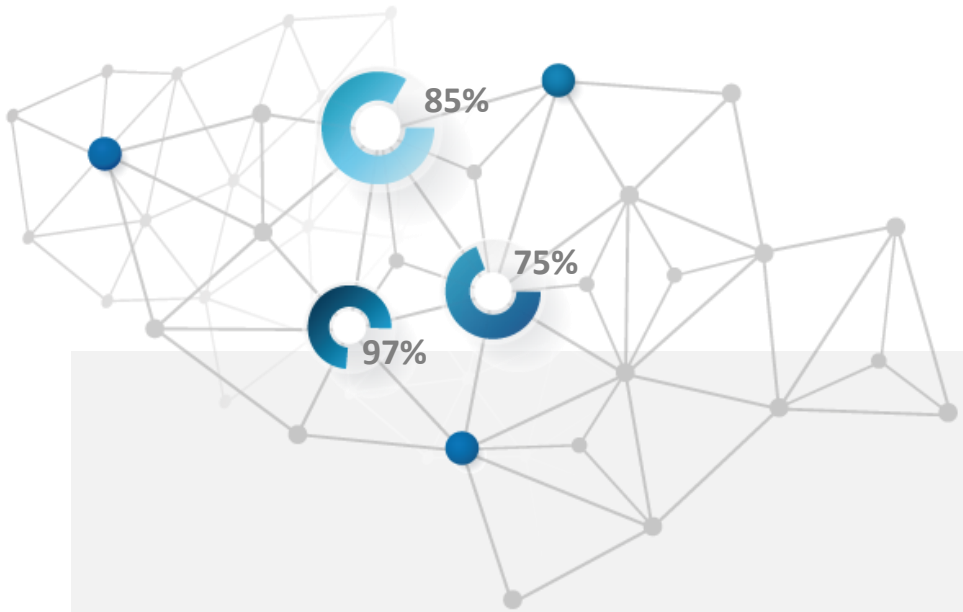
Anonymisierung und De-Identifizierung von medizinischen **Dokumenten** mit dem Averbis DeID-Tool

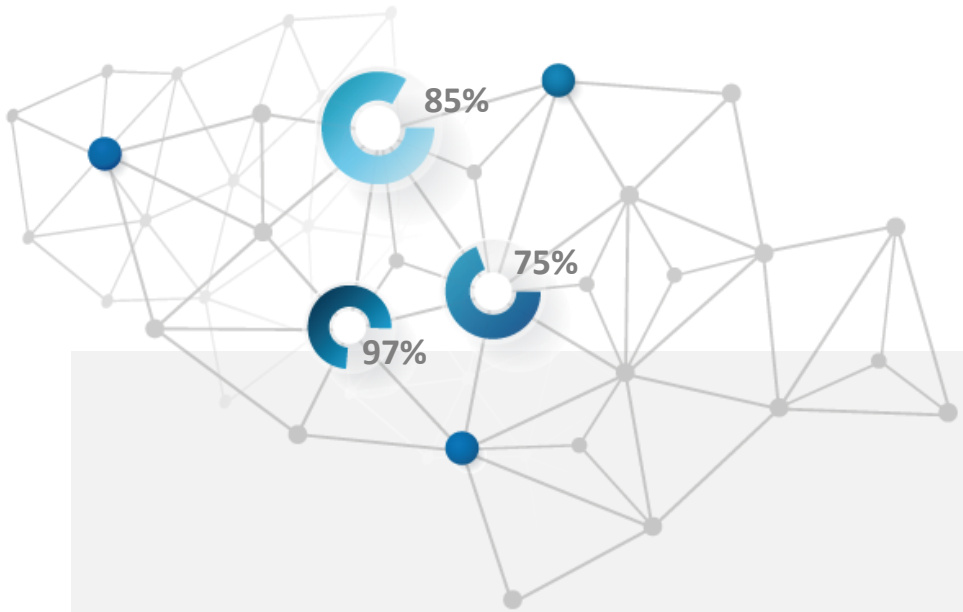


Stefan Schulz

11. Juli 2019

averbis
text analytics

averbis
text analytics



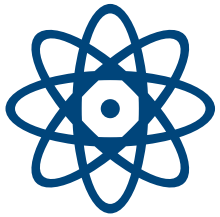
Technologien:	Natural Language Processing, Machine Learning
Industrien:	Healthcare, Pharma, IP
Produkte:	Information Discovery, **Health Discovery**, Patent Monitor
Firmensitz:	Freiburg im Breisgau
Gründungsjahr:	2007

VISION

> **Softwarelösungen, die letztendlich die Unterschiede zwischen strukturierten und unstrukturierten Daten aufheben.**

FIND

Turn text into actionable information



UNDERSTAND

Understand and automate cognitive processes

PREDICT

Enable better business critical decisions

KUNDEN

HEALTHCARE

IQVIA

Krebsregister Baden-Württemberg

TriNetX

RHÖN-KLINIKUM AG

St George's Healthcare NHS
NHS Trust

UNIVERSITÄTS KLINIKUM FREIBURG

PHARMA

Boehringer Ingelheim

Roche

Bayer HealthCare

NOVARTIS

Deloitte.

pwc

PATENTS/IP

CENTREDOC

SERVIER

Europäisches Patentamt
European Patent Office
Office européen des brevets

XPAT

ip-search

Bayer HealthCare

OTHERS

DEUTSCHE NATIONAL BIBLIOTHEK

Thieme

BMW Group

MINI

HRS
Das Hotelportal

caritas

BertelsmannStiftung

WISSENSCHAFTLICHE PROJEKTE

- EU: DEBUG-IT, MANTRA, SEMCARE, euCASES, ASSESS-CT
- BMBF:
  - MI-I: SMITH, MIRACUM, DIFUTURE
  - TOPOS, XplOit
- ZIM: PH^3
- BMWI: Cloud4Health, Klinische Datenintelligenz (KDI)
- THESEUS: RadMining

MEDIZIN INFORMATIK INITIATIVE

miracum
Medical Informatics in Research and Care in University Medicine

SMITH
Smart Medical Information Technology for Healthcare

debugIT

topos+

DIFUTURE
Data Integration for Future Medicine

XplOit
i:DSem

semcare

THESEUS
Neue Technologien für das Internet der Dienste

cloud4health

TECHNOLOGIEN

De-Identifizierung von Dokumenten

TEXT MINING

'A Lung Adenocarcinoma *biomarker panel comprising an microRNA'*

MACHINE LEARNING

Oncology Therapy

- Use of Cabazitraxel in patients with metastatic NSCLC (WO-2015036507-A1)
- Treatment of Melanoma (EP-3021848-A2)

HEALTH discovery

SEMANTIC SEARCH

Lung Cancer

'A Lung Adenocarcinoma *biomarker panel comprising an microRNA'*

TERMINOLOGIES

- Lung Cancer
  - Lung Adenocarcinoma
  - Pulmonal Mass

HEALTH DISCOVERY PLATTFORM

ACTION

DOMAIN SPECIFIC APPLICATIONS

- Diagnosis Support
- Structured Coding
- Predictive Analytics
- Quality Management
- DRG Coding
- Patient Recruitment

INTELLIGENCE

COGNITIVE TECHNOLOGIES

- Terminology Management
- Text Mining
- Semantic Search
- Machine Learning

PLATFORM

FRAMEWORK

- UX
- Web Services
- Data Connectors
- Monitoring
- Scalability
- User Management

TEXT MINING – DISCHARGE PIPELINE

Diagnoses

- ICD10 Codes
- Context

Lab Values

- Parameters
- Values
- Units
- Normalisation
- Interpretation*

Drugs

- Ingredients
- Brand Names
- Strengths
- Dose Forms
- Dose Schemes

Temporal Aspects

- Length of Stay
- Creation Date

AVERBIS DE-ID (DE-IDENTIFIZIERUNG VON COMPUTERLESBAREN TEXTEN)

HALS- NASEN- OHREN- KLINIK
mit Abteilung für Phoniatrie und Pädaudiologie
- Kopf-Halschirurgie, plastische Gesichtschirurgie -

KLINIKUM LUDWIGSHAFEN

Direktor: Prof. Dr. med. K. W. Delank

OPERATIONSBERICHT

Datum: 08.03.2010 BZ: OA Sirra

Patient: Tomm-Bub, Burkhard, geb. 25.12.67

Fallnummer: 0001011138 Station: HO 02 / AK

Aufenthaltsart: Stationär

Diagnose: Klinisch hochgradiger V.a. T2 Tonsillencarcinom li.

Operation: Panendoskopie in ITN

Operateure: OA Scheuermann/Dr. Kerscher

Anästhesie: Dr. Wickenhäuser

OPS: 1-620.1, 1-630.1, 1-610.2, 1-611.0, 1-611.1, 8-701, 1-422 x 4

Zuerst Intubation des Pat. mit dem starren Tracheo Bronchoskop durch mich. Inspektion des Bronchialbaumes mit der 30 °-Optik und der 90 °-Optik, bis findet sich bds. komplett unauffällige Verhältnisse ohne Hinweis auf einen Tumor. Zurückziehen des starren Rohres unter Kontrolle der 30 °-Optik. Die Trachea ist unauffällig. Die Subglottis ist ebenfalls unauffällig. Intubation des Pat. mit dem Woodbridge-Tubus durch mich. Anschließend Durchführen der Hypopharyngoskopie. Hier sind beide Vallecul, die linguale Epiglottisfläche, beide Sinus piriformes, der Pharynxdachpalt und der oberste Anteil des Oesophaguseinganges unauffällig. Einsehen der Glottisebene. Dies gelingt nur teilweise. Daher mit der 30 °-Optik Inspektion der ventralen Anteile des Kehlkopfes. Es liegen insgesamt vollkommen reizlose Verhältnisse vor. Abschließend mit dem starren Bronchoskop Intubation des Oesophagus. Kontrolle des Oesophagus mit der 30 °-Optik. Es finden sich allseits unauffällige Verhältnisse.

Einverzicht des Mc-Ivor-Spatels. Es findet sich ein Tonsillentumor li. Exulceration. Er betrifft die cranialen Anteile der Tonsille mit dem umgebenden weichen Gaumen. Der Tumor reicht bis knapp vor die Grenze harten/weicher Gaumen, nach medial bis an die Mediane heran. Nach caudal überschreitet er die Tonsille selbst nicht. Der gesamte Tumor ist gegenüber dem horizontalen Unterkieferast verschieblich. Der NRR ist bel. Es findet sich keine Exulceration des Tumors im naso-pharyngealen Schleimhautbereich. Es werden Proben aus dem Tumor selbst, auch aus dem Übergang zwischen weichen und harten Gaumen, li. peritonsillär und vom caudalen Anteil des hinteren Gaumenbogens genommen.

Bei positiver Histologie ist die transorale Oropharynxteilresektion indiziert. Aufgrund der Größe des Defektes mit Verlust nahezu der Hälfte des weichen Gaumens ist eine Defektrekonstruktion mit mikrovaskulär anastomosiertem Radialislappen notwendig. Zusätzlich sollten eine funktionelle Neck-diss. bds. und die Tracheotomie durchgeführt werden.

OA Scheuermann

Optical character recognition

Klinischer Befund:

72-jährige Patientin in ausreichendem AZ, Körpergröße 1,57 m, Körpergewicht 72,8 kg. Haut und sichtbare Schleimhäute gut durchblutet. Kein Ikterus, keine Zyanose. Mundhöhle o.B.. Keine Struma, keine peripheren Lymphknotenschwellungen. Über Herz und Lungen war der klinische Befund unauffällig. RR bds. 160/80 mm Hg, Pulsfrequenz regelmäßig. Leber und Milz nicht vergrößert. Nierenlager klopfschmerzfrei. Keine Varizen, keine Oedeme. Fußpulse beiderseits tastbar. MER seitengleich.

Papier → Bild (pdf, jpg) → Text (txt, doc,)

OCR-ARTEFAKTE

Englischer Text

Laserdruck, Scan Graustufen, 8 Bit Farbtiefe

Jude the Obscure is the last of Thomas Hardy's novels, begun as a magazine serial and first published in book form in 1895.

Its hero Jude Fawley is a lower-class young man who dreams of becoming a scholar.

The two other main characters are his earthy wife, Arabella, and his intellectual cousin, Sue. Themes include class, scholarship, religion, marriage, and the modernization of thought and society.

```
Englischer Text

Jude the Obscure is the last of Thomas Hardy's novels, begun as a magazine serial and first
published in book form in 1895.
IIS hero Jude Fawley is a lower-class young man who dreams of becoming a scholar.

The two other main characters are his earthy wife, Arabeiia, and his intellectual cousin,
Sue. Themes include class, scholarship, religion, marriage, and the modernization of
thought and society.
```

Englischer Text

Kopie auf Thermokopierer mit Artefakten

Jude the Obscure is the last of Thomas Hardy's novels, begun as a magazine serial and first published in book form in 1895.

Its hero Jude Fawley is a lower-class young man who dreams of becoming a scholar.

The two other main characters are his earthy wife, Arabella, and his intellectual cousin, Sue. Themes include class, scholarship, religion, marriage, and the modernization of thought and society.

```
En|lischcr Text

V Jude the Obscure is the last of 'ljhomns | begun as a magazine serial and first
published in book [arm in 1895.
IIS ]^i6I`O Jude Fawley `iS` am luwuii·Cl|aS$^|hy<jiing'|nian iviilio dreams ofbccoming a scholar.

The two other main qhdrabiers | wife, Arabeiia, and his intellectual cousin,
Sue. Themes includepiassqvScholarship, religion, marriage, and the modernization of
thought and society.
```

https://www.linux-magazin.de/ausgaben/2006/12/abc-schuetzen/3/

AVERBIS DE-ID (DE-IDENTIFIZIERUNG VON COMPUTERLESBAREN TEXTEN)

averbis deidentification

Erkennt und maskiert identifizierende Inhalte von medizinischen Texten

Grundbegriffe:

- Personenidentifizierende Merkmale (IDAT)
- De-Identifizierung: Identifizierende Inhalte werden erkannt und maskiert:
- Anonymisierung: Inhalte sind nicht mehr den Originaltexten zuzuordnen
- Tagging: automatische oder manuelle Markierung einer Textpassage, z.B. <name>Stefan Schulz</name>
- Pseudonymisierung: Über eine Vertrauensstelle ist eine Re-Identifizierung möglich
  nicht: Ersatz identifizierender Merkmale wie Namen durch "Fantasienamen"

HIPAA „Safe Harbor“ identifiers*

(1) Names
(2) Geographic subdivisions
(3) Dates (except years)
(4) Phone numbers
(5) Vehicle identifiers
(6) Fax numbers
(7) Device identifiers and serial numbers
(8) Mail addresses
(9) URLs
(10) Social security numbers
(11) IP addresses
(12) Medical record numbers
(13) Biometric identifiers
(14) Health plan beneficiary numbers
(15) Photos
(16) Account numbers
(17) Certificate/license numbers
(18) Any other identifying number

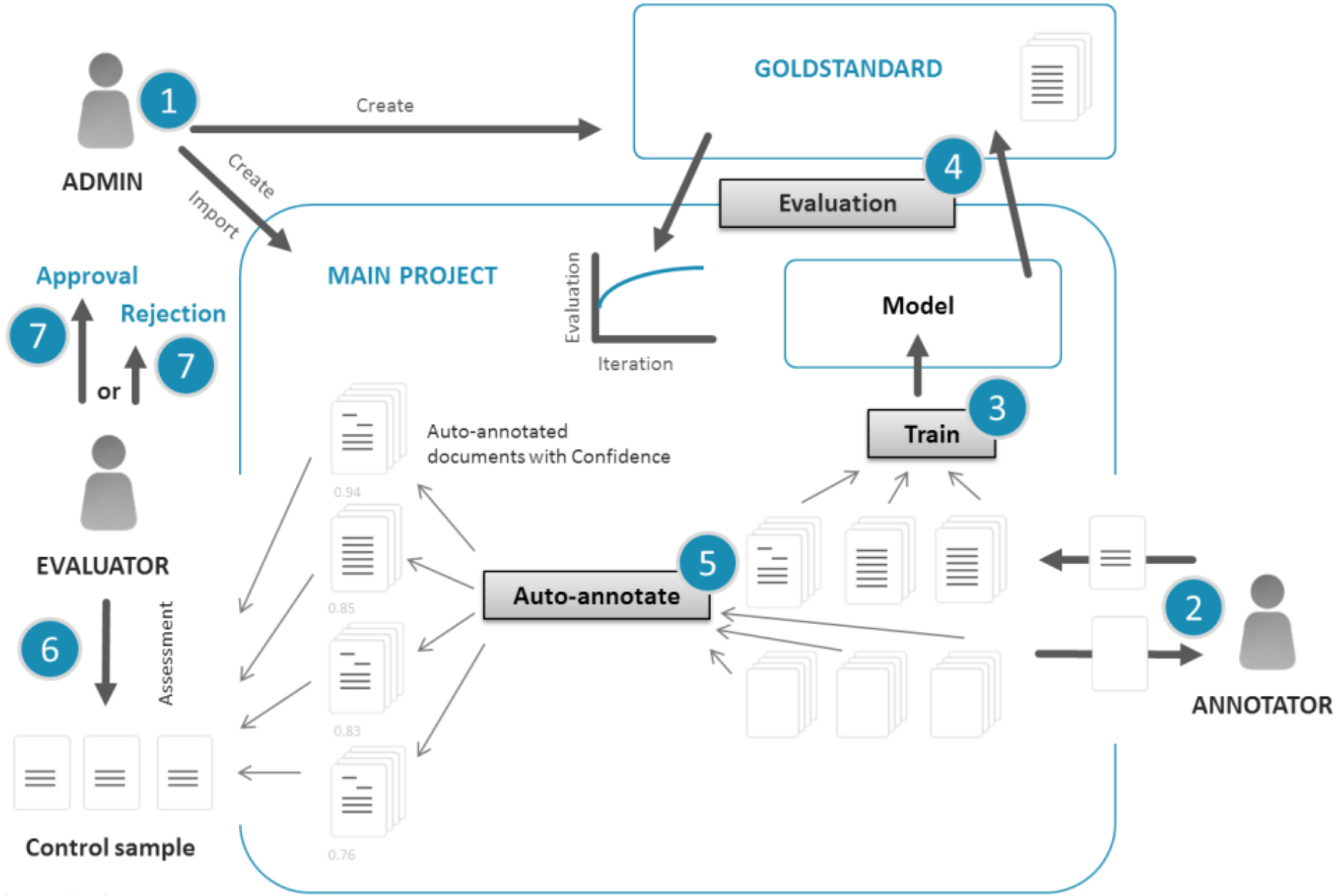
DE-IDENTIFIZIERUNGSPROZESS IN AVERBIS DE-ID

Drei Personengruppen sind beteiligt:

- Administrator: Verwaltung von Benutzerzugängen und Dokumentenkorpora, Import von Doumenten, Export deidentifizierter Versionen
- Annotator: können nur die Dokumente der ihnen zugewiesenen Projekte bearbeiten. Sie annotieren dabei Dokumente und initiieren iterativ Training und die Evaluation des Modells.
- Evaluator: begutachtet die automatisch annotierten, sowie die manuell nachbearbeiteten Dokumente und entscheidet: Freigabe oder Nachbearbeitung.

Ablauf:

- Dokumente werden importiert
- Dokumente werden vom System vorannotiert (CRF Tagger)
- Annotationen werden von Annotatoren überprüft und modifiziert
- Modell wird trainiert, dadurch wird die automatische Annotation verbessert
- Dokumente werden exportiert (anonymisiert oder mit Tags versehen)


GOLDSTANDARD
1
Create
ADMIN
Create
Import
4
Evaluation
MAIN PROJECT
Evaluation
Iteration
Model
Approval
Rejection
7
or
7
3
Train
Auto-annotated
documents with Confidence
0.94
0.85
0.83
0.76
5
Auto-annotate
2
ANNOTATOR
EVALUATOR
6
Assessment
Control sample

DE-IDENTIFIZIERUNG: TECHNIKEN

- Metadata
- Name Lists
- Patterns
- Machine Learning

Export:

Annotations-Tags:

Betrifft Patienten <name>Luise Kisselbach</name>, geb. <date>17.7.1921</date>

Sehr geehrter Kollege <name>Müller</name>

De-identifiziert:

Betrifft Patienten XXXXXXXXX, geb. XXXXXXXXX

Sehr geehrter Kollege <name>XXXXXX</name>

AVERBIS DEID

deID
averbis deidentification

Erkennt und maskiert identifizierende Inhalte von medizinischen Texten

Universitätsklinikum Bad Krozingen Abteilung Neurologie Prof. Dr. Dr. Lutz Schmeisser
Ärztlicher Direktor der Neurologie Telefon 01258/5698245 Fax 01258/5698246

Universitätsklinikum Bad Krozingen
Abteilung: Neurologie
Station 3

Herrn Dr. Schröder Neurologie, Station 3 Universitätsklinikum Bad Krozingen
Nachrichtlich: Herrn Dr. Ludwig Wedel Neurologische Station Rehabilitationskliniken Münstertal Belchenstraße 7 56895 Münstertal

Bad Krozingen, am 28.02.2016

Sehr geehrter Herr Dr. Wedel,

Wir überweisen Ihnen Herrn Jakob Bleifuss, geb. am 17.12.1990, wohnhaft in Veilchengasse 27a, 56895 Heitersheim.

Jakob Bleifuss stürzte am 15.2.16 beim Skifahren abseits der Piste schwer und blieb bewusstlos mit dem Kopf in einer Schneewehe stecken. Seine Begleiter befreiten ihn aus dem Schnee und reanimierten Herrn Bleifuss bis zum Eintreffen der Rettungskräfte.

Aufgrund der unterbrochenen Sauerstoffzufuhr kam es zu einer schweren Schädigung des Gehirns. Herr Bleifuss ist zum jetzigen Zeitpunkt (Stand 27.02.2016) nicht in der Lage zu sprechen und zu schlucken, weshalb am 17.2.16 eine PEG-Sonde zur künstlichen Ernährung gelegt wurde. Außerdem hat er Schwierigkeiten bei der Feinkoordination der Hände und kann nur mit Rollator ein paar Schritte gehen.

Annotations:

RE, LI, ML	Div.	"Universitätsklini..."
RE, LI, ML	Name	"Prof. Dr. Dr. Lut..."
ML	Div.	"Ärztlicher Direktor"
RE, ML	Con.	"01258/5698245"
RE, ML	Con.	"01258/5698246"
RE, LI, ML	Div.	"Universitätsklini..."
RE, ML	Div.	"Station 3 Herrn"
RE, LI, ML	Name	"Dr. Schröder"
ML	Div.	"Neurologie"
RE, LI, ML	Div.	"Station 3 Univers..."
RE, LI, ML	Name	"Dr. Ludwig Wedel"
ML	Div.	"Neurologische"
RE, LI, ML	Div.	"Rehabilitationskl..."
RE, LI, ML	Loc.	"Belchenstraße 7 5..."
LI, ML	Date	"28.02.2016"

Original

Krankenhaus der Samariter Holzhausen

Röntgenabteilung, Vorstand Prim. Univ. Prof. Dr.Dr. Gotthard Vogler

CT Abdomen und kl. Becken

Name: Mustafa Üstün, * 21.06.67

Aufnahmezahl: 1933309807

Abteilung: Chirurgie

Station: A31. OG. Viszeralchirurgie B /

Zi: 119

dikt. Arzt: OA Dr. Huber Karina

WinA. 06/07/2011

Getaggt

<location>Krankenhaus der Samariter</location>
<location>Holzhausen</location>

Röntgenabteilung, Vorstand <name>Prim. Univ. Prof. Dr.Dr. Gotthard Vogler</name>

CT Abdomen und kl. Becken

Name: <name>Mustafa Üstün</name>, * <date>21.06.67</date>

Aufnahmezahl: <id>1933309807</id>

Abteilung: Chirurgie

Station: <division>A31. OG. Viszeralchirurgie B</division> /

Zi: 119

dikt. Arzt: <name>OA Dr. Huber Karina</name>

WinA. <date>06/01/2011</date>

Getaggt

<location>Krankenhaus der Samariter</location>
<location>Holzhausen</location>

Röntgenabteilung, Vorstand <name>Prim. Univ. Prof. Dr.Dr. Gotthard Vogler</name>

CT Abdomen und kl. Becken

Name: <name>Mustafa Üstün</name>, * <date>21.06.67</date>

Aufnahmezahl: <id>1933309807</id>

Abteilung: Chirurgie

Station: <division>A31. OG. Viszeralchirurgie B</division> /

Zi: 119

dikt. Arzt: <name>OA Dr. Huber Karina</name>

WinA. <date>06/01/2011</date>

Anonymisiert

XXXXXXXXXX XXXXXXXXXX

Röntgenabteilung, Vorstand Prim. Univ. Prof. Dr.Dr. XXXXXXXX XXXXXXXXXXXXXX

CT Abdomen und kl. Becken

Name: XXXXXXX XXXXXXXXXX, * X.X.X

Aufnahmezahl: XXXXXXXXXXXX

Abteilung: Chirurgie

Station: XXXXXXXXX XXXXXXXXXXX

Zi: 119

dikt. Arzt: OA Dr. XXXXXXXXX XXXXXXXXX

WinA. XX/XX/XXXX

Getaggt

<location>Krankenhaus der Samariter</location>
<location>Holzhausen</location>

Röntgenabteilung, Vorstand <name>Prim. Univ. Prof. Dr.Dr. Gotthard Vogler</name>

CT Abdomen und kl. Becken

Name: <name>Mustafa Üstün</name>, * <date>21.06.67</date>

Aufnahmezahl: <id>1933309807</id>

Abteilung: Chirurgie

Station: <division>A31. OG. Viszeralchirurgie B</division> /

Zi: 119

dikt. Arzt: <name>OA Dr. Huber Karina</name>

WinA. <date>06/01/2011</date>

Anonym und Verfremdet (separater Prozess)

Kantonsspital Friedrichshafen

Röntgenabteilung, Vorstand Prim. Univ. Prof. Dr.Dr. Gerhard Voigtländer

CT Abdomen und kl. Becken

Name: Manuel Überreuter, * 1.07.69

Aufnahmezahl: 9983209971

Abteilung: Chirurgie

Station: Station Sauerbruch

Zi: 119

dikt. Arzt: OA Dr. Heilmann Kristina

WinA. 16/07/2013

GÜTE DER AUTOMATISCHEN VERFAHREN

- Text-Mining ist kein 100%-Verfahren
- Abhängig von den Spezifika des Dokumentenkorpus
- Güte der automatischen Verfahren i.A. mit manueller De-Identifizierung vergleichbar
- Studie Radiologie Erlangen :
  - Training verbessert kontinuierlich die De-Identifizierungsperformance
  - Ab etwa 100 korrigierten Texten wurde die Detektion als zuverlässig angesehen

AUSBLICK AVERBIS DE-ID FÜR NAKO ENDPOINTVALIDIERUNG

Derzeit Neuimplementierung als Komponente von Averbis Health Discovery

- 1. Release Ende des Jahres

Nutzerwünsche können mit eingearbeitet werden

- Features
- Batch-Processing
- Post-Processing

KONTAKT

Stefan Schulz, Averbis GmbH: stefan.schulz@averbis.com