
**FOM Hochschule für Oekonomie & Management**

Hochschulzentrum Münster

# Hausarbeit

im Studiengang Big Data & Business Analytics

im Rahmen der Lehrveranstaltung

**Analyse semi- & unstrukturierter Daten**

über das Thema

## CAPTUM

- **Characterisation of Type IIb autoimmune chronic spontaneous urticaria markers** -

von

Fiete Ostkamp, Tim Lapstich und Artur Gergert

| | |
|---|---|
| Betreuer : | Prof. Dr. Rüdiger Buchkrämer |
| Matrikelnummern : | 557851, , 562394 |
| Abgabedatum : | 24. August 2021 |

# Inhaltsverzeichnis

# Abbildungsverzeichnis

# Tabellenverzeichnis

# Abkürzungsverzeichnis

| | |
|---|---|
| **DL** | Deep Learning |
| **FN** | Falsch-Negativ |
| **FP** | Falsch-Positiv |
| **IR** | Information Retrieval |
| **KDT** | “Knowledge Discovery in Textual Databases“ |
| **MeSH** | Medical Subject Heading |
| **ML** | Machine Learning |
| **NER** | Named Entity Recognition |
| **NLP** | Natural Language Processing |
| **POS** | Part-of-Speech |
| **TP** | Richtig-Positiv |

# Symbolverzeichnis

# 1 Einleitung

## 1.1 Zielsetzung

Die chronische spontane Urtikaria gehört zu der Gruppe chronischer Urtikaria Erkrankungen. Gekennzeichnet ist diese durch das Wiederauftreten von Quaddeln und/oder Angioödemen über einen Zeithorizont von mehr als sechs Wochen[1] Die geschätzte weltweite Prävalenz chronischer Urtikaria Erkrankten beträgt schätzungsweise 1%. Es liegt lediglich eine geschätzte Prävalenz vor, da es Schwierigkeiten bei der Klassifizierung, der Identifizierung sowie der Diagnose der Erkrankung gibt. Dies ist vor allem auf erhebliche Verzögerungen bei der Diagnose sowie unzureichende Kenntnisse über die chronische Urtikaria zurückzuführen[2]

Die oben angeführte Problematiken wurden zum Anlass genommen ein Projekt zu initiieren, welches den Auftrag verfolgt einen Beitrag zur bekämpfung der chronischen spontanen Urticaria Krankheit zu leisten. Begleitet wird das Projekt von Ärzten und Spezialisten der Charité in Berlin.

Die vorliegende Hausarbeit behandelt das Teilprojekt „Information Retrieval". Ziel dieses Teilprojektes war es aus einem Text-Corpus mit insgesamt über 500 medizinische Fachartikel automatisiert Informationen aus den Texten zu extrahieren, um das Wissen über die Krankheit, erfolgreiche Behandlungsmöglichkeiten etc. zu erweitern.

## 1.2 Aufbau der Arbeit

Zunächst betrachten wir die theoretischen Hintergründe des Natural Language Processing (NLP) und der Named Entity Recognition (NER) um die Grundlage für den zweiten Teil der Arbeit zu schaffen. Dabei gehen wir in Kapitel 2.1 auf die Vorverarbeitung des Textkorpus ein, um dann im Kapitel 2.2 auf die Typen und Aufgaben von Named Entity Recognition (NER) einzugehen.

Im Praxisteil der Arbeit (3) gehen wir dann auf die Erstellung eines eigenen Modells zur Named Entity Recognition (NER) ein. Dabei beschreiben wir den Prozess der Annotierung

[1] **savic.2020**.
[2] **savic.2020**.

(3.1.6 bis 3.1.4) und des Trainings um dann eine Evaluierung des Annotierungsprozesses (3.1.5) und des Trainings (3.1.7) vorzunehmen.

Im abschließenden Teil (4) ziehen wir ein Fazit über das Projekt und fassen die wesentlichen Erkenntnisse der Erstellung des domänenspezifischen NER-Modells und der Analyse von Beziehungen zwischen Entitäten zusammen. Weiter blicken wir auf mögliche Verbesserungspotentiale des gewählten Vorgehens und darüber hinaus zusätzliche Analyseaspekte die von uns nicht berücksichtigt wurden.

# 2 Theoretische Grundlagen der Named Entity Recognition

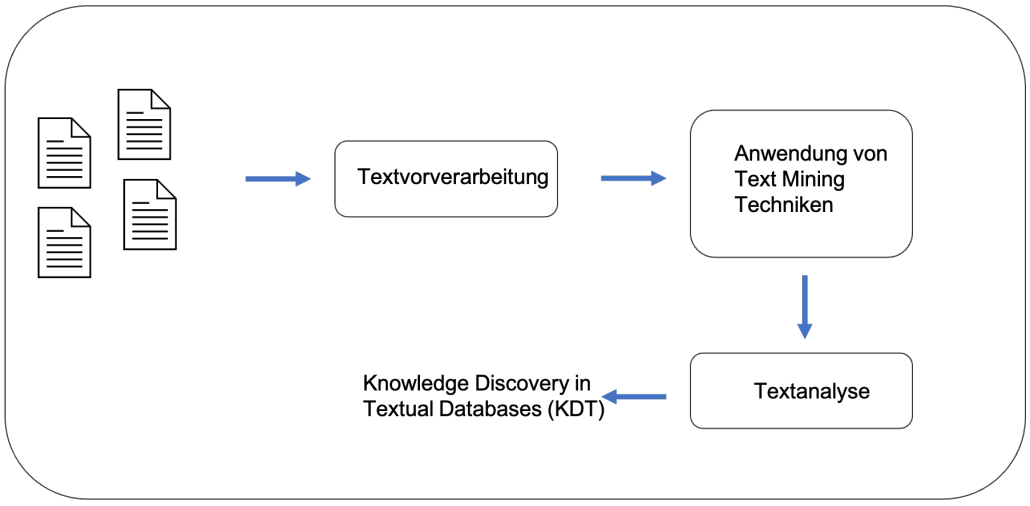
## 2.1 Text Preprocessing

Text Mining bezeichnet den Prozess, um wesentliche bekannte aber auch unbekannte Informationen aus Textdaten zu generieren[3] Die Verarbeitung von unstrukturierten Textdaten wird auch als “Knowledge Discovery in Textual Databases“ (KDT) bezeichnet und spielt eine signifikate Rolle in Anwendungsgebieten wie

- Information Retrieval
- Information Extraction
- Natural Language Processing[4]

Im Wesentlichen geht es in allen o. g. Anwendungsgebieten um die Wissen durch das Mining der Texte zu generieren.

**Abbildung 1: Text Mining Prozess**




*Quelle: Eigene Darstellung in Anlehnung an* **mohan.2015**

Wie der Abbildung 1 entnommen werden kann, stellt die Vorverarbeitung von Volltextdaten bei nahe zu jeder Aufgabe im NLP einen essentiellen und kritischen Schritt dar, da

[3] **mohan.2015**.
[4] **mohan.2015**.

hierbei die fundamentale Basis für die Weiterverarbeitung sowie die Entwicklung der Modelle geschaffen wird.[5] Der Begriff der Textvorverarbeitung umfasst dabei die Anwendung unterschiedlicher Techniken/Methoden, bei denen die Textdokumente für die eigentlichen Zielsetzungen vorbereitet werden. Gängige Techniken für die Vorbereitung der Texte für die nachgelagerten Analysen können folgendermaßen aufgeteilt werden:[6]

- Inhalte Extrahieren und Bereinigen
- Annotationen
- Normalisieren

Zu Beginn der Volltextanalysen stehen häufig die Rohfassungen der Texte zur Verfügung. Diese gilt es im ersten Schritt technisch einzulesen. Hierbei werden auch Daten mit eingelesen, dessen Informationsgehalt gering ist. Beispielshaft zu nennen sind hier HTML tags, Werbung, etc beim Auslesen einer Website[7] oder Grafiken, ASCII-Codes in PDF-Dateien. Demnach ist das Ziel bei dem **Extrahieren und Bereinigen der Inhalte** die Rohdaten soweit zu säubern, bis sich schließlich die reinen Texte als Resultat ergeben. Nachdem die Texte um die technischen Störfaktoren bereinigt wurden, ist die **Tokenization** eine typische Technik der Textextraktion. In dem Prozess der Tokenisation wird der gesamte zu analysierende Text in einzelne Wörter, Phrasen, Symbole, etc. geteilt. Hierbei wird das Ziel verfolgt die Bedeutung einzelner Wörter innerhalb eines Satzes zu analysieren. Die Tokens dienen nämlich als Eingabewerte für viele weitergehende Prozessschritte.[8]
In jedem Text befinden sich Wörter, die wenig Informationsgehalt bei der Textanalyse bieten. Solche Wörter werden auch als **Stop Words** bezeichnet. Beispiele für solche Stop Words sind Artikel oder Präpositionen wie „der“, „die“, „das“, „ein“, „in“, „mit“, etc. Im Analyseprozess stellt jedes unterschiedliche Wort eine eigene Dimension dar. Durch die Entfernung der Stop Words wird somit die Dimensionshöhe reduziert bei gleichzeitiger Beibehaltung des Informationsgehaltes des jeweiligen Satzes/Textes.[9]
Neben der klassischen Methode, die Stop Words auf Basis einer vordefinierten Liste zu entfernen, sind diverse mathematische und nicht-mathematische Methoden entwickelt worden, um Stop Words in Texten zu identifizieren und zu bereinigen.[10]

[5] **gurusamy.2014**.
[6] **pahwa.2018**.
[7] **pahwa.2018**.
[8] **gurusamy.2014**.
[9] **mohan.2015**.
[10] Detailiertere Informationen zu unterschiedlichen Methoden für die Entfernung von Stop Words können**mohan.2015** entnommen werden.

Die Annotationen eines Textes sollen die Funktion des jeweiligen Wortes im Kontext des gesamten Satzes identifizieren. Eine gängige Methode ist dabei das so genannte Part-of-Speech (POS)-Tagging. Jener ist ein Prozess, der die Zuweisungen einzelner Wörter zur POS bzw. zu lexikalischen Klassenmarkern wie Nomen, Verben, Adjektiven, usw. vornimmt.[11] Bei der Erkennung der Funktion eines Satzes treten folgende Hauptprobleme auf:

- Mehrdeutige Wörter
- Unbekannte Wörter

Ersteres stellt das wichtigste Problem dar. Es gibt Wörter, für die es mehr als einen Tag geben kann. Dieses Problem wird durch einen Fokus des Wortes im Satzkontext gelöst.[12] Andersrum existieren Wörter, die zwar denselben Tag haben, jedoch unterschiedliche Bedeutungen im Satzkontext einnehmen.[13] Eine Lösung ist hierbei die Betrachtung des einzelnen Wortes, statt dem Kontext. Das menschliche Auge kann solch eine Differenzierung schnell vornehmen, während es für eine Maschine mühselige Arbeit und einen Lernprozess darstellt. Wie zuvor angeführt, ist es für die automatische POS-Erkennung notwendig ein Modell zu trainieren. Hier wurden mit der Zeit diverse Methodiken entwickelt, die sich auf oberster Ebene in überwachte und unüberwachte Methoden aufteilen. Bei den überwachten Ansätzen wird das POS-Modell auf Basis eines Datensatzes trainiert, bei dem die POS-Werte bekannt sind, während das Modell bei den unüberwachten Ansätzen die POS-Werte selbst induziert, da dem Trainings-Datensatz keine bekannten Werte vorliegen. [14]

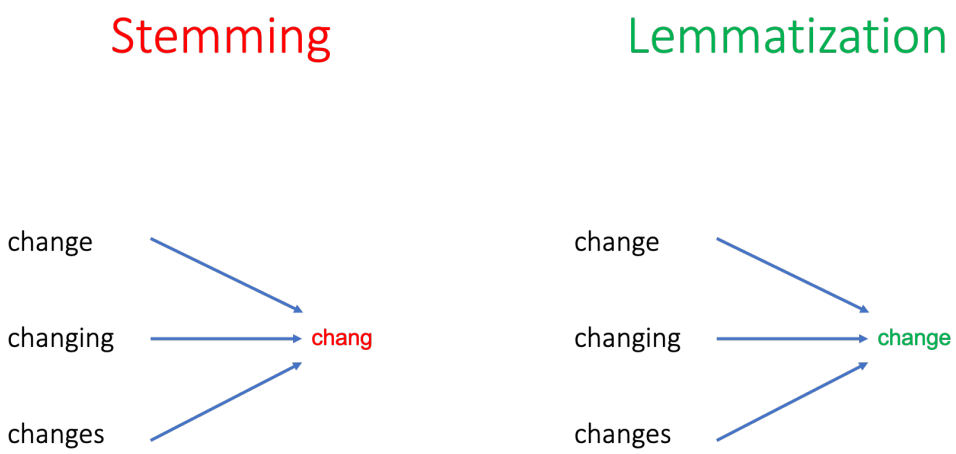
Bei dem **Normalisieren** von Texten wird das Ziel verfolgt ähnliche Wörter zu vereinheitlichen bzw. diese auf einen Standard zu bringen. Dieser Prozess soll vor allem die Dimensionen reduzieren, um die Berechnungen zu vereinfachen und gleichzeitig die Effizienz durch die Standardiesierung der Wörter erhöhen.
Bei der Normalisierung von Wörtern wird häufig auf die Techniken des **Stemming** oder der **Lemmatization** zurückgegriffen. Das Stemming ist ein Prozess, der zugrundeliegende Wörter auf den Wortstamm herunterbricht.[15] Dieser Wortstamm ist im Ergebnis häufig kein echtes Wort, sondern oftmals eine Buchstabenkombination bzw. ein Präfix, den viele Wörter gemeinsam haben.[16] Es existiert eine Vielzahl an Stemming-Algorithmen, dessen Performance vom jeweiligen Einsatzbereich abhängt, sodass noch kein Standard etabliert

[11] **kumawat.2015**.
[12] **gurleenkaursidhu.2013**.
[13] **gurleenkaursidhu.2013**.
[14] Eine detailierte Übersicht über vorhandene POS- Methoden sind in **kumawat.2015** zu finden.
[15] **khyani.2021**.
[16] **khyani.2021**.

ist.[17]

Die Idee bei der Lemmatization ist das so genannte “Lemma“ oder auch die Vokabularform eines Wortes zu identifizieren.[18] Der Prozess ist ähnlich dem des Stemmings, jedoch mit dem Unterschied im Ausgabewert. Während beim Stemming der Ausgabewert der Wortstamm ist und oftmals kein echtes Wort, ist der Ausgabewert bei der Lemmatization das Grundwort aus beispielsweise einem Wörterbuch.[19] Die Abbilung 2 verdeutlicht die Unterschiede der Lemmatization und des Stemmings anhand des englischen Wortes “change“ bzw. dessen Abwandlungen.

**Abbildung 2: Stemming vs. Lemmatization**




*Quelle: Eigene Darstellung in Anlehnung an* **khyani.2021**

Ein vorab durchgeführtes POS-Tagging kann die Ergebnisse der Lemmatization optimieren. Wenn das Wort „schloss“ nämlich in einem Text als Nomen durch das POS-Tagging erkannt wird, dann handelt es sich dabei je nach Kontext entweder um ein Schloss zum verriegeln oder um ein Schloss als Gebäude. Wird es dagegen als Verb identifiziert, so wird mit einer hohen Wahrscheinlichkeit zum Lemma „schließen“umgewandelt.

[17] Eine ausführliche Analyse unterschiedlicher Stemming-Algorithmen kann **jivani.2011** entnommen werden.
[18] **khyani.2021**.
[19] **khyani.2021**.

## 2.2 Named Entity Recognition

Mithilfe der NER wird das Ziel verfolgt automatisiert Eigennamen in Texten zu identifizieren, dessen semantische Typen wie beispielsweise Personen, Ort, Organisationen vordefiniert wurden.[20] Die NER kann nicht nur ausschließlich für die Extraktion von Informationen aus Texten genutzt werden, viel mehr spielt sie eine wesentliche Rolle in einer Vielzahl von Anwendungen aus dem Gebiet des NLP wie beispielsweise dem Textverständnis, dem Information Retrieval (IR), automatisierter Textzusammenfassungen und Übersetzungen, Fragenbeantwortungen, etc. In der Forschung existiert eine Vielzahl an Definitionen für die zu erkennenden Eigennamen, die hauptsächlich in folgende zwei Kategorien aufgeteilt werden können:

- Generische (z. B. Personen und Ort)
- Domänenspezifische (z. B. Proteine, Enzyme und Gene)[21]

Bei den in der NER angewandten Techniken wird zwischen

- Regelbasierte Ansätze
- Unüberwachte Ansätze
- Merkmalsbasierte überwachte Lernansätze
- Deep-Learning Ansätze

unterschieden, wobei die drei erstgenannten den traditionellen Ansätzen zugehörig sind.[22] Regelbasierte Systeme beruhen auf manuell erstellten Regeln. Die zugrundeliegenden Regeln können hierbei z. B. aus domänenspezifischen Ortsverzeichnisen oder syntaktisch-lexikalischen Mustern abgeleitet worden sein.
Das Clustering ist ein typischer Ansatz für unüberwachte NER-Systeme.[23] Auf Basis von Kontextähnlichkeiten werden geclusterte Gruppen generiert aus denen schilieβlich die Entitäten extrahiert werden. Die Idee bei dieser Technik ist mithilfe eines großen Corpus lexikalische Muster und Statistiken zu berechnen, um daraus auf im Text benannte Entitäten schließen zu können.[24]
Im Teilbereich der überwachten Ansätze ist NER hauptsächlich eine Klassifikationsaufgabe. Ausgehend von annotierten Datensätzen werden Merkmale entwickelt, um jedes

[20] **nadeau.2007**.
[21] **li.2018**.
[22] **li.2018**.
[23] **nadeau.2007**.
[24] **li.2018**.

Trainingsbeispiel zu repräsentieren.[25] Für die Entwicklung der Modelle kommen dann Algorithmen des Machine Learning (ML) zu Einsatz, um aus den gegeben annotierten Datensätzen Vorhersagemodelle für noch ungesehene Daten zu erlernen.[26] Essentiell in überwachten NER-Systemen ist die Entwicklung der Merkmale. Merkmalsvektoren abstrahieren dabei den Text, bei der ein Wort durch einen oder mehrere boolische, numerische oder nominale Werte dargestellt wird.[27]
Neben den eben erläuterten traditionellen Methoden für die NER wurden in den letzten Jahren Ansätze im Bereich des Deep Learning (DL) entwickelt, welche sich bewährt haben und Spitzenergebnisse erzielen.[28] Der Einsatz von DL hat wesentliche Vorteile gegenüber den traditionellen Methoden. Zum einen ist es durch die besondere Architektur und den Verarbeitungsmöglichkeiten im Bereich des DL über mehrschichtige künstliche neuronale Netze möglich nicht-lineare Zusammenhänge zu erkennen und zu lernen und zum anderen erleichtern DL-basierte Modelle durch ihre Automation und Selbstständigkeit beim Lernen die Arbeit.[29]

Die NER umfasst zwei Teilaufgaben: Typenidentifikation und Grenzerkennung Die Bewertung eines entwickelten NER-Systems wird in der Regel durch den Vergleich mit den mennschlich getätigten Annotationen vorgenommen. Der Vergleich kann entweder über eine exakte oder durch eine partielle Übereinstimmungsevaluation vorgnommen werden.[30] Bei der exakten Übereinstimmungsevaluation wird geprüft, ob das System sowohl den richtigen Typen als auch die Grenzen korrekt identifiziert.[31] Bei der partiellen Übereinstimmungsevaluation genügt es, wenn, unabhängig von den Grenzen, eine korrekte Identifikation des Typen vom System vorgenommen wird, solange es eine Überschneidung zwischen den ermittelten Grenzen und den wahren Grenzen gibt. Bewertungskennzahlen für die Qualität des NER-Systems sind Precision, Recall und der F-Score. Um diese zu ermitteln, wird vorab die Anzahl der Falsch-Positiv (FP), Falsch-Negativ (FN) und Richtig-Positiv (TP) Zuordnungen ermittelt.

- FP: Eine Entität wurde vom System erkannt, welche in Wirklichkeit keine ist
- FN: Eine Entität wurde nicht vom System erkannt, welche in Wirklichkeit jedoch eine darstellt
- TP: Eine Entität wurde vom System korrekt erkannt

[25] **li.2018**.
[26] **li.2018**.
[27] **nadeau.2007**.
[28] **li.2018**.
[29] **li.2018**.
[30] **li.2018**.
[31] **tjongkimsang.2003**.

Die Precision stellt sich dar als

$$Precision = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})} \tag{1}$$

und kann als Verhältnis von richtig erkannten Entitäten zur Gesamtheit an identifizierten Entitäten interpretiert werden.

Der Recall

$$Recall = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \tag{2}$$

bezieht die richtig erkannten Entitäten auf die Gesamtheit aller möglich gewesenen Entitäten.

Der ausgeglichene F-Score vereint die Precision und den Recall zu einem harmonischen Mittel:

$$F - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3}$$

### 2.2.1 Disease Named Entity Recognition (DNER)

32

### 2.2.2 Chemical Disease Relations (CDR)s

33

### 2.2.3 Bestehende Datensätze für medizinische Named Entity Recognition

### 2.2.4 Etablierte Modelle im Bereich der biologischen Named Entity Recognition

-

## 2.3 Named Entity Linking

[32] vgl. *Li, J.* et al., 2016, S.2.
[33] vgl. ebd., S.2.

# 3 Praxis

## 3.1 Erstellung eines domänenspezifischen NER Modells

Wie im Kapitel 2.2 erläutert, lässt sich zwischen allgemeinen und domänenspezifischen NER-Modellen unterscheiden.[34] Domänenspezifische Modelle sind dabei solche, die für einen dedizierten Themenbereich erstellt wurden und daher in diesem Kontext besonders gute Ergebnisse erzielen. Es gibt bereits spezialisierte Modelle im Bereich der Medizin die mit einem medizinischen Korpus trainiert wurden. Dazu zählen zum Beispiel BioBERT, ScispaCy oder Y[35]. Diese beinhalten bereits zahlreiche medizinische Konzepte und Entitäten, allerdings mussten wir feststellen, dass auch sie für unseren Anwendungsfall *out of the box* nicht alle erforderlichen Entitäten der hier betrachteten Domäne erkennen. Daher haben wir uns entschlossen ein eigenes Modell für das Themengebiet der Urtikaria-Forschung auf Basis eines sciSpacy Modells zu erstellen.

### 3.1.1 Labeling von Trainingsdaten

Das Annotieren von Trainingsdaten hat einen zentralen Anteil bei der Erstellung eines eigenen Modells. Akkurate Daten sind essentiell für die Genauigkeit des resultierenden Modells. Neben der Menge der zum Training zur Verfügung stehenden Daten ist es unterlässlich, dass diese widerspruchsfrei sind.

### 3.1.2 Auswahl der Labels

Die Auswahl der Labels muss für den Anwendungsfall angemessen sein. Je nachdem was die Zielstellung des Projektes ist, können unterschiedliche Labels erforderlich sein, um die notwendigen Zusammenhänge abzubilden. In der vorliegenden Arbeit wurden 5 Entitäten ausgewählt. Diese sind Disease, Treatment, Biomarker, Diagnostic und Person. Die hier getroffene Auswahl basiert zum einen auf in der einschlägigen Literatur gewählten Labels und zum anderem auf den im Rahmen des CAPTUM Projektes umrissenen Entitäten.[36,37]

[34] vgl. *Nouvel*, *D.*, 2016, S.47.
[35] *Li*, *J.* et al., 2020, S.12.
[36] vgl. *Li*, *J.* et al., 2016, S.
[37] vgl. *Eickhoff*, *C.*, *Kim*, *Y.*, *White*, *R. W.*, 2020, S.

### 3.1.3 Richtlinien zur Annotation

Um möglichst homogene, gleichmäßige Annotationen über mehrere Personen hinweg zu erhalten ist es sinnvoll Richtlinien für die Annotation festzulegen. Dies ist häufig ein iterativer Prozess, da zu Beginn einer solchen Aufgabe nicht immer vorab klar ist, an welchen Stellen Unklarheit besteht.[Citation]

Auch wir haben Richtlinien aufgestellt, die sich zum einen mit den von uns gewählten Entitäten selber und zum anderen mit der Art und Weise der Annotation beschäftigen. Die begrifflichen Definitionen sind vor allem vor dem Hintergrund wichtig, dass die beteiligten Personen keinen medizinischen Hintergrund haben und auf diese Art und Weise zunächst ein gemeinsames Grundverständnis davon geschaffen werden soll, was zum Beispiel einen *Biomarker* oder *Diagnostik* ausmacht. An diesem Beispiel wird auch deutlich, dass Entitäten nicht immer trennscharf sind. Biomarker wird demnach definiert als „ein messbarer Indikator für das Vorhandensein oder die Schwere eines Krankheitszustands oder eines anderen physiologischen Zustands eines Organismus.“[38], medizinische Diagnose als „process of determining which disease or condition explains a person's symptoms and signs.“ Diagnose wird als Prozess des Nachweises einer Krankheit verstanden und Biomarker als Gegenstand des Nachweises. In diesem Falle ist nicht immer einfach zu unterscheiden ob es im vorliegenden Satz um das Vorhandensein eines Stoffes/eines Markers nun im diagnostischem Sinne zu verstehen ist, oder als Marker gekennzeichnet werden sollte.[39]

### 3.1.4 Verwendete Software

Für die Annotation der Entitäten wurde Doccano verwendet.[40] Dabei handelt es sich um ein open source Tool zur Text Klassifikation, Sequenzlabeling und Sequenz zu Sequenz Aufgaben. Zur Unterstützung der Annotierung und um den Aufwand zu reduzieren, wurde auf das Auto Labeling Feature der Software zurückgegriffen. Dabei kann eine Backendanwendung konfiguriert werden, die auf Basis eines trainierten Modells Vorschläge zur Annotation unterbreitet. Für diese Funktionalität wurden von uns zunächst die ersten 100 Textabschnitte ohne Backend annotiert und die gewonnenen Daten dann für das initiale Modell verwendet.[41]

[38] **wiki2021**.
[39] vgl. *Neves*, *M.*, 2014.
[40] vgl. *Hiroki Nakayama*, 2021, S.
[41] vgl. *Neves*, *M.*, *Leser*, *U.*, 2014.

### 3.1.5 Analyse zum Inter-Annotator Agreement

Das Inter-annotator agreement (IAA) ist ein Maß der Übereinstimmung von Annotationen die von mehreren Personen getätigt wurden. Von dem Score lassen sich allgemein Rückschlüsse ziehen, wie zuverlässig der Annotierungsprozess ablief.[42] Der Grundgedanke dabei ist, dass ein hoher Score für die Reproduzierbarkeit der Ergebnisse spricht und Ausdruck ist von der Klarheit der Richtlinien.

- Cohens K[43]
- Fleiss K[44]

### 3.1.6 Labeling von Trainingsdaten

Das Labeling der Daten wurde von zwei Personen ohne medizinischen Hintergrund ausgeführt. Um dennoch eine möglichst hohe Qualität der Annotationen zu erhalten, haben die teilnehmenden Personen bei Unklarheit im Internet recherchiert. Wie im Abschnitt **??** beschrieben ist, wurde dadurch dennoch eine verhältnismäßig gute Genauigkeit erzielt. Insgesamt wurden dadurch X Textabschnitte annotiert, die zufällig aus dem Gesamtcorpus von 465 einzigartigen Texten ausgewählt wurden. Dies entspricht einem Anteil von circa 0.01%, relativ gesehen zu der Gesamtzahl von 57764 Abschnitten.

- Notwendige Menge an Annotationen
- Aufstellung der Label (Ontologie?)
- Richtlininien zur Annotation
- Auswahl der Software (Doccano, Spacy)

### 3.1.7 Training des Modells

Für das Training des Modells werden die annotierten Daten aus den vorherigen Schritten verwendet. Für das Training des Modells wurde die Natural Language Processing Bibliothek spaCy[45] verwendet. Diese wurde im Hinblick auf ihre umfangreichen Funktionalitäten und benutzerfreundlichkeit ausgewählt.

[42] vgl. *Ide*, *N.*, *Pustejovsky*, *J.*, 2017, S.298.
[43] vgl. *Cohen*, *J.*, 1960, S.
[44] vgl. *Fleiss*, *J. L.*, 1971, S.
[45] *Honnibal*, *M.*, *Montani*, *I.*, 2017.

Das Training erfolgt über den *spacy train* Befehl mit Trainingsparametern die über ein config file definiert werden. Die Parameter für das Training des Modells wurden von uns nicht verändert und orientieren sich somit an den empfohlenen Werten.[46]

Für einige NLP tasks wie die *text classification* können unterschiedliche Architekturen gewählt werden, für die Named Entity Recognition ist nur der *Transition Based Parser* verfügbar.

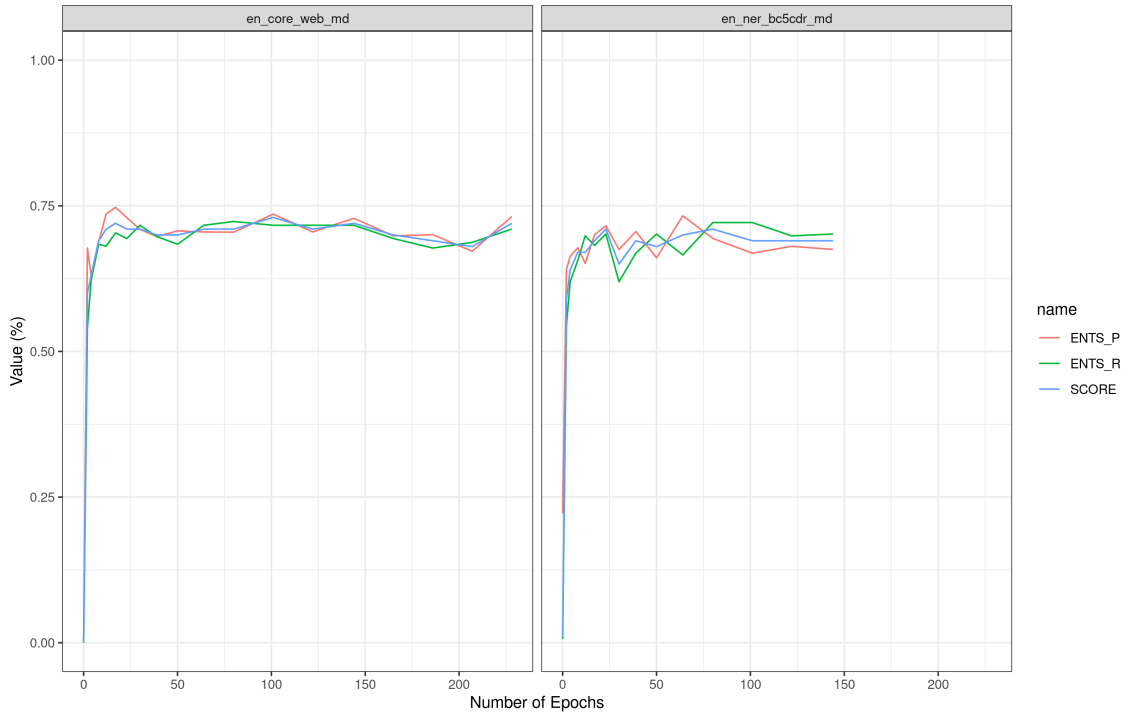
- Warum SpaCy
- Modellarchitektur
- Trainingsparameter

Bei der zugrundeliegenden Architektur handelt es sich um einen sogenannten Transition Based Parser. Diesem liegt das Konzept von Übergängen zwischen Wörtern zugrunde. Auf Basis der Annahme, dass der Part-of-Speech tag eines Wortes abhängig ist von bisherigen Wörtern (dem sogenannten *state*) eines Satzes, lassen sich Wahrscheinlichkeiten für Übergänge vom letzten hin zum nächsten Tag aufstellen.[47] Diese werden in eine sogenannte *transition matrix* überführt, die die Wahrscheinlichkeiten enthält für die Nachfolge eines Tags auf einen anderen.

Wie im Kapitel 2.2 beschrieben, werden zur Evaluierung eines Modells der f-Score verwendet, der aus XY berechnet wird. Dieser setzt sich aus Precision und Recall, also dem Anteil der XY zusammen.

[46] *Ostkamp*, *F.*, 2021, vgl.
[47] *Honnibal*, *M.*, 2013.

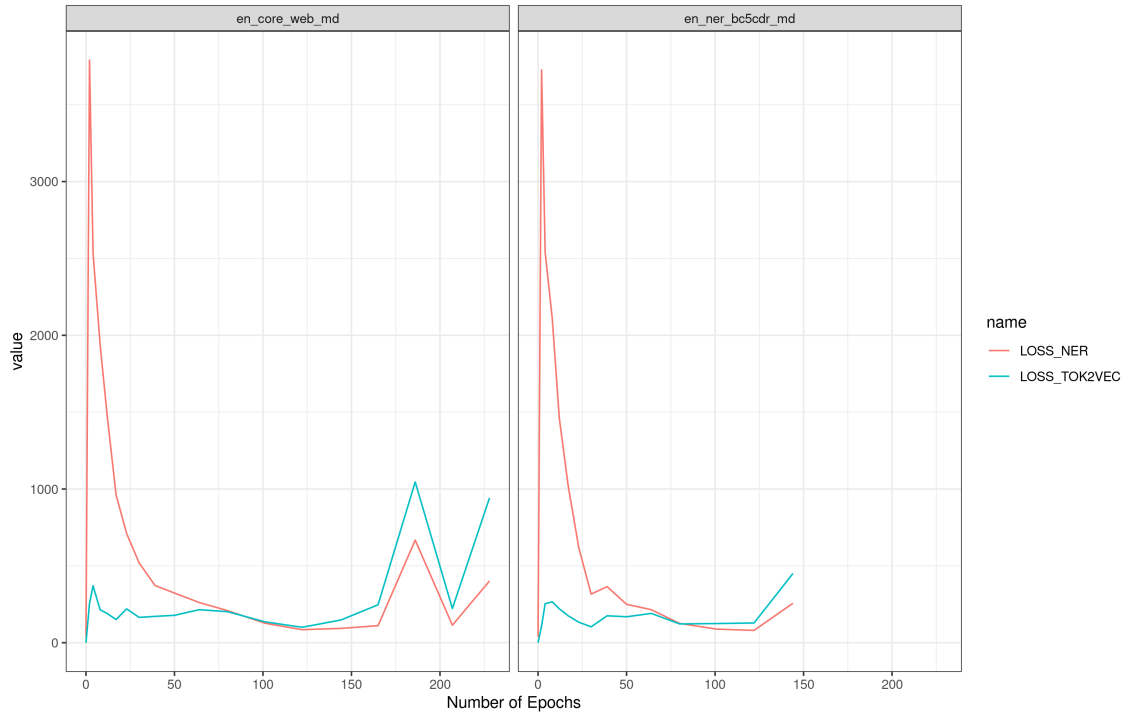
**Abbildung 3: Precision, Recall und Score**




In Abbildung 3 sind Precision, Recall und der f-Score des Trainings für zwei Basismodelle dargestellt. Diese sind das spaCy-eigene „en_core_web_md“ und das im Rahmen der BioCreative 5 Challenge erstellte „en_ner_bc5cdr_md“. Für beide Modelle gilt, dass sich das Verhältnis von Precision und Recall über mehrere Trainingsepochen hinweg nicht wesentlich verbessert beziehungsweise stagnativ bei circa 70% einpendelt. Die Wahl des Modells macht unter dieser Betrachtung folglich nicht wesentlich einen Unterschied. Dies deutet darauf hin, dass das Modell beim Training nicht wesentlich konvergiert, das heißt in XY übergeht.

Die Wahl des Basismodells für das weitere Annotieren sowie nachfolgende Schritte fiel dennoch auf das scispacy Modell, da dieses insgesamt mit mehr Daten trainiert wurde. Daher gehen wir davon aus, dass dieses Modell besser generalisieren wird, das heißt bei unbekannten Daten bessere Klassifikationen treffen wird.

**Abbildung 4: Verluste (*losses*) bei NER und TOK2VEC**




48

Was sind Verluste??

Zum weiteren Verständnis des Modells und der Vorhersagen, ist es sinnvoll, dieses feiner zu analysieren. So kann die Genauigkeit des Modells weiter für jede Entität aufgeschlüsselt werden.

48 *Tsai*, *R. T.-H.* et al., 2006.

**Abbildung 5: F-Score je Entität**

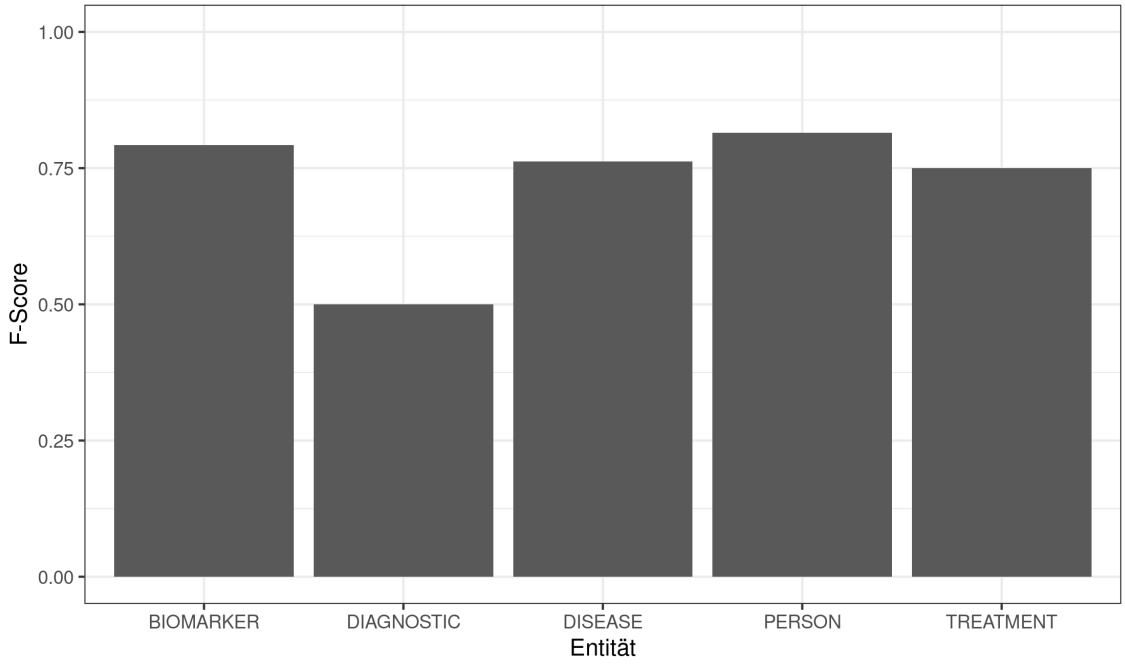



Abbildung 5 stellt den F-Score je Entität dar. Es ist erkennbar, dass die Entitäten Biomarker, Disease, Person und Treatment einen sehr homogenen Score haben, Diagnostic jedoch deutlich darunter liegt. Ursächlich hierfür könnte die relativ geringe Anzahl an Trainingsdaten für das Diagnostic Label sein. Insgesamt liegt die Anzahl bei X, währenddessen bei den anderen Labels jeweils mindestens Y Textsamples vorhanden sind. Zur weiteren Verbesserung des F-Scores und somit des Modells, sollten folglich mehr Annotationen mit dem Diagnostic Label gesammelt werden.

# 4 Fazit

In der vorliegenden Arbeit wurde ein eigenes NER-Modell erstellt und zur Analyse eines medizinischen Korpus genutzt. Dazu wurden [X] Textabschnitte annotiert und zur Named Entity Recognition (NER) und Relationship Extraction genutzt.

An dem Projekt waren keine Personen mit fachlichem - das heißt medizinischem - Hintergrund geteiligt. Auch wenn die Verlässlichkeit der Annotationen durch Metriken wie Cohen's Kappa gestützt werden, lässt sich dadurch nicht direkt die Qualität der Daten ableiten.

Im Laufe des Projektes ist deutlich geworden, dass es bereits geeignetere, domänenspezifische Software zur Annotation im medizinischen Kontext gibt. Doccano verfügt über eine

benutzerfreundliche Oberfläche und erlaubt das Labeling von Entitäten, ermöglicht jedoch keine Verknüpfung der Labels mit medizinischen Datenbanken beziehungsweise Medical Subject Heading (MeSH) Codes. In diesem Zusammenhang wurde für andere Datensätze in dieser Domäne, wie dem *NCBI-Disease* oder *BC5CDR*, PubTator[49] verwendet.[50,51]

[49] *Wei*, *C.-H.*, *Kao*, *H.-Y.*, *Lu*, *Z.*, 2013.
[50] vgl. *Doğan*, *R. I.*, *Leaman*, *R.*, *Lu*, *Z.*, 2014, S.9.
[51] vgl. *Li*, *J.* et al., 2016, S.4.

# Anhang

# Literaturverzeichnis

*Cohen*, *Jacob* (1960): A Coefficient of Agreement for Nominal Scales, in: Educational and Psychological Measurement, 20 (1960), Nr. 1, S. 37–46, [Zugriff: 2021-08-12]

*Doğan*, *Rezarta Islamaj*, *Leaman*, *Robert*, *Lu*, *Zhiyong* (2014): NCBI Disease Corpus: A Resource for Disease Name Recognition and Concept Normalization, in: Journal of Biomedical Informatics, 47 (2014), S. 1–10, [Zugriff: 2021-08-10]

*Eickhoff*, *Carsten*, *Kim*, *Yubin*, *White*, *Ryen W.* (2020): Healthcare NER Models Using Language Model Pretraining, in: Proceedings of the 13th International Conference on Web Search and Data Mining, WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston TX USA: ACM, 2020-01-20, S. 901–902, [Zugriff: 2021-08-16]

*Fleiss*, *Joseph L.* (1971): Measuring Nominal Scale Agreement among Many Raters. In: Psychological Bulletin, 76 (1971), Nr. 5, S. 378–382, [Zugriff: 2021-08-12]

*Hiroki Nakayama* (2021): Doccano: : Text Annotation Tool for Human, o. O.: doccano, 2021-08-11, URL: https://github.com/doccano/doccano [Zugriff: 2021-08-11]

*Honnibal*, *Matthew*, *Montani*, *Ines* (2017): spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing, o. O., 2017

*Ide*, *Nancy*, *Pustejovsky*, *James* (Hrsg.) (2017): Handbook of Linguistic Annotation, Dordrecht: Springer Netherlands, 2017, [Zugriff: 2021-08-12]

*Li*, *Jiao*, *Sun*, *Yueping*, *Johnson*, *Robin J.*, *Sciaky*, *Daniela*, *Wei*, *Chih-Hsuan*, *Leaman*, *Robert*, *Davis*, *Allan Peter*, *Mattingly*, *Carolyn J.*, *Wiegers*, *Thomas C.*, *Lu*, *Zhiyong* (2016): BioCreative V CDR Task Corpus: A Resource for Chemical Disease Relation Extraction, in: Database, 2016 (2016), baw068, [Zugriff: 2021-08-10]

*Neves*, *M.*, *Leser*, *U.* (2014): A Survey on Annotation Tools for the Biomedical Literature, in: Briefings in Bioinformatics, 15 (2014), Nr. 2, S. 327–340, [Zugriff: 2021-08-14]

*Neves*, *Mariana* (2014): An Analysis on the Entity Annotations in Biological Corpora, in: F1000Research, 3 (2014), S. 96, [Zugriff: 2021-08-10]

*Nouvel*, *Damien* (2016): Named Entities for Computational Linguistics, in: (2016), S. 187

*Ostkamp*, *Fiete* (2021): Doccano Annotation Server with Spacy Backend, o. O., 2021-08-10, URL: https://github.com/FieteO/doccano_spacy [Zugriff: 2021-08-18]

*Tsai*, *Richard Tzong-Han*, *Wu*, *Shih-Hung*, *Chou*, *Wen-Chi*, *Lin*, *Yu-Chun*, *He*, *Ding*, *Hsiang*, *Jieh*, *Sung*, *Ting-Yi*, *Hsu*, *Wen-Lian* (2006): Various Criteria in the Evaluation of Biomedical Named Entity Recognition, in: BMC Bioinformatics, 7 (2006), Nr. 1, S. 92, [Zugriff: 2021-08-14]

*Wei*, *Chih-Hsuan*, *Kao*, *Hung-Yu*, *Lu*, *Zhiyong* (2013): PubTator: A Web-Based Text Mining Tool for Assisting Biocuration, in: Nucleic Acids Research, 41 (2013), Nr. W1, W518–W522, [Zugriff: 2021-08-13]

## Internetquellen

*Honnibal*, *Matthew* (2013): A Good Part-of-Speech Tagger in about 200 Lines of Python · Explosion, <https://explosion.ai/blog/part-of-speech-pos-tagger-in-python> (2013-09-18) [Zugriff: 2021-08-17]

*Li*, *Jing*, *Sun*, *Aixin*, *Han*, *Jianglei*, *Li*, *Chenliang* (2020): A Survey on Deep Learning for Named Entity Recognition, arXiv: 1812.09449 [cs], <http://arxiv.org/abs/1812.09449> (2020-03-18) [Zugriff: 2021-08-08]

## Ehrenwörtliche Erklärung

Hiermit versichere ich, dass die vorliegende Arbeit von mir selbstständig und ohne unerlaubte Hilfe angefertigt worden ist, insbesondere dass ich alle Stellen, die wörtlich oder annähernd wörtlich aus Veröffentlichungen entnommen sind, durch Zitate als solche gekennzeichnet habe. Ich versichere auch, dass die von mir eingereichte schriftliche Version mit der digitalen Version übereinstimmt. Weiterhin erkläre ich, dass die Arbeit in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde/Prüfungsstelle vorgelegen hat. Ich erkläre mich damit einverstanden/nicht einverstanden, dass die Arbeit der Öffentlichkeit zugänglich gemacht wird. Ich erkläre mich damit einverstanden, dass die Digitalversion dieser Arbeit zwecks Plagiatsprüfung auf die Server externer Anbieter hochgeladen werden darf. Die Plagiatsprüfung stellt keine Zurverfügungstellung für die Öffentlichkeit dar.

Münster, 24.8.2021

(Ort, Datum)

(Eigenhändige Unterschrift)