Klassifikation von Dokumenten im medizinischen Umfeld

Marcel Canclini

MASTERARBEIT

eingereicht an der
Zürcher Hochschule für angewandte Wissenschaften

MAS Data Science

in Winterthur

Betreuer: Dr. Mark Cieliebak

Januar 2018

Kurzfassung

Im Gegensatz zu anderen Branchen in der Schweiz ist die Digitalisierung im Gesundheitswesen wenig fortgeschritten. Längst wurde zwar das Potenzial erkannt und wurden Bestrebungen in diese Richtung unternommen, die Vielzahl verschiedener Akteure erschwert jedoch ein rasches Vorankommen in diesem Vorhaben. Unterschiedliche Prioritäten und Ressourcen führten dazu, dass die digitale Transformation bei grösseren Gesundheitsinstitutionen bereits deutlich weiter fortgeschritten ist als in den Praxen der niedergelassenen Ärzte. Aktuelle Bemühungen laufen diesbezüglich meist darauf hinaus, dass ein zusätzlicher Aufwand für die niedergelassenen Ärzte entsteht.

Die Digitalisierung lässt sich durch moderne Technologien wie Machine Learning massgeblich vorantreiben und die eHealth Strategie der Schweiz unterstützen. Durch solche intelligenten Systeme lassen sich manuelle Aufwände reduzieren. Als Resultat dieser Arbeit werden daher medizinische Dokumente wie Befunde, Überweisungen oder Rezepte durch Textklassifikation automatisch einem Geschäftsprozess zugeordnet. Das bietet die Basis für eine weitere Strukturierung der Daten.

Die grundlegenden Daten liefern Dokumente aus der gerichteten Kommunikation zwischen verschiedenen Gesundheitsorganisationen, welche manuell klassifiziert wurden. Durch ein Topic Modelling und entsprechende Visualisierung wird aufgezeigt, dass medizinische Dokumente ein ziemlich eindeutiges Vokabular besitzen. Es können sehr klar getrennte Gruppen gebildet werden, welche sich mit den vordefinierten Geschäftsprozessen decken. Mit diesem Wissen werden Modelle zur automatischen Klassifikation der Dokumente gesucht. Auf extrahierten Features werden unterschiedliche Klassifikatoren trainiert und miteinander auf definierten Metriken verglichen. Jene zwei Klassifikatoren mit den vielversprechendsten Resultaten werden weiter auf den Daten optimiert. Dieses Vorgehen führt zu einem Modell, welches eine erstaunlich hohe Präzision in der automatischen Zuordnung der Dokumentenklassen erreicht.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

1.1 Digitale Transformation im Gesundheitswesen der Schweiz

Dank modernster Forschung sind die technologischen Fortschritte in der Medizin enorm gross. Richtet man den Blick weg von der Medizin in die administrative Realität einer Arztpraxis, erkennt man schnell, dass die Digitalisierung im Schweizerischen Gesundheitswesen noch längst nicht überall Einzug gehalten hat. Gemäss dem eHealth Barometer 2017 [4, Grafik 30] führen erst 49% der Praxisärzte/PraxisärztInnen die Krankengeschichte ihrer Patienten vollständig elektronisch. Fax ist vielerorts das Kommunikationsmittel der Wahl und gemäss der SISA Studie 2015 [3] ist man von der papierlosen Praxis weit entfernt.

Bestrebungen zur Digitalisierung der Branche laufen. In den vergangenen Jahren wurden diverse eHealth Umsetzungsprojekte lanciert. Unklarheit bezüglich Standards und die Vielzahl verschiedener Akteure mit unterschiedlichen Interessen erschweren aber meist eine erfolgreiche Umsetzung. Auszug aus dem eHealth Barometer [4]:

> Während die Spitäler gegenüber eHealth-Bestrebungen in vielerlei Hinsicht Speerspitze und Motor zugleich sind, steht die Praxisärzteschaft eher auf der Bremse. Dies ist auch die Gruppe Gesundheitsfachpersonen, welche das geringste Potential durch eHealth sieht. Dabei ist aber gerade die Ärzteschaft – im Spital und in den Praxen – essenziell für den Erfolg von eHealth in der Schweiz.

So kommt es, dass grössere Gesundheitsorganisationen, welche über mehr Ressourcen verfügen, bereits deutlich weiter sind in der digitalen Transformation, als die niedergelassenen Ärzte.

1.2 Problemstellung

Spitäler und besonders radiologische Institute sind bei der Digitalisierung ihrer Prozesse den Arztpraxen einen Schritt voraus. Sie wünschen bei Zuweisungen von Patienten bereits eine gewisse Strukturierung der medizinischen und administrativen Informationen. Der Aufwand fällt beim Arzt als Auftraggeber an, der Nutzen jedoch liegt beim Empfänger. Die Motivation für den Arzt, eine manuelle Strukturierung vorzunehmen, ist somit gering.

Betrachten wir dazu den Geschäftsprozess *Radiologische Anmeldung* genauer: Radiologische Institute wünschen bei Überweisungen von niedergelassenen Ärzten, dass ein spezifisches (Web)Formular ausgefüllt wird[1]. Denn durch die Vielzahl von Überweisungen (Anmeldungen) wird täglich ein enormer administrativer Aufwand generiert. Wären die Informationen der Anmeldung (Patientendaten, Modalität, Fragestellung, etc.) strukturiert vorhanden, könnte der interne Prozess für das Radiologische Institut deutlich effizienter gestaltet werden. Beispielsweise wäre eine automatische Terminvergabe möglich, wenn die Modalität (MRI, Röntgen, CT, etc.) bei der Anmeldung bekannt wäre. Aus Sicht des Empfängers ist es somit nachvollziehbar, dass strukturierte Daten verlangt werden. Für den Absender jedoch ist das Ausfüllen eines empfängerspezifischen Formulars umständlich, da Informationen jeweils manuell aus seiner Praxissoftware in das entsprechende Formular übernommen werden müssen. Für den Arzt ist ein allgemeines Überweisungsschreiben mit entsprechenden Textbausteinen, welche automatisch aus seiner Praxissoftware generiert werden, deutlich einfacher und schneller. Trotzdem könnte auch der Auftraggeber von einer Strukturierung profitieren. Aufgrund einer direkten Terminvergabe wäre es möglich, unmittelbar eine Folgekonsultation mit dem Patienten zu vereinbaren und den koordinativen Aufwand zu verringern. Dazu genügen in der Praxissoftware strukturiert vorhandene Daten nicht. Die Software muss auch die Formulare der verschiedenen Empfänger unterstützen, damit eine strukturierte Übermittlung möglich ist. Dazu fehlen Standards, welche von allen Akteuren gleichermassen verwendet und akzeptiert werden. Zwar wurden auf Bundes- und Kantonsebene eHealth Standards definiert, in der Realität herrscht aber grosse Abweichung und bei über 80 verschiedenen Praxis-Softwaresystemen und hunderten Gesundheitsorganisationen ist es kaum vorstellbar, einen gemeinsamen Nenner zu finden.

Zukunftsgerichtete Lösungsansätze müssen der Arbeitsweise in der Praxis angepasst sein, manuelle Tätigkeiten reduzieren und Mehrwert generieren. Voraussetzungen, welche mit dem Einsatz von intelligenter Software erfüllt werden können.

[1] Anmeldeformular Radiologie Hirslanden Zürich: https://www.hirslanden.ch/de/klinik-hirslanden/centers/institut-fuer-radiologieundnuklearmedizin/anmeldung1.html

1.3 Anwendungsfall BlueConnect

Die Unterschiede bezüglich des Standes der Digitalisierung zwischen den Gesundheitsorganisationen werden vor allem bei der Kommunikation sichtbar. Versendet die Eine Dokumente noch per Fax, verarbeitet die Andere ihre Daten bevorzugt elektronisch. Da setzt BlueConnect[2], ein bestehendes Produkt der BlueCare AG, an. Es unterstützt den überinstitutionell gerichteten Dokumentenaustausch im Gesundheitswesen. Dokumente werden in BlueConnect aus verschiedenen Kanälen digitalisiert, mittels Informationsextraktion strukturiert und den nachgelagerten Systemen weitergeben. Dies ermöglicht den an der Kommunikation beteiligten Parteien die entsprechenden Prozesse zu automatisieren. Wird beispielsweise ein Patient durch einen niedergelassenen Arzt in ein Spital überwiesen, kann die Spitalsoftware anhand der administrativen Informationen aus BlueConnect automatisch einen neuen Fall eröffnen. Umgekehrt wird die Ablage eines Austrittsberichts vom Spital automatisch beim richtigen Patienten in der Krankengeschichte der Praxissoftware abgelegt. Diverse weitere solcher Prozesse werden durch BlueConnect vereinfacht, wobei sich die Informationsextraktion heute ausschliesslich auf administrative Daten wie Patient, Absender und Empfänger beschränkt.

Wie vorgängig bereits erläutert, werden neben den administrativen aber auch die inhaltlichen, medizinischen Daten strukturiert benötigt. Entsprechend der verschiedenen Geschäftsprozesse wie *Radiologische Anmeldung*, *Spitalaustritt*, *Rezept*, etc. werden von den Empfängerorganisationen unterschiedliche Inhalte verlangt. Denn bei der Anmeldung in einem Radiologischen Institut sind andere Informationen relevant als bei der strukturierten Übermittlung eines Rezepts an eine Apotheke.

[2]BlueConnect Produktbeschreibung: https://www.bluecare.ch/blueconnect

1.4 Zielsetzung

Bevor eine inhaltliche Extraktion und Interpretation der Daten durchgeführt werden kann, muss der entsprechende Geschäftsprozess identifiziert werden. Dieser definiert, welche Daten strukturiert vorliegen müssen. Erst dann können diese dem Empfänger zur automatischen Verarbeitung zur Verfügung gestellt werden. Durch den Einsatz von Textanalyse, speziell der Textklassifikation, werden medizinische Dokumente automatisch den verschiedenen Geschäftsprozessen zugeordnet. Dadurch wird die Basis geschaffen, dass manuelle Prozesse wie das Ausfüllen von Formularen in der Arztpraxis entfallen. Dennoch bietet die Dokumentenklassifikation den Arztpraxen bereits einen direkten Mehrwert. Bisher werden eingehende Dokumente in der Arztpraxis manuell sortiert und der entsprechenden Kategorie des Geschäftsprozesses (Austrittsbericht, Labor, Medikation, etc.) zugeordnet. Folglich können Dokumente automatisch richtig kategorisiert in der Krankengeschichte des Patienten abgelegt werden.

Das Ziel dieser Arbeit ist eine qualitativ hochwertige Zuordnung der Dokumente zu ihrem jeweiligen Geschäftsprozess (Klasse). Dazu werden aktuell eingesetzte Ansätze zur Aufbereitung der Dokumente geprüft und verschiedene Machine Learning Klassifikatoren miteinander verglichen. Als Resultat wird ein Modell zur Verfügung stehen, welches mit einer hohen Präzision Dokumente einer der definierten Klassen zuordnet. Der Service soll einerseits das Produkt BlueConnect erweitern und zusätzlich als Dienstleistung Akteuren im Gesundheitswesen zur Verfügung stehen. Die Informationsextraktion ist nicht Teil dieser Arbeit.

Kapitel 2

Technische Grundlagen

2.1 Textklassifikation

Die Textklassifikation als Teilaspekt des Machine Learnings gibt es schon seit den 60er Jahren, wobei in den letzten Jahren immer mehr Anwendungen in diesem Bereich erfolgten. Dies kann durch die grosse Verfügbarkeit von digitalen Texten erklärt werden. Grosse und frei zu nutzende Textsammlungen wie Wikipedia, Twitter oder Amazon, ermöglichen es im Bereich der Textklassifikation zu forschen und zu arbeiten.

Die Klassifikation von medizinischen Daten ist ein weniger erforschtes Feld. Ein Grund dafür kann sein, dass es kaum öffentliche medizinische Textdokumente gibt. Der Datenschutz ist ebenfalls ein grosses Hindernis. Die Aufbereitung eines entsprechenden Korpus ist somit aufwändig. Die vorliegende Arbeit stützt sich daher auf die Erkenntnisse von nicht medizinischen Texten.

Bei der Textklassifikation handelt sich um einen supervised Ansatz. Das bedeutet, es liegen Daten vor, welche bereits einer bestimmten Klasse zugeordnet sind. Es kann somit ein Modell trainiert werden, welches die zugrundeliegenden Muster in den Daten identifiziert, um unbekannte Daten der richtigen Klasse zuzuordnen.

Beispiele für Anwendungsfälle in der Textklassifikation:

- **Sentiment Analyse**, zur Beurteilung der Stimmung in einem Text. Anwendungen finden sich hier bei der Analyse von Tweets oder Filmreviews.
- **Spam Filter**, wobei ein Text als *Spam* oder *nicht Spam* klassifiziert wird.
- automatische **Spracherkennung** aufgrund des Textes.
- **News Kategorisierung**, wo News Artikel einem bestimmten Thema zugeordnet werden.

Dem gegenüber steht das Textclustering, bei welchem eine Ansammlung

von Textdokumenten ohne eine vorbestimmte Klasse zur Verfügung stehen.

Die für diese Arbeit verfügbaren Daten sind in mehrere Klassen unterteilt worden. Es handelt sich somit nicht um ein binäres Klassifikationsproblem sondern um ein *multiclass* Problem. Ein Dokument kann genau einer von mehreren Klassen zugeordnet werden. Dies ist nicht zu verwechseln mit einer multi-*label* Klassifikation, wo ein Dokument gleichzeitig mehreren Klassen zugeordnet werden kann.

Im Bereich von *multiclass* Textklassifikation stellte Large Scale Hierarchical Text Classification (LSHTC) [7] im Zeitraum von 2009 bis 2014 eine Reihe von Wettbewerben auf. Dazu wurden verschiedene Datensätze und Aufgaben zur Verfügung gestellt. Die Wettbewerbe hatten zum Ziel, einzelne Ansätze zur Klassifikation von Daten im grossen Umfang mit mehreren Tausend Klassen zu beurteilen. Im Rahmen von LSHTC wurde unterschieden zwischen Aufgaben mit multiclass, multilabel sowie hierarchischen Klassen. Entgegen der ersten drei Wettbewerbe wurden im vierten (LSHTC4) einzelne Texte nicht vor-verarbeitet, sondern im Original der Aufgabenstellung mitgegeben. Es war somit nicht mehr eine reine Klassifikationsaufgabe auf Basis von vordefinierten Features, sondern eine vollumfängliche Aufgabe von den Rohdaten bis zur Zuordnung der entsprechenden Klassen, was in etwa der vorliegenden Aufgabe entspricht. Die Gewinner des vierten Wettbewerbs [8] beschreiben die Aspekte der Datenaufbereitung sowie der Klassifikatoren. Die Datenaufbereitung basiert auf verschiedenen Varianten von TF-IDF, was ausschlaggebend zur Wahl für die Feature Extraktion in dieser Arbeit ist.

2.2 Termgewichtung mit TF-IDF

Um Dokumente einander gegenüberstellen zu können und ein statistisches Modell zur Trennung von Klassen zu bekommen, braucht es eine dazu geeignete Repräsentation. Ein häufig verwendeter Ansatz ist das TF-IDF Mass [8, 9]. Dieses beurteilt gegenüber einem Bag-of-Word Ansatz nicht nur die Häufigkeit eines Terms, sondern auch die Wichtigkeit innerhalb des gesamten Textkorpus.

Das TF-IDF Mass besteht aus zwei Teilen, der Term Frequency (TF) und der Inverse Document Frequency (IDF). Dabei gibt TF die Vorkommenshäufigkeit $tf(t, D)$ eines Terms t innerhalb eines Dokuments D an.

Die Inverse Document Frequency $idf(t)$ wird bestimmt durch die Betrachtung der gesamten Anzahl Dokumente N. Terme, welche in weniger Dokumenten enthalten sind werden höher gewichtet. Mit $df(t)$ werden dazu die Anzahl Dokumente bestimmt, in welchen der Term t vorkommt.

$$idf(t) = log\frac{N}{df(t)}$$

Die Wichtigkeit w eines Terms t nimmt somit zu, wenn die Häufigkeit im Dokument steigt, und sie nimmt ab, wenn die Dokumentenhäufigkeit zunimmt.

$$w(t, D) = tf(t, D) * idf(t)$$

Die mittels TF-IDF gewichteten Terme werden bei der Textklassifikation und beim Textclustering als Features verwendet.

2.3 Klassifikatoren

Für die Aufgabenstellung werden 5 Klassifikatoren initial miteinander verglichen. Die zwei Erfolgversprechendsten werden danach optimiert. Diese Beiden, Support Vector Machine und XGBoost, werden daher ausführlicher erklärt. Im initialen Vergleich kommen zudem ein multinominal Naive Bayes, ein K-nearest-neighbor (KNN) sowie ein Random Forest Klassifikator zum Einsatz.

2.3.1 Support Vector Machine

Support Vector Machines (SVM) versuchen mittels einer Hyperebene die durch Vektoren repräsentierten Trainingsdaten optimal zu trennen. Dabei wird der Abstand zwischen der Hyperebene und den nächsten Datenpunkten maximiert, so dass möglichst viel Platz für unbekannte Datenpunkte vorhanden ist.

Eine Eigenheit ist, dass die Hyperebene nicht gekrümmt werden kann. Die Daten müssen daher linear getrennt werden können. Um dies zu ermöglichen verwenden SVMs sogenannte Kernels, welche die Daten in einen höherdimensionalen Raum transformieren.

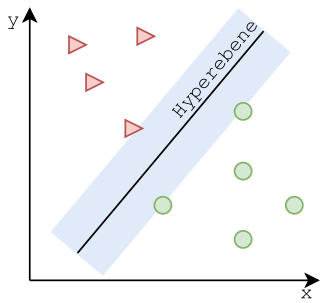


Abbildung 2.1: Optimale Hyperebene zur Trennung von zwei Klassen

Support Vector Machines eignen sich für viele Arten der Klassifikation. Unabhängig von der Dimensionalität der Daten liefern SVMs gute Ergebnisse und eignen sich daher besonders für Textdokumente [5].

2.3.2 Gradient Tree Boosting (XGBoost)

Gemäss einem Blog Beitrag bei Kaggle.com [10], werden auf Machine Learning Problemen mit strukturierten Daten sehr gute Resultate mittels Gradient boosted Trees erzielt. Im Speziellen wird die *parallel tree boosting* Implementierung XGBoost [2] sehr häufig angewendet und gewinnt auch viele der Wettbewerbe [15]. Mittels XGBoost wurde auch ein Wettbewerb auf Textdokumenten gewonnen [13], wobei die Texte zuvor mittels TF-IDF gewichteter Terme vektorisiert wurden.

Aber wie funktioniert das zugrundeliegende Gradient Tree Boosting? Tree Boosting, welches bereits durch AdaBoost eingeführt wurde, basiert auf dem Konzept von sogenannten *weak Learners*, welche nur leicht besser als zufälliges Raten sind, in ihrer *verstärkten* Summe jedoch zu erstaunlich guten Resultaten führen. Häufig handelt es sich bei den *weak Learners* um Decision Stumps, also eine einzige Ebene eines Entscheidungsbaums. Dabei werden die Daten anhand eines einzigen Features an einem bestimmten Wert getrennt. Dies führt zu einfachen Systemen, welche schnell ausgewertet werden können. Durch Boosting der stärkeren und abschwächen der schlechteren Learner kommt das Gesamtsystem zu sehr guten Resultaten.

Gradient Boosting ergänzt nun diesen Ansatz um eine ableitbare Loss Funktion. Das Modell kann durch Hinzufügen von weiteren *weak Learners* diese Loss Funktion minimieren.

XGBoost optimiert das System im Hinblick auf die Performance [11] und Skalierbarkeit. Diese Neuerungen umfassen Parallelisierung, den Umgang mit Daten, welche nicht in das Memory passen und das Auffinden des besten Splits anhand von Quantilen. Die Details zu den jeweiligen Anpassungen sind im XGBoost Paper [2] einzeln beschrieben. Durch diese Anpassungen ist XGBoost enorm skalierbar was die Anzahl zu verarbeitender Datensätze, als auch die Geschwindigkeit betreffen.

2.4 Latent Dirichlet Allocation

Bei Latent Dirichlet Allocation (LDA) [1] handelt es sich um ein generatives Wahrscheinlichkeitsmodell für Textdokumente. Das Modell geht davon aus, dass jedes Dokument aus einer Mischung von zugrundeliegenden Themen besteht. Das Modell versucht dabei die Grenzen zwischen den im Vornherein festgelegten Anzahl Themen in den Dokumenten zu identifizieren. Die häufigsten Wörter werden verworfen und jene Wörter identifiziert, welche die Dokumente optimal zu einem Thema binden und gleichzeitig sich zu den anderen Themen möglichst stark unterscheiden.

2.5 Bewertungsmetriken

Um eine Vergleichbarkeit von verschiedenen Klassifikatoren zu haben, braucht es einheitliche Masse zur Beurteilung der jeweiligen Resultate.

Die Aufgabe des Klassifikators liegt in der Zuordnung eines Dokuments zu einer definierten Klasse. Bei diesem *multiclass* Problem gibt es genau eine richtige Klasse pro Dokument. Die Vorhersage eines Klassifikators kann somit richtig oder falsch sein.

Als Standardmasse für die Beurteilung haben sich die Relevanzmasse **precision**, **recall** und **F1** durchgesetzt. Nachfolgend werden die einzelnen Masse kurz erläutert.

Zur Berechnung aller 3 Masse müssen zuerst die true positives (TP), true negatives (TN), false positives (FP) sowie die false negatives (FN) pro Klasse gezählt werden. TP ist die Anzahl richtig dieser Klasse zugeteilter Dokumente. TN sind die Dokumente, welche richtigerweise als nicht dieser Klasse zugehörig bestimmt wurden. FP, oder auch *type I error*, sind Dokumente, welche fälschlicherweise als dieser Klasse zugehörig bestimmt wurden und FN (*type II error*) sind somit die Dokumente, welche eigentlich zu der gesuchten Klasse gehörten, durch den Klassifikator aber einer anderen Klasse zugeordnet wurden.

Als **Recall** oder **Sensitivität** wird die True Positive Rate (TPR) bezeichnet. Diese gibt den Anteil der korrekt einer Klasse zugeordneter Dokumente, an der Gesamtheit aller dieser Klasse zugehöriger Dokumente an.

$$TPR = \frac{TP}{TP + FN}$$

Mit der **Precision**, auch positive predictive Value (PPV), gibt man den Anteil der richtig zugeordneten Dokumente an der Gesamtheit aller dieser Klasse zugeordneter Dokumente an. Fälschlicherweise als dieser Klasse zugehörig markierter Dokumente führen somit zu einem tieferen Wert.

$$PPV = \frac{TP}{TP + FP}$$

Ein weiteres häufig verwendetes Mass ist die **accuracy**, welche alle korrekt identifizierten Dokumente der Gesamtheit der Dokumente gegenüberstellt.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Der **F1 Score** ist das harmonische Mittel von Precision und Recall. Durch die Verbindung der beiden wichtigen Masse wird der F1 Wert häufig als alleiniger Performancewert eines Klassifikators verwendet.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Die oben genannten Masse werden somit pro Klasse berechnet. Da aber ein Klassifikator für die Gesamtheit der Klassen bewertet werden soll, müssen die einzelnen Masse entsprechend gemittelt werden.

Um die ungleiche Verteilung der Klassen zu berücksichtigen, werden die einzelnen Masse nicht nur gemittelt, sondern auch noch anhand der Klassengrösse (Support) gewichtet.

2.6 Wilcoxon Signed Rank Test

In *Statistical Comparisons of Classifiers over Multiple Data Sets* [6] wurden verschiedene Verfahren zum statistischen Vergleich von Modellen einander gegenüber gestellt. Demšar kam zum Schluss, dass nicht-parametrische Tests, wie der Vorzeichen- oder der Friedman Test, sicherer in der Anwendung sind als Methoden, welche eine Unabhängigkeit der zugrunde liegenden Daten voraussetzen wie dies bei einem t-Test oder ANOVA der Fall ist.

Der Vorzeichen-Test zählt wie häufig das Modell X_1 die besseren Resultate als Modell X_2 geliefert hat und er geht davon aus, dass dies aus einer Binomialverteilung kommt. Mittels eines Hypothesentests wird dann geprüft, ob der Unterschied signifikant ist.

Der Wilcoxon Signed Rank Test (WSR) erweitert den einfachen Vorzeichen-Test um einen gewichteten Rang der einzelnen Differenzen zwischen X_1 und X_2. Neben der Richtung wird somit auch die Stärke der Differenz beurteilt.

WSR stellt die Nullhypothese H_0 auf, dass die Mediane der beiden Stichproben unterschiedlich sind. Bei einer zweiseitigen Hypothese lautet dies somit: $H_0 : \tilde{x}_1 = \tilde{x}_2$

Der WSR Test geht von einer symmetrischen, unabhängigen und identischen Verteilung (i.i.d) der Differenzen D_i aus.
$D_i = X_{i,1} - X_{i,2}$

Die absoluten Werte der Unterschiede D_i werden nach ihrem Wert geordnet (Rang). Die Summe der Ränge, in welchem das erste Modell besser ist, wird als R^+ summiert. R^- entsprechend umgekehrt. Mit $R = min(R^+, R^-)$ wird der kleinere Wert als Teststatistik verwendet und gegen die Normalverteilung geprüft: $R \sim \mathcal{N}(0, 1)$.

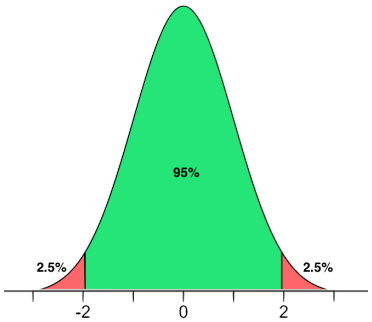


Abbildung 2.2: Normalverteilung mit 5% Verwerfungsbereich (rot)

Bei einem Verwerfungsbereich von 5% bestätigt ein Wert innerhalb der 95% die Nullhypothese. Mit einem Wert ausserhalb kann H_0 verworfen werden.

Kapitel 3

Daten

3.1 Herkunft

BlueConnect[1] ist ein bestehendes Produkt der Firma BlueCare AG, welches den überinstitutionellen, gerichteten Dokumentenaustausch im Gesundheitswesen unterstützt. Die zu verarbeitenden Dokumente werden durch medizinische Institutionen wie Hausarztpraxen, Spitalabteilungen, etc. erstellt. Die Erstellung der Dokumente findet ausserhalb von BlueConnect statt. Die erstellten Dokumente werden über verschiedene Wege wie Fax, eMail, etc. in die zentrale Plattform für die weitere Verarbeitung geladen. Die Dokumente werden auf der Plattform durch Informationsextraktion strukturiert und den nachgelagerten Systemen übergeben.

3.2 Data Labelling

Die aus dem BlueConnect Prozess entstandenen Daten sind in ihrer Form noch ungeeignet für die Erstellung von Modellen. Um ein Modell zu trainieren braucht es *Trainingsdaten*. Diese, auch als *Ground Truth* oder *Gold Standard* bezeichneten Daten, werden optimalerweise durch Fachpersonen erstellt. Dies erfolgt in einem manuellen Prozess, indem die noch nicht klassifizierten Daten einer vordefinierten Klasse zugeordnet werden (Data Labelling). Die Erstellung der manuell klassifizierten Daten ist ein zeitraubender Prozess. Dennoch gilt, je mehr Daten, desto bessere Modelle können erstellt werden. Es ist daher von grossem Vorteil, den Prozess des Data Labellings durch entsprechende Werkzeuge zu optimieren.

Zum Labelling der Dokumente wurde durch das BlueConnect Team eine Webapplikation geschrieben, welche dem Benutzer zufällige Dokumente aus der Gesamtpopulation der Dokumente zusammen mit den möglichen Klassen anzeigt. Nach einer Zuordnung erscheint direkt das nächste noch

[1] BlueConnect Produktbeschreibung: https://www.bluecare.ch/blueconnect

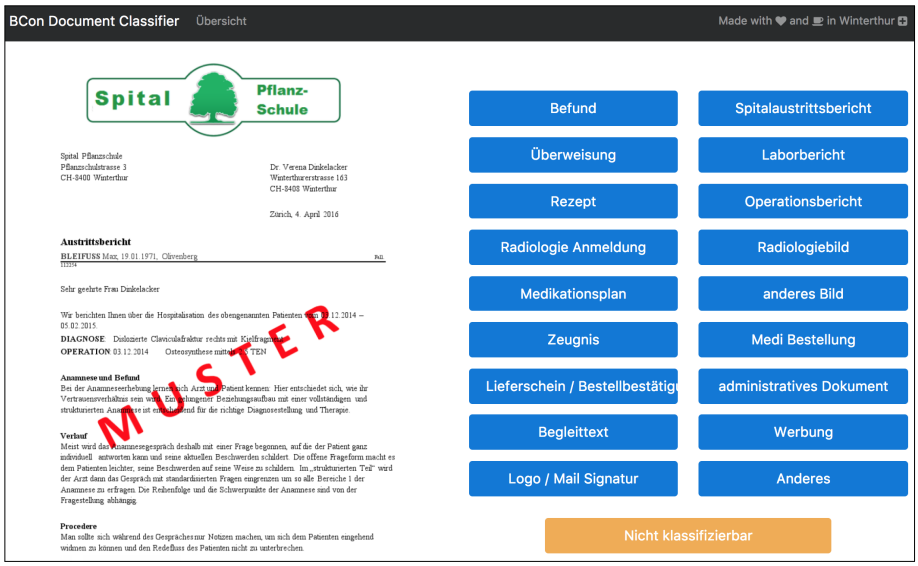


Abbildung 3.1: Beispiel aus der BlueConnect Document Labelling Applikation. Links das zu klassifizierende Dokument und rechts die verschiedenen zur Verfügung stehenden Klassen.

nicht zugeordnete Dokument. Abbildung 3.1 zeigt einen Screenshot aus der BlueConnect Document Labelling Applikation.

Die manuelle Klassifikation wurde durch Hilfspersonen der Arztpraxen durchgeführt und **nicht** durch Fachpersonen wie Ärzte oder MPAs. Jedes Dokument wurde genau einer Klasse zugeordnet. Stichproben von Dokumenten zeigen, dass auch Fehlklassifikationen gemacht wurden, welche sowohl das Trainieren der Modelle als auch die Bewertung eines Klassifikators negativ beeinflusst. Diese wurden, wo erkannt, nachträglich korrigiert.

3.2.1 Klassen

Im Data Labelling Prozess wurden die einzelnen Dokumente einer in Tabelle 3.1 beschriebenen Klasse zugeordnet. Diese Klassen basieren einerseits auf den abzubildenden Geschäftsprozessen, der Ablagestruktur in einer Arztpraxis, sowie den BlueConnect internen Anforderungen zur Aufwandsreduktion.

Tabelle 3.1: Beschreibung aller Klassen

Klasse	*Label*	*Beschreibung*
Überweisung	`transfer`	Das Überweisungsschreiben an eine andere Institution beschreibt die Fragestellung zum Patienten.
Allgemeiner Bericht	`results`	Als Antwort auf ein Überweisungsschreiben oder auch als Information an den Hausarzt. Enthält die Beurteilung des Patienten.
Laborbericht	`lab_report`	Spezifischer Bericht aufgrund einer Laboruntersuchung. Enthält häufig tabellarische Informationen und Messwerte.
Begleittext	`text`	Meist ein eMail Text zu einem Dokument, welches gewisse Zusatzinformationen oder zumindest Patienteninformationen beinhaltet.
Mailsignatur / Logo	`mail_sig`	Für den medizinischen Prozess nicht relevante Texte oder Bilder. Diese könnten früh im Prozess aus der Verarbeitung entfernt werden.
Administratives Dokument	`admin`	Umfasst nicht direkt medizinisch relevante Dokumente wie Versicherungskommunikation, Physioverordnungen u.a.
Rezept	`prescription`	Medikamentenrezept an eine Apotheke.
Austrittsbericht	`hosp_leave`	Umfassender Bericht einer Institution (Spital, Klinik) über den stationären Aufenthalt und die weiterführende Behandlung.
Anderes	`other`	Nicht medizinische Dokumente.
Operationsbericht	`ops_report`	Bericht über die durchgeführte Operation, sowie der weiterführenden Behandlung.
Werbung	`spam`	Werbedokumente
Radiologiebild	`rad_pic`	Radiologiebilder wie Röntgen, Sonographie, Ultraschall.
Bild	`other_pic`	Andere Bilder wie Kardiographie, Bildaufnahmen von Narben, etc.
Anmeldung Radiologie	`radiology_reg`	Schreiben oder Formular zur Anmeldung bei einem radiologischen Institut.
Medikationsplan	`med_plan`	Der Medikationsplan beinhaltet alle durch einen Patienten einzunehmenden Medikamente, inklusive deren Einnahmevorschrift.
Bestellung Medikamente	`med_order`	Medikamentenbestellung einer Arztpraxis bei einem Lieferanten.
Lieferschein	`delivery_note`	Lieferschein von Medikamenten oder medizinischen Utensilien.
Zeugnis	`med_certificate`	Arztzeugnis für Arbeitgeber oder Versicherung.
unklassifizierbar	`unclassifiable`	Nicht einsehbare Dokumente.

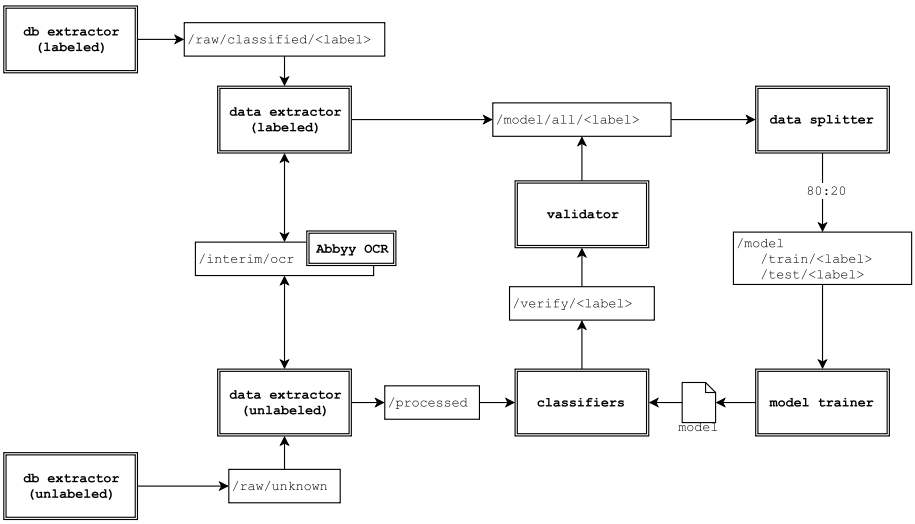


Abbildung 3.2: Kompletter Datenfluss inklusive dem Validierungsprozess von vorklassifizierten Dokumenten mittels eines einfachen Klassifikators.

3.3 Datenaufbereitung

Die Datenaufbereitung nimmt einen grossen Teil der Arbeit an einem Klassifikationsproblem in Anspruch. Gemäss einer Umfrage von CrowdFlower [12, S. 6] im Jahr 2016 macht das Erstellen und Aufbereiten der Daten ca. 80% der Zeit aus. Der Prozess zur Datenbeschaffung und Aufbereitung muss öfters angepasst werden, wenn für die Klassifikation zusätzliche Daten benötigt werden, oder die Struktur der Daten geändert werden soll. Es ist daher von Vorteil die Datenaufbereitung als reproduzierbaren, automatisierten Prozess zu gestalten.

Für die vorliegende Arbeit wurde auf Basis von Cookiecutter[2] eine Struktur definiert und mittels einzelner Python Scripts ein einfacher, reproduzierbarer Prozess erstellt. Die einzelnen Schritte sind in der Abbildung 3.2 zu sehen und werden nachfolgend kurz beschrieben.

3.3.1 Datenextraktion

Abbildung 3.3 zeigt Details der Datenextraktion aus der BlueConnect Datenbank. Dabei werden alle manuell klassifizierten Dokumente identifiziert. Zusammen mit den Metadaten wird das Dokument aus der Datenbank gelesen und in eine definierte Ordnerstruktur geschrieben. Pro Klasse wird ein Verzeichnis erstellt, in welches die Dateien geschrieben werden. Als Name wird der MD5 Hash des Dokuments verwendet. Die Endung wird anhand

[2] python projectstructure for data science: http://drivendata.github.io/cookiecutter-data-science/

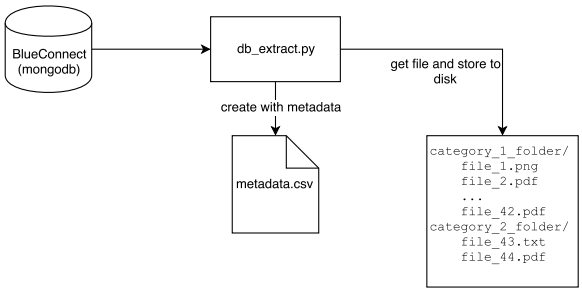


Abbildung 3.3: Detailprozess der Datenextraktion aus der Datenbank und Ablage in der Ordnerstruktur

des Dokumenteninhalts ermittelt und angefügt. Sollten Dokumente aus anderen Systemen klassifiziert werden, ist diese Ordnerstruktur die Grundlage für die weiteren Prozessschritte.

3.3.2 Textextraktion

Die Textextraktion erfolgt auf Basis der PDF Dateien mittels des Linux Tools `pdftotext`[3]. Ist kein lesbarer Text vorhanden, wie dies zum Beispiel bei eingehenden Fax Dokumenten der Fall ist, kommt zusätzlich eine OCR Engine (Optical Character Recognition) zum Zug. Diese erkennt auf den Bilddokumenten den Text und generiert daraus ein lesbares PDF, aus welchem der Text entnommen werden kann.

Dieser Prozess wurde später so angepasst, dass alle Dokumente mittels Abbyy OCR[4] verarbeitet wurden und der zusätzlich erkannte Text dem entsprechenden Dokument hinzugefügt wurde. Die Erkenntnis kam aus der Analyse der ersten Resultate von *Radiologie Anmeldungen*, wobei es sich häufig um ein vorgefertigtes Formular handelt aus welchem nur der Inhalt der Formularfelder extrahiert werden konnte. Somit fehlten die für die Klassifikation wichtigen Begriffe.

3.4 Problem der Dokumentenverteilung

Nach einer ersten Phase der manuellen Klassifikation zeigt sich eine ungleiche Verteilung der Dokumente (Tabelle 3.2). *Berichte* und *Überweisungen* machen fast die Hälfte aller Dokumente aus. Gewisse Dokumenttypen, wie *Zeugnisse* oder *Radiologiebilder* sind kaum vorhanden, als Geschäftsprozess aber auch nicht prioritär. Andere Klassen wie *Medikationsplan*, *Radiologieanmeldung* oder auch *Lieferscheine* sind jedoch auch sehr stark unterbesetzt

[3] pdftotext basiert auf Poppler: https://linuxappfinder.com/package/poppler-utils
[4] Abbyy Finereader für Mac: https://www.abbyy.com/en-eu/finereader/pro-for-mac/

Tabelle 3.2: Verteilung der Dokumente auf die einzelnen Klassen nach manuellem Labelling (Phase 1) sowie nach Vor-Klassifikation durch einen einfachen Klassifikator und Klassenreduktion (Phase 2).

	Phase 1		*Phase 2*
Klasse	*# (%)*	*F1*	*# (%)*
`results`	1023 (33.2%)	0.79	1370 (24.4%)
`transfer`	368 (12.0%)	0.83	485 (8.4%)
`lab_report`	350 (11.4%)	0.87	452 (8.0%)
`text`	283 (9.2%)	0.64	-
`mail_sig`	207 (6.7%)	0.50	-
`admin`	162 (5.3%)	0.44	-
`prescription`	161 (5.2%)	0.90	**917 (16.3%)**
`hosp_leave`	109 (3.5%)	0.57	**452 (8.0%)**
`other`	97 (3.2%)	0.28	1257 (22.4%)
`ops_report`	63 (2.0%)	0.67	**166 (3.0%)**
`spam`	50 (1.6%)	0.43	-
`other_pic`	45 (1.5%)	0.18	-
`radiology_reg`	42 (1.4%)	0.63	**245 (4.4%)**
`med_plan`	31 (1.0%)	0.92	**115 (2.0%)**
`delivery_note`	26 (0.8%)	0.86	**156 (2.8%)**
`unclassifiable`	21 (0.7%)	0.00	-
`rad_pic`	17 (0.6%)	0.00	-
`med_order`	16 (0.5%)	0.57	-
`med_certificate`	8 (0.3%)	0.00	-
	3079		**5615**

aber wichtig. Um eine genügend grosse Anzahl Dokumente für diese Klassen zu bekommen, müssten enorm viele Dokumente manuell klassifiziert werden.

3.4.1 Boosting durch Einsatz eines einfachen Klassifikators

Um auf effiziente Weise genügend Dokumente für die einzelnen Klassen zu bekommen, wurde eine einfache Support Vector Machine auf den in der Phase 1 verfügbaren Dokumenten trainiert und in den Datenprozess eingebunden. Die Anzahl Trainingsdaten sowie die F1 Werte der jeweiligen Klassen sind in Tabelle 3.2 ersichtlich.

Wie in Abbildung 3.2 ersichtlich, wurde das trainierte Modell auf noch ungesehenen Dokumenten aus der Datenbank angewendet. Die daraus klassifizierten Dokumente wurden mittels einer angepassten Klassifikationsapplikation validiert. Dabei wurden nur Dokumente einer bestimmten Klasse angezeigt. Es können somit effizient zusätzliche Dokumente einer bestimmten, unterbesetzten Klasse zugeordnet werden. Dieser Prozess wurde auf den zeitnah relevanten Geschäftsprozessen durchgeführt. Die Resultate dazu sind ebenfalls in Tabelle 3.2 unter Phase 2 zu finden.

Da der Klassifikator Dokumente findet, welche er bereits identifizieren kann, entsteht eine gewisse Einseitigkeit der zur Verfügung stehenden Dokumente. Da für die vorliegende Aufgabenstellung die Präzision das wichtigere Bewertungskriterium ist, wird diese Einseitigkeit so akzeptiert.

3.4.2 Klassenreduktion

Für bessere Resultate werden zusätzlich die kleineren, und vorerst nicht relevanten Klassen zusammengefasst. Dies führt für die weiteren Modelle zu einer Reduktion von 19 auf 10 Klassen. Folgende Klassen werden unter `other` zusammengefasst: `med_certificate`, `text`, `spam`, `mail_sig`, `other_pic`, `unclassifiable`, `med_order`, `admin`.

Die Klassenreduktion führt zusammen mit dem boosted labelling zu einer neuen Verteilung der Dokumente (siehe Tabelle 3.2).

3.5 Data Splitting

Abbildung 3.4 zeigt wie die Gesamtheit der Dokumente mit dem Verhältnis 80/20 in Trainings- und Testdaten unterteilt wird. Die Testdaten dürfen nicht zum Training oder der Evaluation von Modellen gebraucht werden. Die Testdaten werden nur zur abschliessenden Beurteilung eines trainierten Modells verwendet.

Da eine ungleiche Verteilung der Dokumente pro Klasse vorliegt, werden die Dokumente stratifiziert den jeweiligen Sets zugeordnet. Das heisst, in allen Trainings-, Test- und Validierungssets sind proportional gleich viele Dokumente einer Klasse vorhanden, wie in der Gesamtheit der Daten.

Für die Evaluation und die Optimierung werden Modelle mehrmals trainiert, um auf die Verteilung zu schliessen und so eine bessere Aussage zur Qualität des entsprechenden Modells treffen zu können. Dazu werden aus den Trainingsdaten nach Bedarf n Validierungssets erzeugt, und wiederum stratifiziert in 80% Trainings- und 20% Testdaten aufgeteilt. Zusätzlich werden diese in ihrer Reihenfolge gemischt, um eine mögliche sequentielle Abhängigkeit der Daten auszuschliessen.

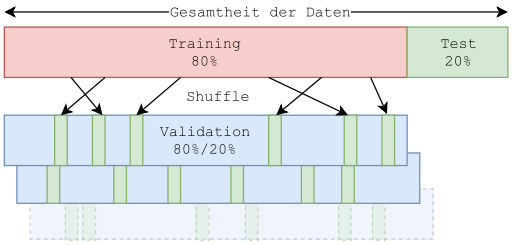


Abbildung 3.4: Aufteilung der Daten in Training- und Testset mit zusätzlicher Erstellung von n Validierungssets aus den Trainingsdaten

Kapitel 4

Explorative Analyse mittels Topic Modelling

4.1 Vorgehen

Bevor mittels einer supervised Klassifikation ein Modell trainiert wird, soll mit einem *un*supervised Ansatz überprüft werden, ob die Dokumente sich in Gruppen unterteilen lassen, welche sich mit den definierten Klassen decken.

Da die Dokumente in einem Vektorraum repräsentiert werden, müssen sich die Dokumentenvektoren einer Klasse von den Vektoren der anderen Klassen unterscheiden lassen. Dies soll mit Hilfe der *Latent Dirichlet Allocation (LDA)* geprüft, und mit *t-distributed stochastic neighbor embedding (t-SNE)* eine geeignete Visualisierung gefunden werden.

Das LDA Topic Modelling wird auf der Gesamtheit der Dokumente erstellt. Da die Klasse `other` aus vielen unterschiedlichen Themen besteht und keine direkte Zuordnung zu einem Geschäftsprozess vorhanden ist, werden diese Dokumente nicht für das Topic Modelling verwendet. Das LDA Modell soll somit 9 verschiedene Themen finden, welche möglichst diesen 9 Klassen entsprechen.

4.2 Topic Modelling

Das Resultat einer LDA Modellierung ist eine Matrix, welche für jedes einzelne Dokument eine Wahrscheinlichkeit der Zuordnung zu einem bestimmten Thema angibt. Anhand dieser kann ein Dokument nicht nur einem Thema zugeordnet werden, es gibt auch an mit welchem Anteil ein bestimmtes Dokument zu einem anderen Thema passt. Denn bei einer genaueren Betrachtung einzelner Dokumente, bestehen diese aus einer Sammlung von mehreren Klassen. Beispielsweise beinhaltet ein *Austrittsbericht* die Informationen eines allgemeinen Berichts, aktuelle Laborwerte (*Laborbericht*), eine Austrittsmedikation (*Medikationsplan*) und unter Umständen noch weitere

Informationen wie eine Nachversorgung.

Das mit LDA erstellte Modell hat so viele Dimensionen, wie Themen gefunden werden sollen. Dies entspricht den 9 eingegebenen Klassen, was sich nicht zur Darstellung eignet. Durch eine Methode der Dimensionsreduktion, wie einer Hauptkomponentenanalyse (PCA) oder t-SNE, können die einzelnen Datenpunkte auf zwei Dimensionen reduziert und in einem Diagramm dargestellt werden.

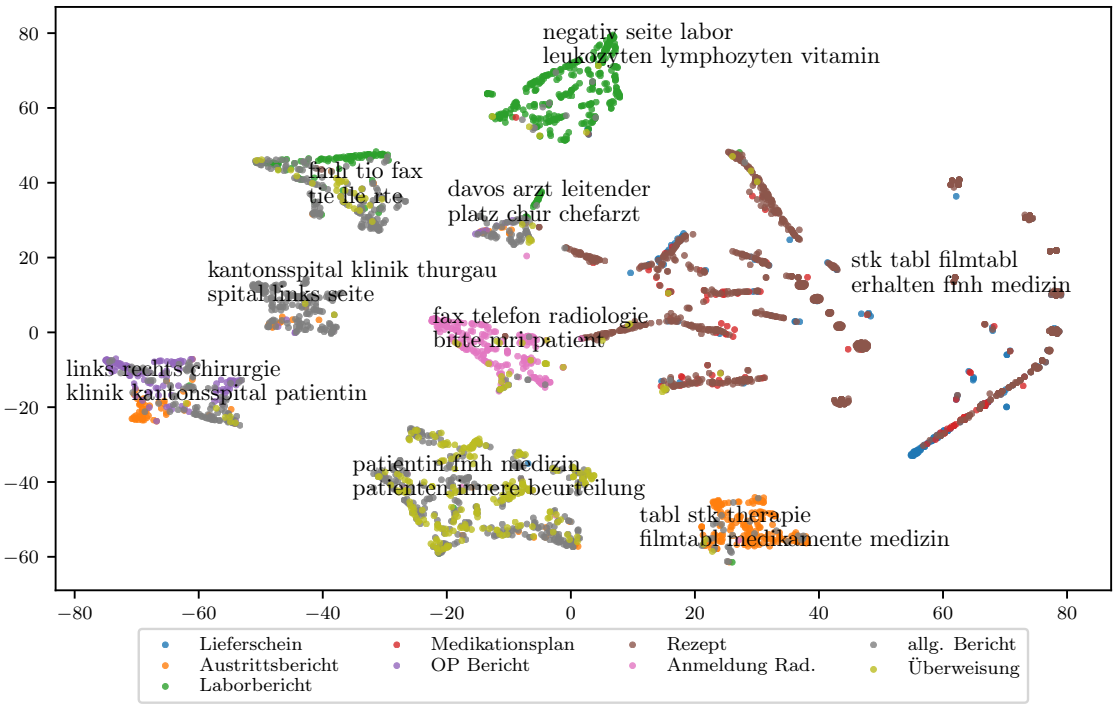
4.3 Dimensionsreduktion und Visualisierung

PCA liefert eine Funktion zur linearen Transformation der hochdimensionalen Features auf einen niedrig dimensionalen Raum. t-SNE hingegen versucht für jeden Punkt eine Repräsentation im niedrig dimensionalen Raum zu finden, so dass ähnliche Punkte nahe beieinander und unähnliche weiter voneinander entfernt dargestellt werden. t-SNE ist daher zur Visualisierung besser geeignet, da die Gruppierungen optisch klarer getrennt dargestellt werden.

Durch nachträgliche Einfärbung der Datenpunkte anhand ihrer effektiven Klasse kann visuell die Qualität des erstellten Modells geprüft werden. Sind die Daten in klar getrennte Punktewolken aufgeteilt, hat das zugrundeliegende LDA Modell eindeutige Themen gefunden. Sind die nah beieinander liegenden Punkte auch noch gleich eingefärbt, stimmt das durch LDA identifizierte Thema mit der realen Klasse der Dokumente überein. In diesem Fall wird auch ein supervised Ansatz mit einer entsprechend hohen Qualität die Klasse identifizieren können. Das Resultat der Analyse ist in Abbildung 4.1 zu sehen.

Abbildung 4.1: Mit t-SNE visualisierte LDA Themen. Die einzelnen Datenpunkte sind anhand ihrer definierten Klasse eingefärbt. Pro Gruppe sind zudem die 6 wichtigsten Wörter des LDA Modells dargestellt. Inspiration und Code-Beispiele von Shuai's Data Blog [14]

.



4.4 Interpretation

Die Darstellung zeigt klare Gruppen. Radiologische Anmeldungen, Austrittsberichte und Rezepte sind klar abgegrenzt. Die Laborberichte (grün) scheinen sich in in zwei Gruppen zu unterteilen. Dies kann zum Beispiel aufgrund der verschiedenen Laborinstitutionen, mit unterschiedlichen Namen und Bezeichnungen zustande kommen. Die allgemeinen Berichte (grau) mischen sich unter die Austrittsberichte. Dies erklärt sich dadurch, dass eine Spitalabteilung nicht nur stationäre Behandlungen mit einem Austrittsbericht, sondern auch ambulante Behandlungen mit einem allgemeinen Befund durchführen. Die Rezepte mischen sich mit den Lieferscheinen (blau) und den Medikationsplänen (rot). Dies erklärt sich durch das gemeinsame Vokabular von Medikamenten. Das LDA Modell hat auch nach Regionen, bzw. nach Institutionen getrennt. Eine Gruppe setzt sich aus Dokumenten der Spitalregion Thurgau zusammen. Eine andere Gruppe formiert sich rund um die Region Davos / Chur. Dies weist auf ein Bias in den Daten hin. Es ist davon auszugehen, dass keine homogene Verteilung der Dokumente

Tabelle 4.1: LDA Themen mit Termen und einer möglichen Beschreibung

#	*Begriffe*	*Beschreibung*
1	stk tabl filmtabl erhalten fmh medizin rezept ean innere allgemeine	Rezepte, Medikationsplan, Lieferscheine
2	negativ seite labor leukozyten lymphozyten vitamin wert einheit mch mcv	Laborberichte
3	patientin fmh medizin patienten innere beurteilung patient freundliche anamnese aktuell	Berichte / Ueberweisung
4	davos arzt leitender platz chur chefarzt spital fmh promenade medizin	Berichte Spital Davos / Chur
5	tabl stk therapie filmtabl medikamente medizin verlauf austrittsbericht eintritt austritt	Austrittsberichte
6	fax telefon radiologie bitte mri patient schaffhausen termin untersuchung fmh	Anmeldungen Radiologie
7	links rechts chirurgie klinik kantonsspital patientin operation fax sprechstunde schmerzen	Operationsberichte
8	fmh tio fax tie lle rte innere lin gastroenterologie	unbekannt. Innere Medizin / Gastro
9	kantonsspital klinik thurgau spital links seite psychiatrische rechts dienste katharinental	Befunde Spitalverbund Thurgau

über die Regionen der Schweiz, sondern eine Fokussierung auf die Region Ostschweiz vorhanden ist.

Tabelle 4.1 zeigt die 10 häufigsten Wörter der einzelnen Themen mit einer möglichen Beschreibung. Viele Begriffe sind Fachbegriffe aus der Medizin und dem Gesundheitswesen.

Kapitel 5

Modellerstellung

5.1 Vorgehen

Auf Basis der aufbereiteten Daten soll nun ein Modell trainiert werden, welches mit einer hohen Präzision unbekannte Dokumente einer der definierten Klassen zuordnen kann.

Ein wichtiger und erster Teil des Modells ist die Extraktion von sinnvollen Features. Auf diesen werden verschiedene Klassifikatoren mit Basisparametern mittels Kreuzvalidierung trainiert und miteinander verglichen. Die zwei besten Modelle werden in einem nächsten Schritt auf den Trainingsdaten optimiert und auf den Testdaten beurteilt. Die daraus resultierenden Werte zeigen die Qualität der Modelle und lassen einen ersten Vergleich zu. Ein Blick auf die Lernkurven ermöglicht eine Aussage zu Over- / Underfitting. Daraus lässt sich ableiten, ob weitere manuell klassifizierte Trainingsdaten eine Verbesserung der Modelle bringen würde. Ein statistischer Vergleich, mittels des Wilcoxon Signed Rank Test, gibt schlussendlich Auskunft, ob ein signifikanter Unterschied zwischen den Modellen existiert.

5.2 Feature Extraktion

Bevor ein Modell trainiert werden kann, müssen aus den einzelnen Dokumenten sinnvolle Features extrahiert werden. Die Basis dafür sind die mittels Abbyy OCR verarbeiteten Textdokumente. Diese werden durch die in Abbildung 5.1 dargestellte Pipeline in Features umgewandelt, welche als Eingangsparameter für den jeweiligen Klassifikator zur Verfügung gestellt werden. Durch diese Pipeline stehen schlussendlich pro Dokument die nach TF-IDF gewichteten Terme zur Verfügung.

Zu Beginn wurden noch weitere Features erstellt, welche jedoch keine besseren Werte ergaben und daher nicht verwendet werden. Diese Features waren: Textlänge der Dokumente, Anzahl Medikamente im Dokument, Anzahl radiologischer Begriffe im Dokument.

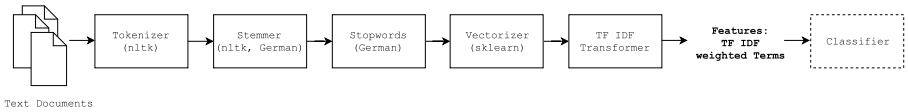


Abbildung 5.1: Bestandteile der Feature Pipeline

5.3 Auswahlverfahren von Klassifikatoren

Zu Beginn sollen verschiedene Klassifikatoren auf den Daten trainiert und anhand der Bewertungsmetriken miteinander verglichen werden. Aufgrund der Resultate werden zwei Klassifikatoren ausgewählt um optimiert zu werden. Dies führt zu einer Auswahl der aufgrund von Standardparmeter am besten funktionierenden Klassifikatoren. Natürlich besteht die Möglichkeit, dass einer der dadurch verworfenen Klassifikatoren, mittels einer Parameteroptimierung, zu besseren Werten als der schlussendlich Gewählte führen könnte.

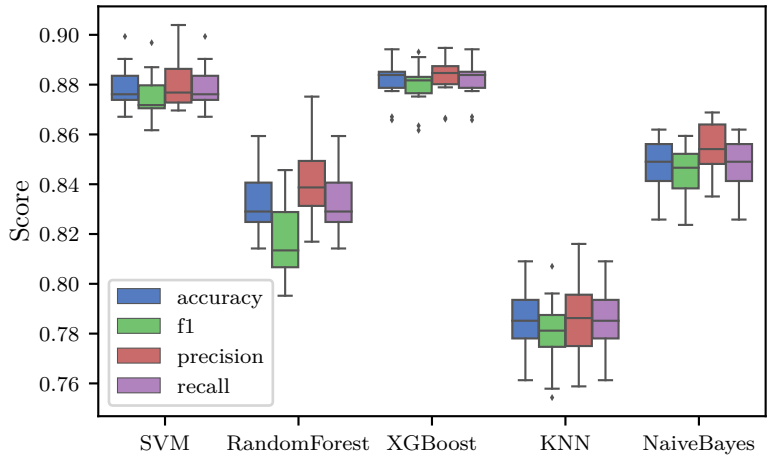
Die ausgewählten Klassifikatoren und deren verwendete *out-of-the-box* Parameter sind in Tabelle 5.1 ersichtlich. Als Kriterien wurden die gewichteten Werte für *F1*, *Recall* und *Precision* sowie die *Accuracy* verwendet.

Tabelle 5.1: Klassifikatoren mit initialen Parametern für einen ersten Vergleich

Classifier	*Parameters*
Support Vector Machine	kernel = 'linear', C = 1.0
Random Forest	n_estimators = 50, class_weight='balanced'
eXtreme Gradient Boosting	-
k-nearest neighbors	n_neighbors=5, weights='distance'
Multinominal Naive Bayes	alpha=0.001

Jeder Klassifikator wurde 10 Mal auf einem, in Kapitel 3.5 beschriebenen, Validierungsset trainiert und getestet. Das Diagramm in Abbildung 5.2 zeigt die Werte pro Klassifikator.

Abbildung 5.2: Resultate verschiedener Klassifikatoren



Es zeichnen sich zwei klare Favoriten ab. Sowohl die Support Vector Machine als auch XGBoost erzielen Werte im ungefähr selben Bereich, wobei der Range des XGBoost kleiner ist, was auf einen stabileren Klassifikator hinweist. Auf Basis dieser Auswahl wird nachfolgend eine Support Vector Machine als auch ein XGBoost Klassifikator auf den Daten optimiert.

5.4 Vergleich von SVM und XGBoost

Die beiden gewählten Klassifikatoren werden nun auf den Trainingsdaten optimiert und auf den Testdaten miteinander verglichen.

Für die Optimierung wird an die Feature Pipeline aus Kapitel 5.2 nun der Klassifikator angehängt und mittels verschiedener Parameterkombinationen aus Tabelle 5.2 der Klassifikator mit den besten Werten ermittelt. Anhand dieser Parameter ergeben sich 576 Parameterkombinationen für die Support Vector Machine und 96 Kombinationen für die XGBoost Pipeline. Jede Parameterkombination wird auf 10 unterschiedlichen Validierungssets trainiert. Dies führt zu 5760 Berechnungen für die Support Vector Machine und 960 Berechnungen für die XGBoost Pipeline.

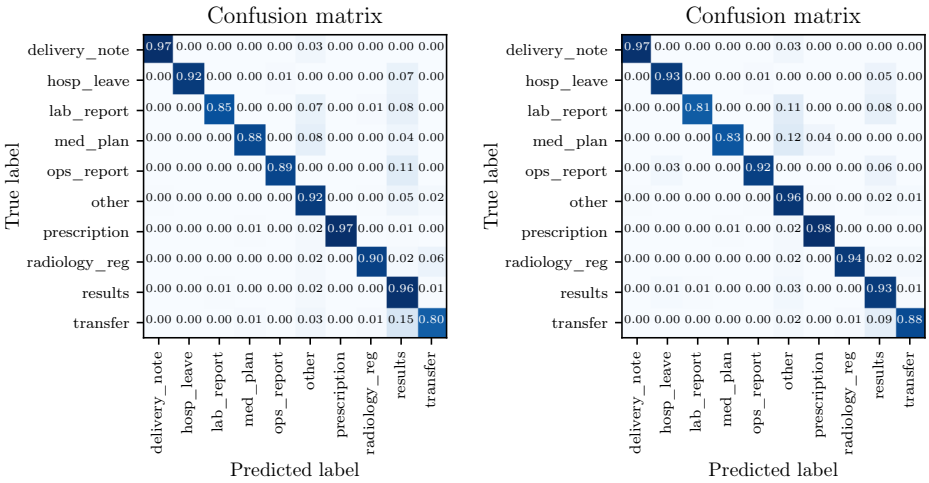
Tabelle 5.2: Parameterbereich für das Training der gesamten Pipeline sowie die besten Werte für SVM und XGBoost

Parameter	*Wertebereich*	*SVM*	*XGBoost*
vectorizer: max_features	(1000, 2000, 5000, 10000)	10000	5000
vectorizer: max_df	(0.5, 0.75, 1.0)	0.5	0.75
vectorizer: ngram_range	((1, 1), (1, 2))	(1, 1)	(1, 1)
tfidf: use_idf	(True, False)	True	False
tfidf: norm	('l1', 'l2')	l2	l1
SVM: kernel	('linear', 'rbf')	linear	-
SVM: C	(0.1, 1, 10)	1	-
XGBoost: -	-	-	-

Mit den besten Parametern wird nun auf den Trainingsdaten ein neues Modell trainiert und auf den Testdaten angewendet. Die Masse zur Bewertung pro Klasse sind in Tabelle 5.3 nebeneinander dargestellt. Zusammen mit der jeweiligen Confusion Matrix in Abbildung 5.3 zeigt sich ein sehr ähnliches Bild der Support Vector Machine und des XGBoost Klassifikators.

Tabelle 5.3: Vergleich der Scores von SVM und XGBoost

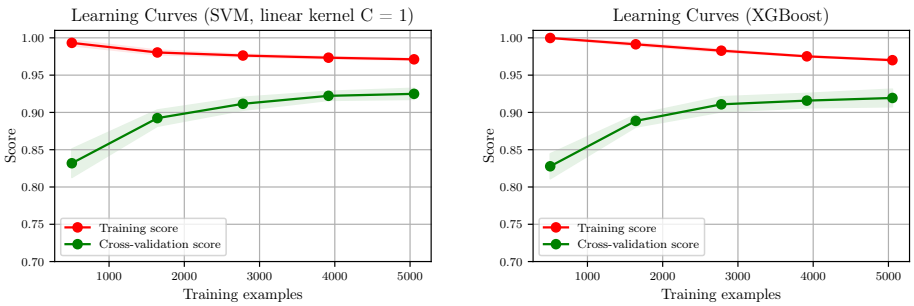
		Support Vector Machine			*XGBoost*		
class	*support*	*Precision*	*Recall*	*F1*	*Precision*	*Recall*	*F1*
delivery_note	32	0.97	0.97	0.97	0.97	0.97	0.97
hosp_leave	92	0.99	0.92	0.96	0.96	0.93	0.95
lab_report	91	0.97	0.85	0.91	0.97	0.84	0.90
med_plan	24	0.91	0.88	0.89	0.95	0.88	0.91
ops_report	36	0.97	0.89	0.93	0.92	0.92	0.92
other	246	0.91	0.92	0.92	0.89	0.94	0.91
prescription	185	1.00	0.97	0.98	0.98	0.97	0.98
radiology_reg	51	0.94	0.90	0.92	0.98	0.92	0.95
results	272	0.85	0.96	0.90	0.88	0.92	0.90
transfer	98	0.89	0.80	0.84	0.91	0.85	0.88
total / avg	**1127**	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**



(a) Support Vector Machine **(b)** XGBoost

Abbildung 5.3: Normalisierte Confusion Matrix der beiden gewählten Klassifikatoren

Anhand der Lernkurve der beiden Modelle kann man sehen wie sie sich im Bezug auf die Anzahl der Trainingsdaten verhalten. Die Kurven in Abbildung 5.4 zeigen ein nahezu identisches Bild zwischen SVM und XGBoost. Bei genauer Betrachtung scheint es, als würde die Support Vector Machine mit leicht weniger Trainingsdaten zu einem besseren Resultat führen. Durch den etwas höheren Training Score kann auf ein leichtes Overfitting auf den Daten geschlossen werden, welches sich mit zusätzlichen Trainingsdaten relativieren würde.



(a) Support Vector Machine **(b)** XGBoost

Abbildung 5.4: Lernkurven der beiden Modelle

5.5 Vergleich mittels Wilcoxon Signed Rank Test

Die gemittelten Werte der beiden Modelle in Tabelle 5.1 sind für die Support Vector Machine und den XGBoost identisch. Bei einigen Klassen ist die Support Vector Machine besser, bei anderen Klassen der XGBoost. Welcher Klassifikator soll nun für einen produktiven Einsatz verwendet werden?

Die Werte in Tabelle 5.1 zeigen das Resultat eines trainierten Modells auf genau einem definierten Testset. Es kann nun sein, dass diese Daten dem einen Klassifikator besser liegen als dem Anderen. Werden aber die gleichen Klassifikatoren mehrmals auf unterschiedlichen Daten trainiert, kann mittels statistischer Methoden eine Aussage gemacht werden, ob das eine Modell gegenüber dem Anderen zu bevorzugen ist.

Mit dem nicht parametrischen Wilcoxon Signed Rank Test (WSR) kann festgestellt werden, ob zwei Stichproben aus Populationen mit gleicher Verteilungen kommen. Die Nullhypothese lautet:

$$H_0 \textbf{ : Die Verteilungsfunktionen der Stichproben } X_1 \textbf{ und } X_2 \textbf{ sind identisch}$$

oder:

$$H_0 : \tilde{x}_1 = \tilde{x}_2$$

Die Stichproben bekommt man, indem beide Modelle auf mehreren unterschiedlichen, stratifizierten Datensätzen trainiert, und auf entsprechenden Testdaten der gewichtete F1 Wert bestimmt wird. Auf diesen F1 Werten wird mittels WSR der p-Wert berechnet.

Für die Bestimmung der F1 Werte werden beide Modelle auf **40** unterschiedlichen Validierungssets trainiert und getestet. Das Resultat des Experiments ergibt folgende Parameter der beiden Stichproben:

$$\text{SVM: } \overline{x} = 0.91, \sigma = 0.015$$
$$\text{XGBoost: } \overline{x} = 0.92, \sigma = 0.014$$

Auf Basis der einzelnen Werte der Stichproben liefert der WSR einen **p-Wert** von $p = 0.0066$. Somit kann H_0 auf dem Signifikanzniveau $\alpha = 5\%$ verworfen werden. Das Modell mit XGBoost als Klassifikator ist dem Modell mit der Support Vector Machine überlegen.

Kapitel 6

Fazit

6.1 Diskussion

Es zeigte sich rasch, dass im Gesundheitswesen ausgetauschte Dokumente eindeutig verschiedenen Themen zugeordnet werden können. Durch ein LDA Topic Modelling konnte aufgezeigt werden, dass die zugrundeliegenden Dokumente ein Vokabular besitzen, welches sich klar trennen lässt. Durch das Trainieren von geeigneten Textklassifikationsmodellen konnte eine sehr hohe Präzision in der automatischen Klassifikation erzielt werden. Ein Modell mit einem F1 Wert von 0.92 über die zehn definierten Klassen übertraf die Erwartungen deutlich. Das gesetzte Ziel, eine qualitativ hochwertige Zuteilung von medizinischen Dokumenten zu ihrem jeweiligen Geschäftsprozess zu ermöglichen, wurde somit erreicht.

Vorgängig zum Projekt wurden die verwendeten Daten in einem separaten Prozess manuell klassifiziert. Ein solches, durch Personen vorgenommenes Labelling, ist generell fehleranfällig und es zeigte sich auch, dass einige Dokumente falsch zugeordnet waren und korrigiert werden mussten. Im Laufe der ersten Modellfindung kamen bei Stichproben aus falsch klassifizierten Testdaten diese widersprüchlichen Dokumente zum Vorschein. Unklare Arbeitsanweisungen führten dazu, dass beispielsweise Physiotherapieverordnungen teils als Auftrag (`transfer`) und teils als administratives Dokument (`admin`) markiert waren.

Wie erwartet haben die Datenbeschaffung und Datenaufbereitung am meisten Aufwand generiert. Was auch gerechtfertigt ist, denn die Qualität der Klassifikation ist direkt damit verbunden: den grössten Einfluss hat nicht der Klassifikator oder die extrahierten Features aus den Texten, sondern die Qualität und Menge der zur Verfügung stehenden Daten. Nach ersten Iterationen von Training und Test hat sich gezeigt, dass bei einer hohen Anzahl von Dokumenten der extrahierte Text nicht auf die Klasse schliessen lässt. Dabei gibt es folgende Fälle:

1. Ein Bilddokument (Fax) wird nachträglich mit Text ergänzt. Es wird nur dieser Text extrahiert, wobei davon nicht auf die richtige Klasse geschlossen werden kann.
2. Ähnlich wie beim ersten Fall gibt es vorgefertigte Formulare, welche zum Beispiel zur Anmeldung bei einem radiologischen Institut verwendet werden. Eine einfache Textextraktion findet nur die im Formular ausgefüllten Felder. Es gibt keine Informationen, welche das Formular, bzw. den Prozess betreffen. Eine Klassifikation ist somit kaum möglich.

Dies konnte mittels vollständiger Anwendung von OCR auf allen Dokumenten behoben werden. Damit wurden auch im Formular hinterlegte Logos oder vorgedruckte Adressen, welche einen Hinweis auf einen möglichen Institutionstypen geben, dem Klassifikator zur Verfügung gestellt. Der Einsatz eines qualitativ hochwertigen OCR Systems wie Abbyy ist grundsätzlich zu empfehlen. Die generierten PDF Dokumente beinhalten dann alle auf dem Dokument vorhandenen Texte. Die bereits in BlueConnect integrierte Extraktion von administrativen Informationen kann genauso davon profitieren, wie es als ideale Grundlage für eine Dokumentenklassifikation oder Informationsextraktion dient.

Kurze Tests mit einschlägigen Features auf den Textdokumenten wie Textlänge, radiologische Ausdrücken oder Medikamentennamen sind bei der Verwendung von TF-IDF zu vernachlässigen. Die Verwendung von TF-IDF mit leicht angepassten Stoppwörtern und Stemming bietet eine sehr gute Ausgangslage für die Klassifikation. Interessant ist, dass für die Support Vector Machine die Verwendung des IDF Wert zu besseren Resultaten geführt hat. Beim XGBoost wurden jedoch die besseren Resultate ohne IDF erzielt.

Mit XGBoost wurde ein optimaler Klassifikator für das Modell gefunden. Der damit erzielte F1 Wert von 0.92 über alle Klassen ist sehr hoch. Erstaunlich sind auch die Werte einzelner Klassen wie 0.98 bei Rezepten oder 0.95 bei radiologischen Anmeldungen. Dies bietet eine sehr gute Ausgangslage zur weiteren Verarbeitung dieser Geschäftsprozesse. Zudem zeigt die Lernkurve der Modelle, dass Potential für noch höhere Werte besteht, indem weitere Dokumente manuell klassifiziert und so die Anzahl der verfügbaren Trainingsdaten erhöht wird.

Aktuell findet der eingesetzte Klassifikator XGBoost generell eine grosse Verbreitung bei der Lösung von Machine Learning Problemen. Er ist als Klassifikator zudem sehr interessant, da er keine Hyperparameter benötigt. Im direkten Vergleich zur Support Vector Machine, wo neben dem Kernel (2 Parameter) auch der Penalty Parameter C (3 Parameter) gefunden werden musste, führt dies zu rund 83% weniger zu berechnenden Kombinationen. Gerade bei der Verwendung mehrere Kreuzvalidierungen pro Parameterset führt dies zu einer deutlich schnelleren Optimierung mit XGBoost.

6.2 Ausblick

Das vorliegende Modell kann Textdokumente zuverlässig unterscheiden. Die zugrunde liegenden Daten beinhalten zusätzlich Bilddokumente, wobei zwischen allgemeinen Bildern und radiologischen Bildern unterschieden wurde. Das bestehende System könnte also um eine Bildklassifikation erweitert werden. Radiologische Bilder wie Röntgen oder CT könnten als Geschäftsprozess weiter verarbeitet werden. Es gibt heute bereits Forschungsbereiche welche mittels Machine Learning, dem Arzt Hilfestellungen geben bei der Beurteilung von Bildern dieser Art.

Dem in der Einleitung erwähnten Anwendungsfall und der damit verbundenen Informationsextraktion steht jetzt nur noch wenig im Weg. Mit dem vorliegenden Modell können die relevanten Dokumente identifiziert werden. Diese Dokumente müssen jedoch auch wieder manuell vorbereitet werden. Der Annotationsprozess ist im Bereich der Informationsextraktion aufwändig, da einerseits verschiedene Informationen gefragt sind, anderseits auch angegeben werden muss, wo im Dokument sich diese Informationen befinden. Ein solches Modell kann auf einzelnen Geschäftsprozessen zu deutlichen Einsparungen und Prozessoptimierungen führen. Denkbar wäre eine schrittweise Umsetzung, angefangen bei dem Geschäftsprozess mit dem grössten Potential bzw. dort wo die Empfänger mit ihren IT Systemen bereit sind. Beispielsweise bei radiologischen Anmeldungen.

Ein weiteres interessantes Forschungsgebiet ist die Gliederung von Dokumenteninhalten. Wie sich beim Topic Modelling zeigt, besteht ein Dokument aus mehreren Themen. Damit könnte beispielsweise beim Geschäftsprozess *Austrittsbericht* das Dokument in die darin enthaltenen Bereiche Diagnose, Befund, Medikation, Labor und Nachsorge aufgeteilt werden. Mittels einer solchen Unterteilung lassen sich patientenzentriert Informationen aus unterschiedlichen Dokumenten auffinden und zusammenfügen, was gerade bei der Medikation von grosser Bedeutung ist. Eine solche Gliederung von Textinhalten kann zudem zur Erstellung von strukturierten eHealth Austauschformaten beitragen und somit einen zusätzlichen Schritt in der digitalen Transformation unterstützen.

Quellenverzeichnis

Literatur

[1] David M. Blei, Andrew Y. Ng und Michael I. Jordan. „Latent Dirichlet Allocation“. In: *Journal of Machine Learning Research* 3 (Jan. 2003), S. 993–1022 (siehe S. 8).

[2] Tianqi Chen und Carlos Guestrin. „XGBoost: A Scalable Tree Boosting System“. In: KDD ’16 (2016), S. 785–794. URL: http://doi.acm.org/10.1145/2939672.2939785 (siehe S. 8).

[3] Sima Djalali. „Wer eHealth sucht, findet einen Haufen Papier“. In: *saez.2015.03985* (Okt. 2015). URL: https://doi.emh.ch/10.4414/saez.2015.03985 (siehe S. 1).

[4] Lukas Golder u. a. *Swiss eHealth Barometer 2017: Akteure im Gesundheitswesen*. URL: https://www.e-healthforum.ch/index.php?apid=503931 (siehe S. 1).

[5] Thorsten Joachims. „Text Categorization with Suport Vector Machines: Learning with Many Relevant Features“. In: *Proceedings of the 10th European Conference on Machine Learning*. ECML ’98. London, UK, UK: Springer-Verlag, 1998, S. 137–142. URL: http://dl.acm.org/citation.cfm?id=645326.649721 (siehe S. 7).

[6] Alexandre Lacoste, Francois Laviolette und Mario Marchand. „Bayesian Comparison of Machine Learning Algorithms on Single and Multiple Datasets“. In: Proceedings of Machine Learning Research 22 (Apr. 2012). Hrsg. von Neil D. Lawrence und Mark Girolami, S. 665–675. URL: http://proceedings.mlr.press/v22/lacoste12.html (siehe S. 10).

[7] Ioannis Partalas u. a. „LSHTC: A Benchmark for Large-Scale Text Classification“. In: *arXiv:1503.08581 [cs]* (März 2015). arXiv: 1503.08581 [cs] (siehe S. 6).

[8] Antti Puurula, Jesse Read und Albert Bifet. „Kaggle LSHTC4 Winning Solution“. In: *arXiv:1405.0546 [cs]* (Mai 2014). arXiv: 1405.0546 [cs] (siehe S. 6).

[9] Juan Ramos. *Using TF-IDF to Determine Word Relevance in Document Queries*. Jan. 2003 (siehe S. 6).

Online-Quellen

[10] *A Kaggle Master Explains Gradient Boosting.* URL: http://blog.kaggle.com/2017/01/23/a-kaggle-master-explains-gradient-boosting/ (besucht am 27. 12. 2017) (siehe S. 8).

[11] *Benchmarking Random Forest Implementations.* URL: http://datascience.la/benchmarking-random-forest-implementations/ (besucht am 27. 12. 2017) (siehe S. 8).

[12] *CrowdFlower DataScience Report 2016.* URL: http://visit.crowdflower.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf (besucht am 28. 12. 2017) (siehe S. 14).

[13] *Dato Winners' Interview: 1st place, Mad Professors.* URL: http://blog.kaggle.com/2015/12/03/dato-winners-interview-1st-place-mad-professors/ (besucht am 27. 12. 2017) (siehe S. 8).

[14] Shuai. *Topic Modeling and t-SNE Visualization.* URL: https://shuaiw.github.io/2016/12/22/topic-modeling-and-tsne-visualzation.html (besucht am 12. 12. 2017) (siehe S. 20).

[15] *XGBoost: Machine Learning Challenge Winning Solutions.* URL: https://github.com/dmlc/xgboost/tree/master/demo#machine-learning-challenge-winning-solutions (besucht am 27. 12. 2017) (siehe S. 8).

Selbständigkeitserklärung

Mit der Abgabe dieser Abschlussarbeit versichert der/die Studierende, dass er/sie die Arbeit selbständig und ohne fremde Hilfe verfasst hat (Bei Teamarbeiten gelten die Leistungen der übrigen Teammitglieder nicht als fremde Hilfe).

Der/die unterzeichnende Studierende erklärt, dass alle zitierten Quellen (auch Internetseiten) im Text oder Anhang korrekt nachgewiesen sind, d.h. dass die Abschlussarbeit keine Plagiate enthält, also keine Teile, die teilweise oder vollständig aus einem fremden Text oder einer fremden Arbeit unter Vorgabe der eigenen Urheberschaft bzw. ohne Quellenangabe übernommen worden sind.

Winterthur, 25. Januar 2018

Marcel Canclini