Einsatz von maschinellem Lernen zur Erkennung von SARS-CoV-2-bezogenen Dokumenten am Luzerner Kantonsspital

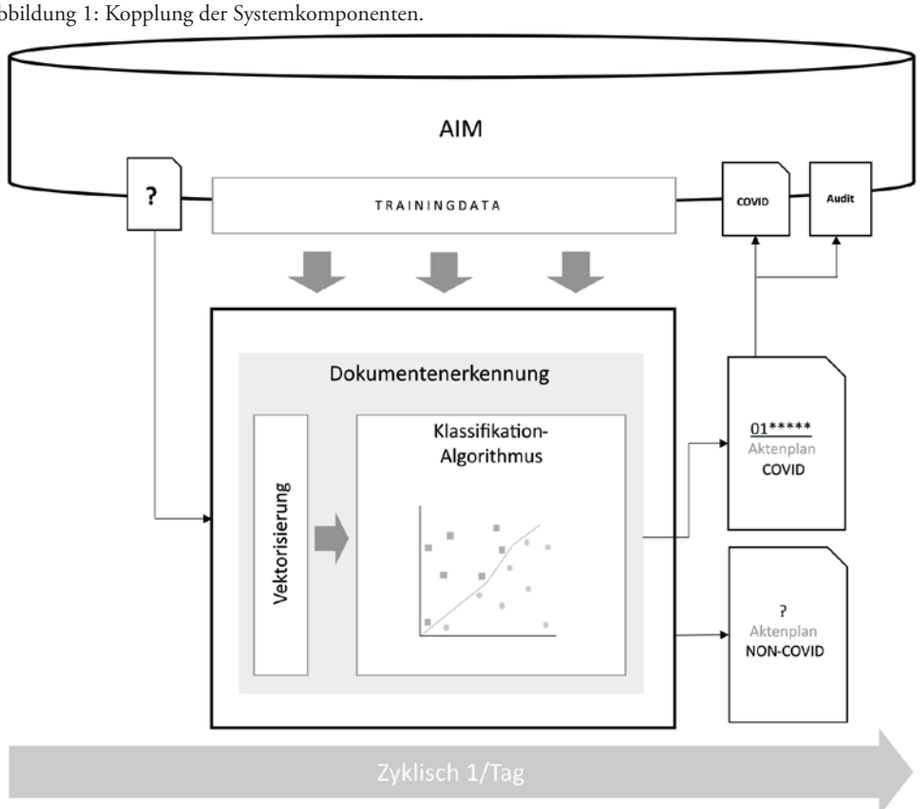
Entscheidende Informationen schnell und zielgerecht finden

Ein vom Medizin-IT-Unternehmen synedra IT GmbH, Innsbruck, bereitgestellter Algorithmus unterstützt die Betreuenden des elektronischen Universalarchives, indem es den Inhalt der ins Archiv einströmenden Dokumente analysiert und den Bezug zu SARS-CoV-2 Erkrankungen identifiziert.

COVID-19 wurde erstmals am 25. Februar 2020 in der Schweiz bestätigt. Seitdem hat sich das Virus in der Schweiz ausgebreitet und es wurden zahlreiche Massnahmen ergriffen, um seine Ausbreitung zu kontrollieren. In der Schweiz wurden mehrere COVID-19-Instrumente eingeführt, um die Verbreitung des Virus einzudämmen und die Bevölkerung zu schützen. Nachfolgend sind die wichtigsten Dokumente generisch aufgelistet, die jedoch in der Praxis in ganz unterschiedlicher Form und Ausprägung in Erscheinung traten.

- COVID-19-**Impfausweis**: Dieses Dokument bestätigt, dass eine Person gegen COVID-19 geimpft wurde. Es dient als Nachweis für den Impfstatus und kann für den Zugang zu Veranstaltungen oder Reisen erforderlich sein.
- COVID-19-**Testzertifikat**: Es kann als Nachweis für den Teststatus verwendet werden und in einigen Fällen für den Zugang zu Veranstaltungen oder Reisen erforderlich sein.
- COVID-19-**Zertifikat** für genesene Personen: Dieses Dokument bestätigt, dass eine Person COVID-19 überstanden und daher vermutlich eine Immunität gegen das Virus hat. Dieser Nachweis kann für den Genesungsstatus verwendet werden und in einigen Fällen für den Zugang zu Veranstaltungen oder Reisen erforderlich sein.

Abbildung 1: Kopplung der Systemkomponenten.



Angesichts der grossen Mengen an Patientenakten und klinischen Dokumenten ist es für das medizinische Fachpersonal oftmals eine Herausforderung, entscheidende Informationen schnell und zielgerecht aufzufinden. Eine hohe Datenqualität ist das Fundament einer jeden Informationssuche. Infolge des rasanten Anstiegs an diversen Nachweisen und Dokumenten im Zusammenhang mit SARS-CoV-2 waren auch die Betreuenden des elektronischen Universalarchives (eArchivs) am Luzerner Kantonsspital (LUKS) gefordert, die Datenqualität im eArchiv aufrechtzuerhalten.

Das LUKS setzt maschinelles Lernen ein, um die Klassifizierung von patientenorientierten Dokumenten zu verbessern und Informationen leichter zugänglich zu machen.

Hintergrund

Im Juni 2021 startete das LUKS in Zusammenarbeit mit seiner Archivherstellerin synedra IT GmbH ein gemeinsames Proof-of-Concept (PoC) Projekt, um SARS-CoV-2-bezogene Patienten-Dokumente zu erkennen und folglich zu klassifizieren. Dafür wurde ein automatisierter Prozess entwickelt, der zum einen die Datenqualität des eArchivs sicherstellt und zum anderen die Betreuenden entlasten soll. Das Projekt gewährte einen Einblick in die Anwendung von neuen technischen Ansätzen sowie deren Eignung für den Einsatz im Betrieb einer Klinik.

Realisierung mit Natural Language Processing und überwachtem Lernen

Zur Realisierung des Vorhabens wurden Methoden des Natural Language Processing (NLP) mit Methoden des überwachten Lernens kombiniert. NLP befasst sich mit der Interaktion zwischen Computern und der menschlichen Sprache und ermöglicht es, diese digital zu erfassen und zu




Sie machen mehr aus der wachsenden Menge von Patientendaten. Gemeinsam entwickelten (v.l.n.r.) Clemens Hörtenhuemer, M.Sc, und Valentin Muhr, MSc MA, synedra information technologies GmbH, mit Dr. med. Guido Schüpfer, PhD, MBA HSG, CMO, und Philipp Wessner, lic. rer. publ. HSG, EMBE HSG, Leiter IT-Dienste & Information Governance LUKS Gruppe, eine umfassende Dokumentenanalyse mit Bezug zu SARS-CoV-2 Erkrankungen.

interpretieren. Beim überwachten Lernen wird ein Modell auf Basis von Referenzdaten trainiert, um Vorhersagen zu treffen, wobei beim Training die Abweichung zwischen vorhergesagtem und tatsächlichem Wert minimiert werden soll.

Sowohl NLP als auch überwachtes Lernen sind Teilbereiche des maschinellen Lernens

Die Referenzdaten wurden vom LUKS gezielt manuell selektiert und als Trainingsdaten für das Modell bereitgestellt. Die Trainingsdaten beinhalten sowohl SARS-CoV-2-relevante Dokumente als auch Dokumente ohne Bezug zu SARS-CoV-2. Charakteristika für beide Klassen sind wichtig, um sie bestmöglich unterscheiden zu können.

Technische Realisierung (Ablauf)

Innerhalb der Patienten-Dokumente werden Formulierungsmuster in den Texten gesucht, um eine möglichst zutreffende Aussage über die «Corona-Relevanz» abzuleiten. Dies bewirkt gegebenenfalls Korrekturen der Aktenplanpositionen und verbessert somit die Datenqualität im eArchiv.

Im Inneren des Modells wird der Textinhalt der Dokumente gemäss dem folgenden Ablauf verarbeitet:

1. Texterkennung (Optical Character Recognition)

Beim Hauptteil der ins Archiv einströmenden Dokumente handelt es sich um PDF-Dokumente. Bei vielen dieser Dokumente muss der maschinell verarbeitbare Text erst mittels Optical Character Recognition (OCR) extrahiert werden. Visuelle Strukturen der Dokumente werden hierbei bewusst ausgeschlossen. Die weitere Verarbeitung stützt sich ausschliesslich auf den Textinhalt.

2. Vorverarbeitung (Preprocessing)

Für die Klassifikation sind nicht alle Wörter der Texte gleich relevant. Beispielsweise ist es leicht vorstellbar, dass Wörter wie «Antigentest» oder «Impfpass» deutlich ausschlaggebender sind als die Wörter «und» oder «daher». Letztere zählen zu den Stoppwörtern, die sehr häufig auftreten und gewöhnlich keine Relevanz für die Erfassung des Dokumenteninhalts haben[1]. Für die Klassifikation kann auf Stoppwörter verzichtet werden, folglich werden Stoppwörter aus dem Text entfernt.

Ähnlich verhält es sich mit Konjugationen. Diese stellen eine zusätzliche Varianz dar, die allerdings keine zusätzliche fachliche Unterscheidbarkeit mitliefert. Um diese Varianz zu dämpfen,

1 Daniel Koch: Suchmaschinen-Optimierung: Website-Marketing für Entwickler. Pearson Deutschland, 2007, ISBN 978-3-8273-2469-6, S. 35.

Dem Luzerner Kantonsspital kommt eine avantgardistische Rolle im systematischen Nutzen von Daten zu, was das aktuelle Projekt unterstreicht.

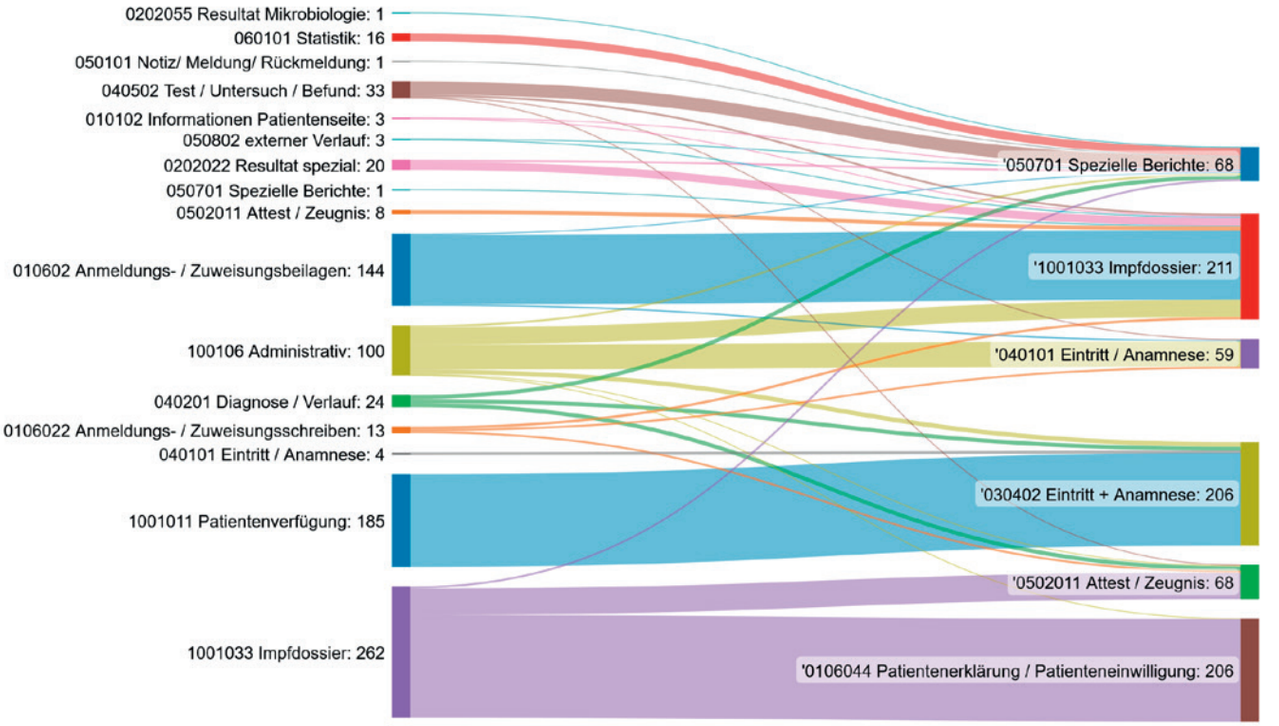


Abbildung 2: Graphik der vorgenommenen Korrekturen durch den NLP-Algorithmus dar. Die ursprünglichen Aktenplanpositionen (links) stehen den neuen Ziel-Aktenplanpositionen (rechts) gegenüber.

werden Verben in ihren Wortstamm zurückgeführt, sogenanntes «Stemming». Konkret wurde dafür der von Martin Porter entwickelte Snowball-Algorithmus[2] angewandt.

3. Vektorisierung (Feature Extraction)

Damit die Texte numerisch analysiert werden können, müssen sie in eine maschinell verarbeitbare Repräsentation überführt werden. Dabei wird der Text auf ein breites Spektrum an Merkmalen hin geprüft. Dazu gehören z. B. Wortkombinationen, sogenannte N-Gramme, die sich aus der Permutation aller vorkommenden Wörter in den Trainingsdaten ergeben. Neben dem binären Auftreten des Merkmals wird auch die Häufigkeit innerhalb des Textes sowie die Signifikanz des Merkmals in der Gesamtheit der Trainingsdaten bestimmt. Anhand dieser Berechnungen wird die Term Frequency Inverse Document Frequency Masszahl (tf-idf) für jedes Merkmal bestimmt.

$$tf - idf_{(t,d,D)} = \frac{count(t,d)}{max\{f_{d,t'} : t' \in d\}} \cdot log\frac{|D|}{\sum d \in D : t \in d}$$

d … Text, D … Menge an Texten in den Trainingsdaten, t … bestimmtes Merkmal

Eine Liste an tf-idf Masszahlen, ihrer Reihung nach korrespondierend mit dem zugehörigen Merkmal, stellt das finale Vektorisierung-Ergebnis des Textes dar.

2 M.F. Porter: Snowball: A language for stemming algorithms. In: Program, 2001

Die Vektorisierung-Komponente des Modells muss vor der operativen Nutzung alle möglichen Merkmale lernen und deren Signifikanz feststellen. Dazu ist eine initiale Selbstkonfiguration anhand der Trainingsdaten notwendig.

4. Klassifikation

Das Herzstück der Verarbeitung ist der Textklassifikationsalgorithmus. Er definiert sich als surjektive Funktion f : X → Y, die im Definitionsbereich (X) vektorisierte Texte annimmt und dazu eine Klasse aus der Zielmenge (Y) zurückliefert.

Es gibt diverse Klassifikationsalgorithmen mit unterschiedlichen Eignungen je nach Einsatzzweck. Beispielsweise werden Künstliche Neuronale Netze gerne für offene «Open Domain» Probleme eingesetzt, also nicht-fachbereichsbeschränkte Einsatzzwecke. Wir haben die Lineare Support Vektor Maschine (SVM) gewählt, die sich für Probleme mit engem fachlichem Korridor etabliert hat und sich durch besondere Laufzeiteffizienz charakterisiert. Am LUKS werden täglich ca. 29 000 Dokumente archiviert, weshalb das Modell nicht nur eine hohe Präzision, sondern auch eine gewisse Durchsatzkapazität erreichen muss, um dem klinischen Alltag gerecht zu werden.

Gleich wie bei der Vektorisierung muss auch bei der Klassifikation der Algorithmus vor der operativen Nutzung initial angelernt werden, um die Funktion f zu erfassen. Ähnlich wie ein Spürhund, der eine Referenzduftnote benötigt, um eine Fährte zu verfolgen, benötigt ein Klassifikations-Algorithmus Beispiele (Trainingsdaten) mit zugeordneten Klassen, also die Werte der Zielmenge Y.

5. Nachverarbeitung (fachliche Verarbeitung)

Der Text eines Dokumentes wird zunächst inhaltsbezogen klassifiziert und ergibt die Aktenplanposition. Nachfolgend wird ein Vergleich zwischen der maschinell festgestellten Aktenplanposition und der bestehenden Aktenplanposition vorgenommen, im Falle einer Abweichung wird diese korrigiert. Zur sicheren Nachvollziehbarkeit wird die Korrektur in einem dokumentbezogenen Audit vermerkt. Eine Korrektur erwirkt das Festhalten von Metadaten, wodurch nachfolgende Analysen ermöglicht werden.

Zum Einsatz kommt das Verfahren bei jeglichen neuen Archivdokumenten. Aber auch Bestandsdokumente konnten damit nachträglich analysiert und gemäss der Dienlichkeit neuerlich begutachtet werden.

Stolpersteine, die es zu meistern galt

Im Zuge der Umsetzung wurden die Projekt-Verantwortlichen mit unerwarteten Herausforderungen konfrontiert:

Im Regelbetrieb des LUKS tauchten Dokumente mit «Corona-Relevanz» gepaart mit Dokumenten

auf, die keinen Bezug zu SARS-CoV-2 aufwiesen. Diese Vermengung von Dokumenten unterschiedlicher fachlicher Herkunft und Domänenorientierung ergibt eine Unschärfe, die im Analyseprozess berücksichtigt werden muss.

Eine Adaption der Methodik von dokumentenbasierter zu seitenbasierter Analyse ermöglichte es, jene gemischten Dokumente zu detektieren. Entspricht die Klasse einer Seite innerhalb eines Dokuments nicht den übrigen Seiten bzw. ist der SARS-CoV-2-Bezug bereits für eine Seite gegeben, wird das Dokument als Konglomerat erfasst. Die Betreuenden des eArchivs prüfen betreffende Dokumente und gehen gezielt Einzelfällen nach, um eventuelle Fehlerquellen in der Erzeugungskette der Dokumente zu identifizieren und diese bestenfalls zu beheben.

Erwähnenswert ist das damit einhergehende Risiko. Erfolgt die Analyse auf der Basis von Einzelseiten, so reduziert sich die Informationsgrundlage, auf die sich die Klassifikation stützt. Die Dichte an charakterisierenden Merkmalen wird verringert und die Robustheit des Verfahrens beeinträchtigt. Die Berücksichtigung von Konglomeraten hat sich dennoch als unerlässlich erwiesen. Durch die Optimierung der Trainingsparameter konnte allerdings auch nach der Umstellung eine äquivalente Modellgüte erzielt werden.

Erfreuliche Ergebnisse

Die erzielten Ergebnisse lassen sich wie folgt zusammenfassen:
- 23 Monate von Konzeption zu Realisierung
- 2500 Dokumente im Training verwendet
- 3 Mio. Dokumente bereits analysiert
- Durchsatz: 4000 Dokumente/Stunde
- 3191 SARS-CoV-2-bezogene Dokumente identifiziert
- 805 Dokumente automatisiert umstrukturiert

Das Verfahren ist seit 12.12.2022 im Einsatz und hat inzwischen über 3 Millionen Dokumente inhaltlich analysiert, 3191 Dokumente mit «Corona-Relevanz» identifiziert und davon 805 automatisiert umstrukturiert. Seit der Realisierung können Verbesserungen bei der Verwaltung von Dokumenten festgestellt werden. Das schnelle Auffinden relevanter Informationen wird erleichtert, wodurch das medizinische Fachpersonal wertvolle Zeit spart.

Autoren

- Clemens Hörtenhuemer, M.Sc, synedra information technologies GmbH
- Valentin Muhr, MSc MA, synedra information technologies GmbH
- Guido Schüpfer, Dr. med., PhD, MBA HSG, CMO LUKS Gruppe
- Philipp Wessner, lic.rer.publ.HSG, EMBE HSG, Leiter IT-Dienste & Information Governance LUKS Gruppe