads to a matrix, where every row is a document, and every column is a word. Every cell thus represents the frequency of occurrence of a specific word in a document.

Term frequency - inverse document frequency

TF-IDF builds on the BoW model by considering the entire corpus of documents to scale the frequency of words. Words that occur frequently in the entire corpus of documents will receive less weight, while words that occur seldom will receive more weight.

$$\text{TF-IDF} = \frac{\text{Number of times a word "X" appears in a Document}}{\text{Number of words present in a Document}} \times \log\left(\frac{\text{Number of Documents present in a Corpus}}{\text{Number of Documents where word "X" has appeared}}\right)$$

Word embeddings

Word embeddings are advanced feature extraction techniques that represent words in a dense vector space. Each word is mapped to a vector using unsupervised learning in such a way that the semantic relationship between words is reflected in the geometric relationship between vectors. Words that are used in similar contexts will thus be close together in this vector space. Common models that generate embeddings include Word2Vec, GloVe, and FastText. [refs].

Contextualized word embeddings

The newer generation language models use contextual word embeddings, where a word can have several different embeddings based on the context. An example is the word 'bank', which can refer to a financial bank or a river bank. These two meanings of the word 'bank' will each have their own embedding. Common models that generate contextual word embeddings include ELmO, BERT, and GPT.

Example of output

Text	the	dog	barked	against
The dog barked loudly	1	1	1	0
The cat rubbed against	1	0	0	1
The dog barked against	1	1	1	1
A cat caught the mouse	0	0	0	0

Table showing the result of applying the bag-of-words technique to four example texts.

Text	the	dog	barked	against
The dog barked loudly	0.1	0.8	0.6	0
The cat rubbed against	0.1	0	0	0.5
The dog barked against	0.1	0.7	0.5	0.3
A cat caught the mouse	0	0	0	0

Table showing an example of the result of applying the TF-IDF formula to four texts.

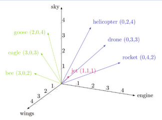

Figure showing a simplified example of an embedding space for words.

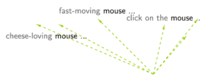

Figure showing a simplified example of a contextualized embedding space for words.

Figure 2: A visualization of common techniques for feature extraction in natural language processing, ordered by increasing complexity.

Modeling applies algorithms to the structured data that arises from the previous steps. There are many different modeling tasks and associated models, the ones presented here are those that emerged as most relevant for the settings discussed in the following chapters (see Figure 3).

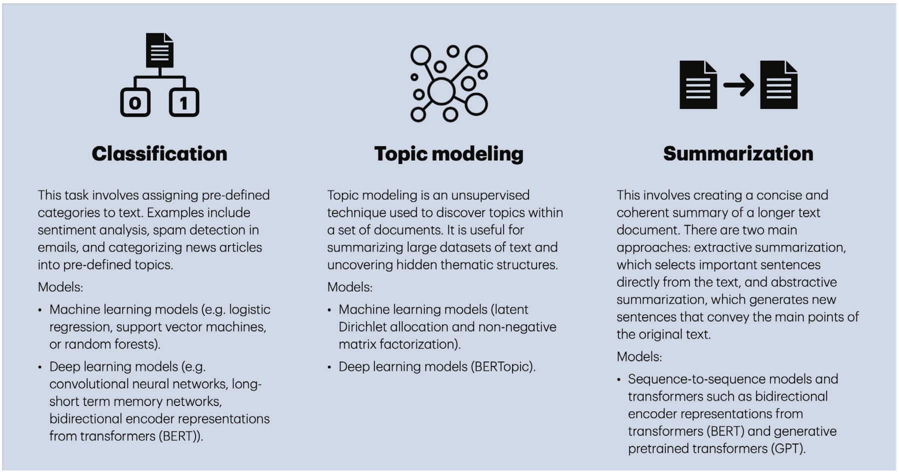

Figure 3: A visualization of the three natural language processing tasks included in this thesis.

Recent advances

In the past decade, large steps have been made in the advancement of the NLP field. The steady increase in computing power has made it possible to build language models with an increasing number of parameters. ELMo (Embeddings from Language Models), introduced in 2018, had approximately 90 million parameters[3]. Its successor, BERT (Bidirectional Encoder Representations from Transformers), introduced in 2018, has a base model with 110 million and a large model with 340 million parameters[4]. This was the first model to use a transformer architecture, a neural network architecture designed to be efficient at capturing relationships and dependencies between elements in a sequence, such as words in a sentence. This new architecture led to huge jumps in performance. The newest language models, such as OpenAI's GPT-3.5 (Generative Pretrained Transformers) and Google's PaLM (Pathways Language Model), also use the transformer architecture and have over 100 billion parameters, coining the new term *large language models* (see Figure 4). With these new possibilities, a key question is how these techniques can be applied to healthcare data and if they can create value for clinical practice.

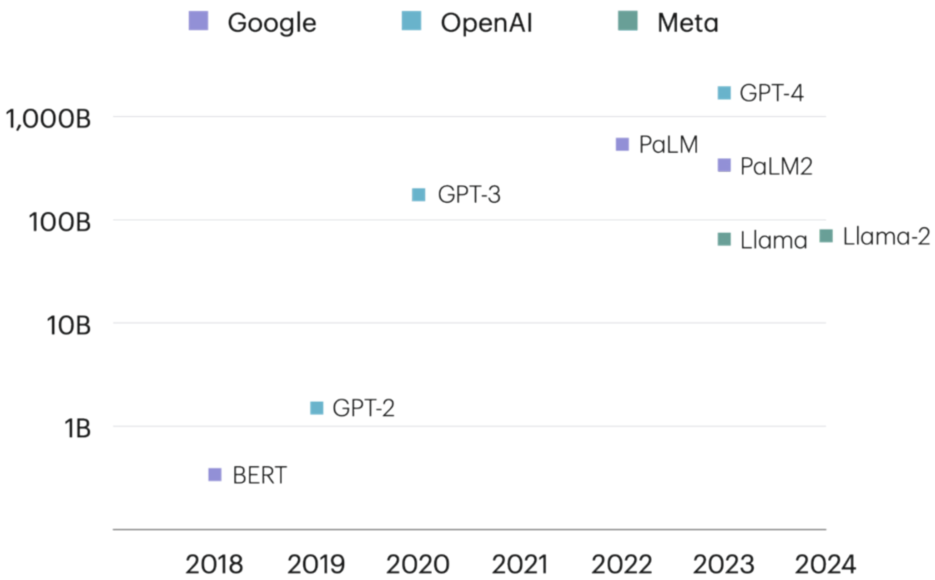

Figure 4: A timeline showing the (large) language models that have been introduced over the past five years and the number of parameters, plotted on a logarithmic scale.

1.2 Healthcare data

To create value at point-of-care, it is essential to use data routinely collected throughout the healthcare process. Routinely collected healthcare data for the purpose of NLP encompasses a broad spectrum of text-based information, including but not limited to:

- **Clinical data**, such as clinical notes, operative reports, and discharge summaries captured within the electronic health record (EHR). This data can be characterized as information from and to healthcare providers, summarizing interactions with patients, letters from and to other healthcare professionals, and notes from nurses, physicians, and other medical professionals.

- **Patient-generated data**, such as patient portal messages and patient experience surveys, characterized as information from patients to the hospital. With the increasing focus on patient-centered care, different efforts have been made to improve capturing the patient's voice and improving patient communication. This data is unique as it provides insights into patients' lives outside of the hospital.
- **Social media data**, from websites such as Reddit, Twitter, or online fora. The use of social media has increased over the past two decades, giving patients the opportunity to share their experiences, knowledge, and questions online. This data is characterized as peer-to-peer.

Previous studies show that free-text clinical data include valuable information not captured in structured data fields. Especially information about the complexity, evolving circumstances, uncertainty, and severity are often captured in free-text fields[5]. Furthermore, free-text clinical notes have previously been found to be more accurate, more reliable in identifying patients with certain diseases, and more understandable to review for other healthcare providers[6]. Similarly, for sources such as patient experience data, healthcare providers prefer answers to open-ended questions over closed-ended questions, as they provide more nuance[7]. All these examples clearly highlight the value of free-text healthcare data. With the volume of healthcare data estimated to grow with an annual rate of 36%[8], of which approximately 80% is unstructured, the need for the application of NLP is high.

1.3 Research questions

Despite the growing body of research on NLP in healthcare, the value of NLP for clinical practice is still unclear and implementation is lagging[9]. This thesis aims to clarify the value of the opportunities presented by NLP tools in clinical practice. It explores various applications of NLP within the healthcare setting and evaluates their value and offers insights into how NLP technologies can be effectively integrated into daily healthcare delivery to contribute to the improvement of healthcare accessibility and affordability. The following research questions will be addressed:

1. **Promising applications:** Which combination of NLP methods and data sources are most promising for enhancing healthcare delivery at the point of care?
2. **Challenges during development:** What are the principal challenges during the development of NLP models that hinder valuable adoption in clinical practice?
3. **Value for clinical practice**: What is the value of NLP applications for daily clinical practice?

1.4 Outline

The research questions will be addressed in the following chapters, which are divided into two parts. Part 1 focuses on the application of NLP in five different healthcare settings. Chapter 2 presents a scoping review about digital scribes in clinical practice. Chapter 3 through 6 describe the development of NLP models in different settings. Part 2 centers around the added value of NLP in clinical practice. Chapter 7 and 8 present pilot studies of applications presented in Part 1. See Table 1 for an overview of the clinical domains, data types and data sources. Figure 5 visualizes the outline and the relations between the chapters.

Table 1: overview of the chapters and their clinical domain, data type, and data source.

Chapter	Clinical domain	Data type	Data source
Part 1: Application of NLP in various healthcare settings			
2	General	Clinical	Scoping review
3	Oncology	Clinical	Electronic health record (Stanford)
4	Otorhinolaryngology	Patient-generated	Patient experience questionnaires (LUMC)
5	Oncology	Patient-generated	Patient portal messages, electronic health record (Stanford)
6	Cardiology	Social media	Online forum (Reddit)
Part 2: Evaluating the added value in clinical practice			
7	Otorhinolaryngology	Patient-generated	Patient experience questionnaires (LUMC)
8	Internal medicine	Clinical	Recorded mock conversations

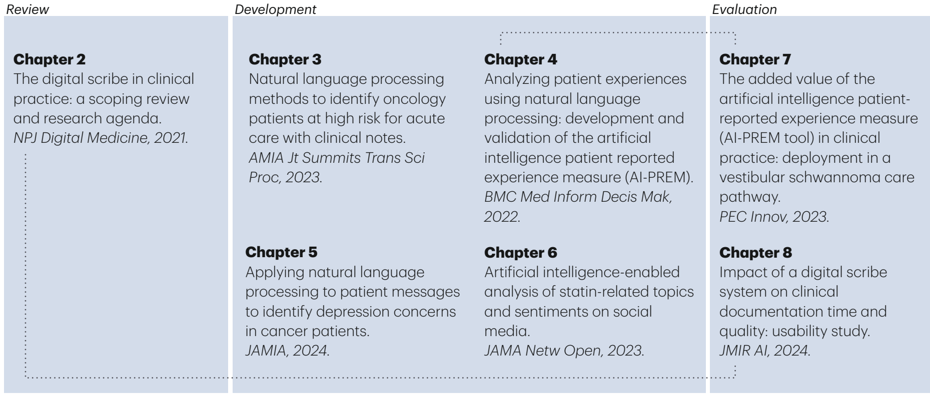

Figure 5: visualization of the outline of this thesis. Chapters that address the same setting are connected by a dotted line.

1.5 Terminology

Although NLP and the broader field of machine learning share many methodologies with biostatistics and epidemiology, different terms for similar concepts are used. In Table 2, key terminology used in this thesis is defined.

Table 2: definition of key terminology.

Term	Meaning
Model	A set of parameters and structure needed for a system to turn input data into output. Examples: logistic regression, support vector machines, neural network.
Label	The assigned category or value that an algorithm aims to predict, often used in supervised learning for training purposes.
Features	The individual measurable properties or characteristics of the data used by a model to make predictions. Similar to risk factors in epidemiology and independent variables in statistics.
Model training	Learning the association between features and labels. Similar to model fitting in statistics.

References

1. Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan W-J, Sinsky CA, Gilchrist VJ. Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. Ann Fam Medicine 2017;15(5):419–426.
2. Vajjala S, Majumder B, Gupta A, Surana H. Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems. O'Reilly Media, Inc.; 2020.
3. Peters, M. E. et al. Deep contextualized word representations. In Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL) 1, 2227–2237 (Association for Computational Linguistics, 2018).
4. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), 1, 4171–4186 (Association for Computational Linguistics, 2019).
5. Ford E, Oswald M, Hassan L, Bozentko K, Nenadic G, Cassell J. Should free-text data in electronic medical records be shared for research? A citizens' jury study in the UK. J Méd Ethics 2020;46(6):367–377.
6. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Méd Inform Assoc 2011;18(2):181–186.
7. Riiskjaer E, Ammentorp J, Kofoed P-E. The value of open-ended questions in surveys on patient experience: number of comments and perceived usefulness from a hospital perspective. Int J Qual Health C 2012;24(5):509–516.
8. Reinsel D, Gantz J, Rydning J. The Digitization of the World: From Edge to Core. IDC; 2018.
9. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019;25(1):30–36.