Universität Heidelberg
Institut für Informatik
Arbeitsgruppe Datenbanksysteme

Bachelor-Arbeit

Kontextbasierte Informationsextraktion aus Datenschutzerklärungen

Name: Björn Ternes
Matrikelnummer: 4015840
Betreuer: Prof. Dr. Michael Gertz
Datum der Abgabe: 31.03.2021

Ich versichere, dass ich diese Bachelor-Arbeit selbstständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe und die Grundsätze und Empfehlungen "Verantwortung in der Wissenschaft" der Universität Heidelberg beachtet wurden.

Abgabedatum: 31.03.2021

Zusammenfassung

Datenschutzerklärungen enthalten essenzielle Informationen über die Verarbeitung der Nutzerdaten. Gleichzeitig sind sie sehr umfassend und in juristischer Fachsprache geschrieben, sodass Laien nicht alle Zusammenhänge verstehen und deshalb eine Datenschutzerklärung nicht lesen. Aus diesem Grund thematisieren diverse Arbeiten die automatische Aufarbeitung von Datenschutzerklärungen in verständlicher Sprache. Ziel dieser Arbeit ist die Untersuchung der Informationsextraktion in Abhängigkeit von der Textgliederung. Die Funktionalität dieser Vorgehensweise wird durch Textanreicherung Abschnitts spezifischer Informationen demonstriert. Nach Erarbeitung eines theoretischen Modells der Datenschutzerklärung und der Textbausteine wurde ein Datensatz von 100 deutschsprachigen Datenschutzerklärungen annotiert. Die Annotation umfasste vier Abschnittsklassen der Textabschnitte (Datenerhebung, Rechtsgrundlagen, Drittdienste und Verantwortlichkeit) und 12 Abschnitt spezifische Entitäten. Zur Klassifizierung der Abschnitte in Klassen wurde ein regelbasierter Ansatz mit einer Support Vector Machine (SVM) verglichen. Abhängig von der Abschnittsklasse wurde dann eine geeignete Methode zur Informationsextraktion evaluiert. Für die Klassen der Drittdienste und der Datenerhebung konnte mit einem F1-score von jeweils 96,30% und 97,90% ein auf dem Datensatz selbst trainiertes Spacy-NER-Modell erfolgreich evaluiert werden. Für die Klasse der Verantwortlichkeit wurde eine Mischung aus einem Regelbasierten und vortrainierten NER-Modell genutzt, wobei ein F1-score von durchschnittlich 90,05% erzielt wurde. Die Extraktion der Klasse Rechtsgrundlagen erfolgte ebenfalls mithilfe eines NER-Modells, welches einen F1-score von 84,65% erreichte. Für die praktische Umsetzung wurden die im theoretischen Modell beschriebenen Hintergrunddokumente und deren Platzhalter durch extrahierte Informationen korreliert und ersetzt. Dadurch entstanden Zusammenfassungen der Datenschutzerklärung, die aus den eingesetzten extrahierten Informationen bestehen. Schlussfolgernd konnte die These bestätigt werden, dass die Nutzung der Gliederung der Datenschutzerklärung zur gezielten Informationsextraktion beiträgt.

Abstract

Privacy policies contain essential information about the processing of user data. However, they are also very comprehensive and written in legal jargon, so that non-experts do not understand all the connections and may therefore not read them. For this reason, various works focus on the automatic processing of privacy policies in understandable language. The aim of this work is to investigate the extraction of information depending on the structure of the text. The functionality of this approach is demonstrated by text enrichment of section-specific information. After developing a theoretical model of the privacy policy and the text components, a dataset of 100 German-language privacy policies was annotated. The annotation included four section classes of text sections (third-party services, data collection, legal basis and responsibility) and 12 section-specific entities. A rule-based approach was compared with a Support Vector Machine (SVM) to classify the sections into classes. Depending on the section class, a suitable information extraction method was evaluated. For the classes of third-party services and data collection, a Spacy-NER model trained on the dataset was successfully evaluated with an F1-score of 96.30% and 97.90% respectively. For the class of responsibility, a mixture of a rule-based and pre-trained NER model was used, achieving an average F1-score of 90.05%. The legal basis class was also extracted using a NER model, which achieved an F1-score of 84.65%. For the practical implementation, the placeholders within the text components described in the theoretical model were correlated with the extracted information and replaced. This resulted in summaries of the privacy policy consisting of the extracted information used. In conclusion, the thesis could be confirmed that the use of the structure of the privacy policy contributes to the targeted extraction of information.

Inhaltsverzeichnis

1 Einleitung **1**

1.1 Motivation . . . 1
1.2 Ziele der Arbeit . . . 2
1.3 Aufbau der Arbeit . . . 4

2 Grundlagen und verwandte Arbeiten **5**

2.1 Textklassifikation . . . 5
2.1.1 Regelbasierte Klassifikation . . . 6
2.1.2 Lineare Klassifikation . . . 6
2.2 Named Entity Recognition . . . 8
2.3 Klassifikator Evaluation . . . 9
2.4 Verwandte Arbeiten . . . 11

3 Dokumentenmodell und Textklassifikation **13**

3.1 Überblick und Zielsetzung . . . 13
3.2 Dokumente- und Hintergrundmodell . . . 14
3.3 Klassifizierung . . . 16
3.4 Entitäten . . . 18
3.5 Textanreicherung . . . 19

4 Experimentelle Evaluation **22**

4.1 Zusammensetzung des Datensatzes . . . 22
4.2 Erstellung des Datensatzes . . . 23
4.3 Implementierung . . . 29
4.3.1 Aufbereitung . . . 29
4.3.2 Klassifikation . . . 30
4.3.2.1 Regelbasierter Ansatz . . . 32
4.3.2.2 Support Vector Machine (SVM) . . . 34
4.3.3 Extraktion von Informationen . . . 35
4.3.3.1 Datenerhebung . . . 35
4.3.3.2 Rechtsgrundlage . . . 36

4.3.3.3 Drittdienste 36
4.3.3.4 Verantwortlichkeit 36
4.3.4 Textanreicherung 37

5 Zusammenfassung und Ausblick **40**
5.1 Zusammenfassung 40
5.2 Ausblick 41

Abbildungsverzeichnis

1.1 Komprimierte Darstellung der Datenschutzerklärung der Zeit Verlagsgruppe 2
1.2 Drittdienste: Von der Datenschutzerklärung zur Anmerkung 3

2.1 Visuelle Darstellung der Textklassifikation 5
2.2 Visuelle Darstellung einer linearen Klassifikation [27] 7
2.3 Support Vektoren [27] 7
2.4 NER-Pipeline [3] 9
2.5 Wechselbeziehung Precision und Recall [1] 10

3.1 Darstellung der Baumstruktur, die sich durch die hierarchische Struktur ergibt $t, n \in \mathbb{N}$ 15
3.2 Abschnitt - Visuelle Darstellung der Kardinalität 15
3.3 Hintergrund und Abschnitt - Visuelle Darstellung der Kardinalität 16
3.4 Rechtsdokument und Rechtsklasse - Visuelle Darstellung der Kardinalität 16
3.5 Abschnitt und Abschnittsklasse - Visuelle Darstellung der Kardinalität . 17
3.6 grafische Darstellung einer Entität 18
3.7 Hintergrundmodell - Visuelle Darstellung der Kardinalität 20

4.1 Zusammensetzung des Datensatzes nach Betreiber 23
4.2 Ja oder Nein? Entscheidung anhand der Gliederung 24
4.3 Visuelle Darstellung des Datensatzaufbaus 25
4.4 Klassifizierung der Abschnitte mithilfe von Doccano 26
4.5 Markierung der Entitäten mithilfe von Doccano 27
4.6 Annahme (Y), Ablehnung (N) oder Teilweise (P) - Annotierung der Rechtsgrundlagen 27
4.7 Vom Rechtsdokument zur Anmerkung 29
4.8 Gliederung der Datenschutzerklärung der Universität Heidelberg 30
4.9 Zusammensetzung Datensatz nach festgelegten Klassen 31

5.1 Darstellung einer Webseite, welche auf Grundlage der Arbeit erstellt werden könnte 41

Tabellenverzeichnis

3.1 Übersicht über Entitäten 19

4.1 Überblick über annotierte Entitäten 28
4.2 Übersicht über den annotierten Datensatz nach Klasse 31
4.3 Ergebnisse regelbasierter Ansatz zur Klassifizierung von Rechtsgrundlagen 33
4.4 Ergebnisse regelbasierter Ansatz zur Klassifizierung von Drittdiensten . . 33
4.5 Ergebnisse regelbasierter Ansatz zur Klassifizierung von Verantwortlichkeit 34
4.6 Ergebnisse einer SVM zur Klassifizierung von Datenerhebung 34
4.7 Ergebnisse einer SVM zur Klassifizierung von Rechtsgrundlage 34
4.8 Ergebnisse einer SVM zur Klassifizierung von Drittdienste 35
4.9 Ergebnisse einer SVM zur Klassifizierung von Verantwortlichkeit 35
4.10 Ergebnisse Extraktion von Kontaktinformationen 36
4.11 Protoyp: Übersicht gewählter Methoden zur Klassifizierung 37
4.12 Übersicht der verwendeten Methoden 38
4.13 Übersicht Textbausteine 38

Abkürzungsverzeichnis

FQDN Fully Qualified Domain Name

URL Uniform Resource Locator

SVM Support Vector Machine

BLSTM Bidirectional Long short-term memory

NER Named Entity Recognition

LER Legal Entity Recognition

MUC Message Understanding Conference

POS-Tagging Part-of-Speech Tagging

1 Einleitung

1.1 Motivation

Datenschutzerklärungen enthalten essenzielle Informationen über die Verarbeitung der Nutzerdaten. Die Pflicht zur Erstellung einer solchen Erklärung ergibt sich aus den Datenschutzprinzipien, die durch die Datenschutzgrundverordnung[1] Artikel 5 in der gesetzlichen Norm gefestigt sind und vorschreiben, dass ein Nutzer an der ihn betreffenden Datenverarbeitung zu beteiligen ist. Diese Beteiligung kann nur garantiert werden, wenn der Nutzer die Aussagen versteht. Der große Umfang der Datenschutzerklärung und die Verfassung in juristischer Fachsprache, welche Laien nicht vollständig verstehen, stellen dabei zwei Hindernisse dar. Beispielsweise umfasst die Datenschutzerklärung von der Zeit Verlagsgruppe ausgedruckt 40 Seiten[2], diese Erklärung ist komprimiert in Abbildung 1.1 zur Veranschaulichung dargestellt. Folglich ist die Bereitschaft der Nutzer, diese langen Texte ohne Aufbereitung zu lesen, gering [9], [25], [15], [24]. McDonald und Cranor haben gar festgestellt, dass ein durchschnittlicher Internetbenutzer über 8 volle Tage im Jahr benötigen würde, betreffende Datenschutzerklärungen zu lesen [23].

Es gibt mehrere Lösungsansätze für die Aufbereitung [19], [18], [22], [7] und Darstellung [19], [17] einer Datenschutzerklärung zum besseren Verständnis des Lesers. Einen Ansatzpunkt für die Informationsextraktion bildet der strukturelle Aufbau einer Datenschutzerklärung. Wegen der gesetzlichen Anforderungen müssen diese bestimmte Inhalte, beispielsweise die Nennung eines Verantwortlichen abdecken, sodass sie oft nach einem ähnlichen Muster von Juristen umgesetzt werden. In diesem Zusammenhang wurde die erfolgreiche kontextbasierte Extraktion zum Beispiel auf italienischen Gesetzestexten umgesetzt [6], [8]. Die Grundlage einer erfolgreichen Informationsextraktion sind repräsentative, umfangreiche und annotierte Datensätze, die zum Training und der Auswahl von Methoden benutzt werden. Bisher gibt es keinen frei verfügbaren Datensatz, der deutschsprachige Datenschutzerklärungen enthält.

[1] `https://eur-lex.europa.eu/legal-content/DE/TXT/HTML/?uri=CELEX:32016R0679`
[2] `https://www.zeit.de/hilfe/datenschutz`

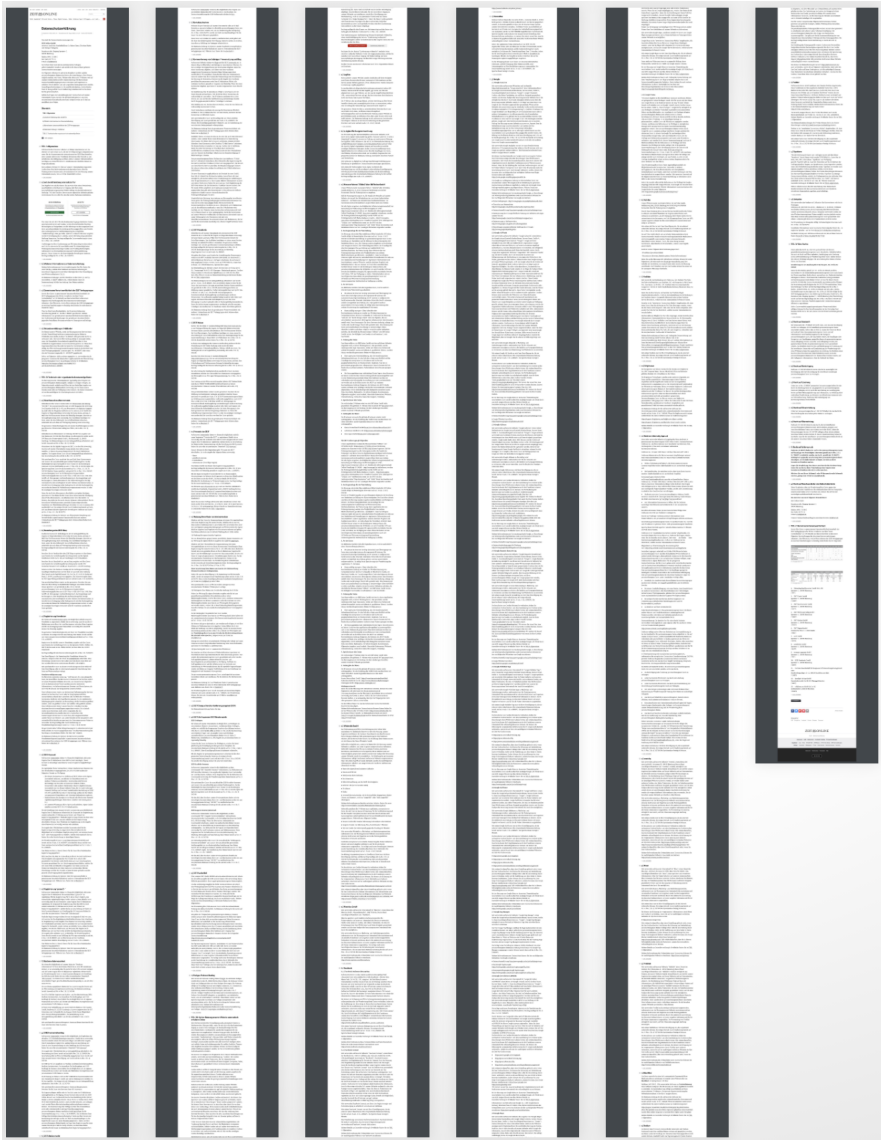


Abbildung 1.1: Komprimierte Darstellung der Datenschutzerklärung der Zeit Verlagsgruppe
Link zur Datenschutzerklärung: `https://www.zeit.de/hilfe/datenschutz`
(abgerufen am 27.03.2021)

1.2 Ziele der Arbeit

Primäres Ziel dieser Arbeit ist die kontextbasierte Informationsextraktion aus Datenschutzerklärungen als Lösungsansatz der zuvor beschriebenen Problematik. Zunächst wird ein abstraktes Modell der Datenschutzerklärung und der Hintergrunddokumente erstellt, um eine generalisierbare Problemlösung für die Datenextraktion und -zuordnung zu formulieren. Anschließend wird ein Datensatz aus 100 deutschsprachigen Datenschutzerklärungen aufgebaut, um das Training der Methoden zu ermöglichen.

Dieser enthält Annotationen von vier Abschnittsklassen (Datenerhebung, Rechtsgrundlagen, Drittdienste und Verantwortlichkeit) und 12 Abschnitts spezifische Entitäten. Die Klassifikation der Abschnitte soll mit einem regelbasierten Ansatz und einer Support Vector Machine (SVM) vergleichend durchgeführt werden, um abschnittsweise eine optimale Methode zu wählen. Danach sollen Methoden zur Extraktion von klassenabhängiger Entitäten evaluiert werden. Zur Demonstration der Funktionalität soll ein Prototyp entwickelt werden, der Textbausteine mit den zugehörigen extrahierten Entitäten aus der Datenschutzerklärung anreichert.

Der Prototyp soll aus semistrukturiertem HTML-Text anhand der verwendete Überschriften-Hierarchie Abschnitte extrahieren und in 4 Kategorien, die sich aus den Abschnittsklassen (Datenerhebung, Rechtsgrundlagen, Drittdienste und Verantwortlichkeit) ergeben, klassifizieren. Darauf folgend soll abhängig der gewählten Kategorie Informationen aus den Abschnitten extrahiert und in vorgeschriebene Textbausteine eingesetzt werden.

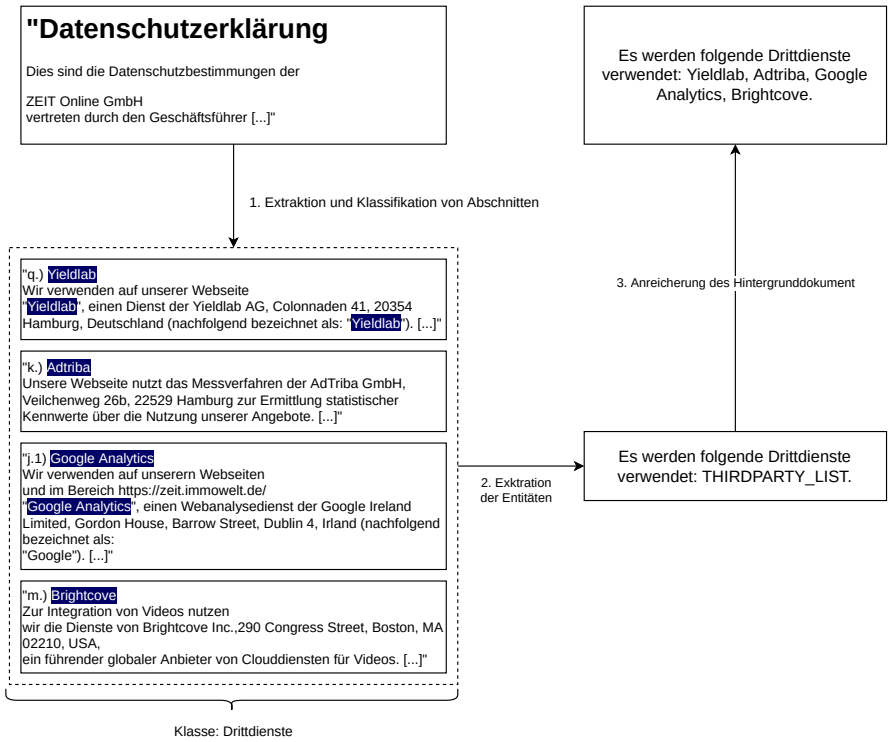


Abbildung 1.2: Drittdienste: Von der Datenschutzerklärung zur Anmerkung
Link zur Datenschutzerklärung: `https://www.zeit.de/hilfe/datenschutz`
(abgerufen am 27.03.2021)

1.3 Aufbau der Arbeit

Zunächst werden in Kapitel 2 grundlegende Aspekte und verwandte Arbeiten eingeführt. Kapitel 3 beschäftigt sich mit der Modellierung der Datenschutzerklärung, der Hintergrunddokumente und Textklassifikation und Kapitel 4 zeigt anschließend die experimentelle Evaluation und die Umsetzung des Prototyps. In Kapitel 5 werden die Ergebnisse dieser Arbeit zusammengefasst und ein Ausblick auf weitere Möglichkeiten gegeben.

2 Grundlagen und verwandte Arbeiten

Nachfolgend werden einige Grundlagen dieser Arbeit eingeführt, bevor in Kapitel 3 eine Modellierung von Datenschutzerklärung, Hintergrunddokumenten und Textklassifikation vorgenommen wird. In Kapitel 2.4 wird auf verwandte Arbeiten eingegangen, die sich mit der Informationsextraktion von Datenschutzerklärungen und der verständlichen Darstellung für den juristischen Laien beschäftigen.

2.1 Textklassifikation

Die Textklassifikation beinhaltet die Zuordnung von Texten in vordefinierten Kategorien. Eine vordefinierte Kategorie entspricht dabei einem Label. Zum Beispiel ist eine gängige Klassifikationsaufgabe die Zuordnung einer E-Mail zu einem der beiden Label (1) Spam oder (2) kein Spam [29], [28], [13], [12], [3, Seite 114]. Im Allgemeinen können k-verschiedene Kategorien $\{1, 2..., k \mid k \in \mathbb{N}\}$ existieren und es gibt keine Ordnung, die auf diesen Kategorien definiert ist [3, Seiten 114-115]. Dies ist zur Verdeutlichung in Abbildung 2.1 dargestellt.

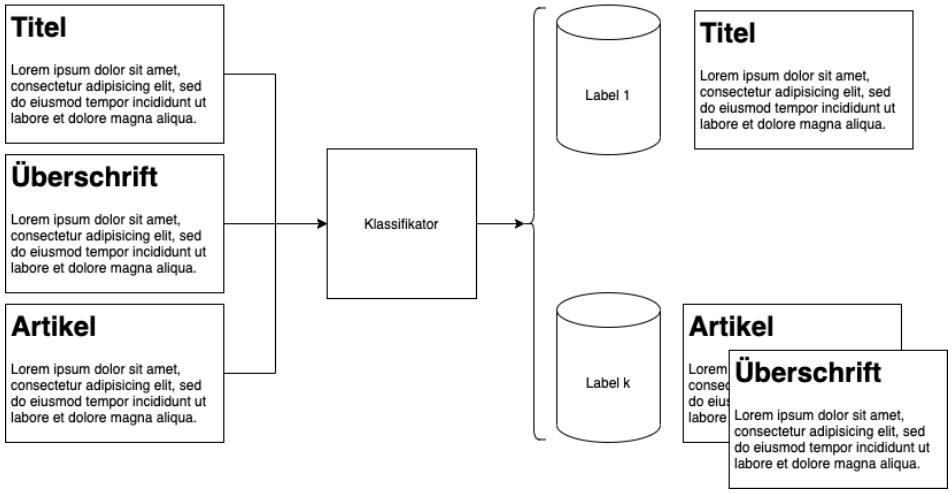


Abbildung 2.1: Visuelle Darstellung der Textklassifikation

Die Klassifikation erfolgt durch den Klassifikator. Dieser erhält Text als Eingabe und ordnet ihm ein passendens Label zu. In Kapitel 2.1.1 und 2.1.2 werden ein regelbasierter Ansatz und ein linearer Ansatz zur Umsetzung eines Klassifikators beschrieben. Unabhängig des Ansatzes wird ein Klassifikator mithilfe von Trainingsdaten trainiert. Aus diesem Grund ist die Textklassifikation überwachtem Lernen zugeordnet [3, Seiten 113]. In der Klassifikation ist der Wertebereich dieser Funktion f() eine diskrete Menge von Werten, beispielsweise $\{Spam, kein\ Spam\}$. Eine Multi-Label Klassifikation besteht, wenn die Menge der Labels größer als 2 ist. Im speziellen Fall der binären Klassifikation kann kann diese Menge so dargestellt werden: $\{0, 1\}$. Gleichzeitig kann die Multi-Label Klassifikation mit einfachen Algorithmen durch mehrfache Anwendung als binäre Klassifikation umgesetzt werden [3, Seite 115].

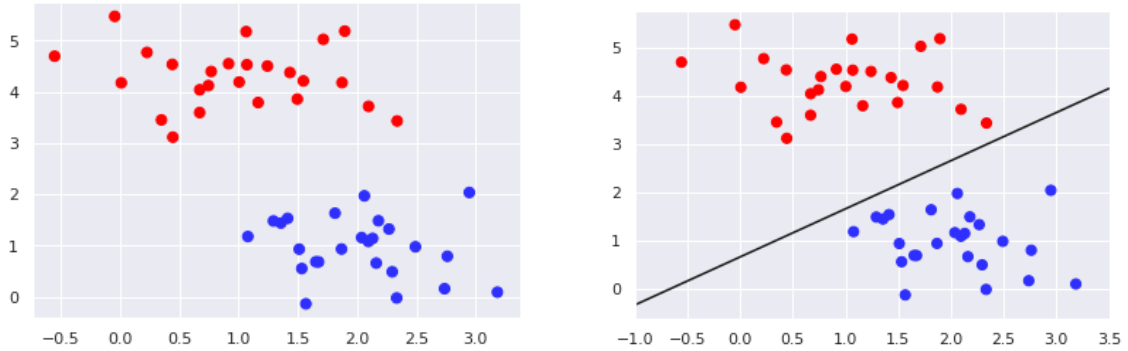
Dies wird an einem Beispiel illustriert: Es gilt, eine Zuordnung von Zeitungsartikeln zu ihrem Ressort {Politik, Kultur oder Sport} herzustellen. Diese Zuordnung lässt sich durch drei binäre Klassifikatoren umsetzen, indem nacheinander ein Zeitungsartikel durch die Entscheidung {Nein: 0, Ja: 1} einem der Ressorts Politik|Kultur|Sport zugeordnet wird.

2.1.1 Regelbasierte Klassifikation

Die regelbasierte Klassifikation erfolgt durch die Aufstellung von Regeln, die sequentiell auf der Texteingabe angewendet werden. Meistens lässt sich eine solche Regel als Vorhandensein einer Teilmenge von Wörtern auf einem Text definieren [3, Seite 147]. Ein Beispiel für eine Regel wäre, dass ein Zeitungsartikel, der das Wort "Dr. Angela Merkel" enthält, der Kategorie Politik zuzuordnen ist.

2.1.2 Lineare Klassifikation

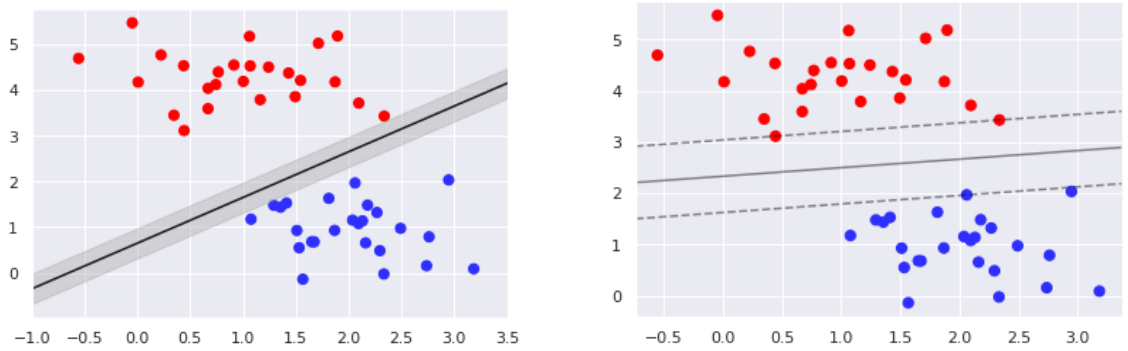
Die lineare Klassifikation funktioniert nach dem Prinzip eine Klassifikation durch eine lineare Funktion durchzuführen. Die lineare Funktion bildet dabei eine Abgrenzung der zu differenzierenden Mengen, nachfolgend Entscheidungsgrenze genannt [3, Seite 159]. Eine vereinfachte Darstellung dieser Vorgehensweise ist in Abbildung 2.2 anhand einer binären Klassifikation mit roten und blauen Datenpunkten gezeigt.



(a) linear klassifizierbarer Datensatz (b) mit eingezeichneter Entscheidungsgrenze

Abbildung 2.2: Visuelle Darstellung einer linearen Klassifikation [27]

Ein Klassifikator, der sich dieses Prinzip zunutze macht, ist die sogenannte Support Vector Machine (SVM). Diese fügt parallel der Entscheidungsgrenze zwei symmetrische Hyperebenen hinzu, sodass die Elemente der zu klassifizierenden Menge entweder rechts oder links von dieser liegen [3, Seiten 177-178].



(a) mit Hyperebenen (b) Support Vektoren

Abbildung 2.3: Support Vektoren [27]

Das Verfahren erhält seinen Namen durch die in Abbildung 2.3 (b) eingezeichneten Support Vektoren, welche Datenpunkte auf der gestrichelten Linie, also der Hyperebene darstellen. Der Erfolg einer SVM lässt sich dadurch begründen, dass für die Einordnung von Datenpunkten nur die Position zu den Support Vektoren wichtig ist. Demnach ändern alle Datenpunkte, die weiter weg von diesen liegen und damit eindeutig einer Kategorie zuzuordnen sind, die Aussage im Laufe des weiteren Trainings nicht. Der technische Grund dafür ist, dass diese Punkte nicht durch die Verlustfunktion beachtet werden, die zum Training des Modells verwendet wird. Folglich spielt ihre Position und Anzahl keine Rolle, solange sie nicht in dem Bereich zwischen den beiden Hyperebenen liegen [27]. Wenn Datenpunkte dazwischen liegen, kann deren Entfernung zur nächstgelegenen

Hyperebene durch die Verwendung der sogenannten Slack-Variable als Fehler der Verlustfunktion quantifiziert werden. Daraus folgt, dass solche "Ausreißer" den Fehler der Verlustfunktion erhöhen, wohingegen die SVM versucht diesen zu minimieren [27].

2.2 Named Entity Recognition

Die Named Entity Recognition beschreibt die Problematik der Extraktion von Eigennamen aus einem Text. Der Text ist eine Folge von Token, wobei die Art und die Beziehung der Token untereinander nicht bekannt sind. Das Ziel der Named Entity Recognition ist die Erkennung von Eigennamen-Token, die mit real-existierenden Konzepten verbunden sind. Ein Beispiel dafür wäre die Erkennung von IBM als Unternehmen oder Hillary Clinton als Person [3, Seite 386].

Die ursprüngliche Definition von Named Entity Recognition wurde erstmalig auf der 6. Message Understanding Conference (MUC) in Columbia, Maryland eingeführt [16]. Die meisten Methoden zur Erkennung benannter Entitäten konzentrieren sich auf die drei Arten von Entitäten, die Person, den Ort und die Organisation. Datum, Uhrzeit, monetäre Werte und Prozente sind auch Bestandteile der Definition, wenngleich sie nicht im klassischen Sinne Eigennamen darstellen [3, Seite 386]. In der juristischen Domäne können zum Beispiel Rechtsprechungen, Aussagen, Eide und Plädoyers als Trainingsgrundlage genutzt werden, um Entitäten, wie beispielsweise die Namen von Gerichten oder Rechtszitate zu erkennen [11], [2], [21], [4].

Named Entity Recognition Algorithmen werden in der Regel im Rahmen des überwachten Lernens trainiert. Dabei besteht der Trainingsdatensatz aus unstrukturiertem Text mit Annotationen aller relevanter Entitäten. Beispielsweise kann eine Annotation von Entitäten wie folgt aussehen:

> <Person> Bill Clinton </Person> wohnt in <Ort> New York </Ort> in einer Nachbarschaft, welche ein paar Meilen von der <Organization> IBM </Organization> und deren Gebäude entfernt ist. Er und seine Frau <Person> Hillary Clinton </Person>, sind nach <Ort> New York </Ort> nach seiner Präsidentschaft zurück gekehrt.

Das Ziel ist, ausgehend von dem Trainingsdatensatz, vergleichbare oder gleiche Entitäten mit unbekannter Position in einem Testdatensatz zu bestimmen.

Die Erkennung von Eigennamen ist ein grundlegendes Problem der Informationsextraktion, weil es den Grundbaustein darstellt, auf dem viele andere Methoden der Informationsextraktion aufbauen. Exemplarisch wäre hier die Relationship Extraction zu nennen [3, Seite 386].

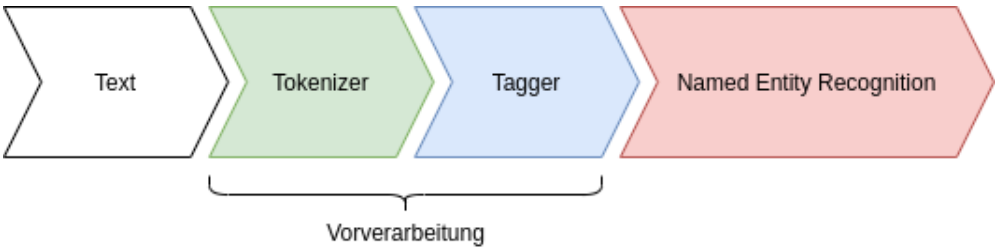


Abbildung 2.4: NER-Pipeline [3]

Der Tokenizer zerlegt den Eingabetext in einzelne Token, die zusammengehörige semantische Einheiten sind. Beispielsweise wäre ein Token "15 Euro".

Der Tagger bildet Part-of-Speech Tagging (POS-Tagging) ab, bei dem Wörter und Satzzeichen zugehörigen Wortarten zugeordnet werden.

2.3 Klassifikator Evaluation

Die Evaluation eines Klassifikators zielt darauf ab, den Erfolg der Methoden auf einem spezifischen Datensatz zu bemessen. Daraus ergibt sich eine Entscheidungsgrundlage, mit der eine geeignete Methode ausgewählt werden kann. Damit ist die Methodenauswahl und -evaluation eng verbunden [3, Seite 241].

Es gilt, mit einer Evaluation die Qualität und den Erfolg eines Algorithmus festzustellen. Es gibt einige Metriken, um den Erfolg einer Implementierung und Methodik zu messen [20]. Im Rahmen dieser Arbeit werden folgende Metriken benutzt: (1) Precision, (2) Recall, (3) Accurancy und (4) F1-score. Diese bilden jeweils einen Wert zwischen 0 und 1 ab. Eine Multiplikation mit 100 ergibt dabei den prozentualen Wert.

Um die Metriken zu berechnen, müssen zunächst alle erkannten Entitäten ("Treffer"), ob richtig oder falsch, in folgende Kategorien eingeteilt werden:

True Positive - TP: Als True Positive werden alle Treffer bezeichnet, die auch erkannt werden sollten.

True Negative - TN: Als True Negative werden alle nicht erkannten Entitäten bezeichnet, die auch nicht erkannt werden sollten.

False Positive - FP: Als False Positive werden alle Treffer bezeichnet, die nicht erkannt werden sollten.

False Negative - FN: Als False Negative werden alle nicht erkannten Entitäten bezeichnet, die erkannt werden sollten.

Anhand dieser Kategorien kann die Leistung einer Methode wie folgt berechnet werden:

$$Precision = \frac{TP}{TP + FP} \tag{2.1}$$

Die Precision gibt den Anteil der korrekten Treffer im Verhältnis zu den gefundenen Treffern an.

$$Recall = \frac{TP}{TP + FN} \tag{2.2}$$

Der Recall gibt den Anteil der korrekten Treffer im Verhältnis zu den Treffern an, die gefunden werden sollen.

Dabei ergibt sich aus Precision und Recall folgende Wechselbeziehung, die in Abbildung 2.5 visuell aufgearbeitet ist.

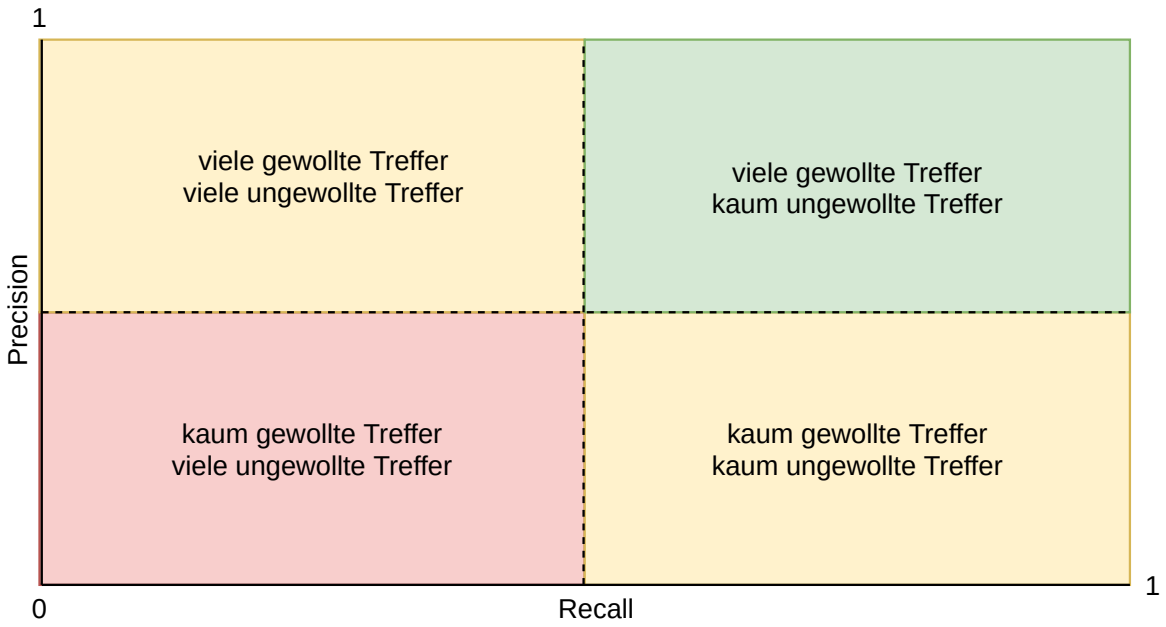


Abbildung 2.5: Wechselbeziehung Precision und Recall [1]

Der Zusammenhang kann durch den F1-score bewertet werden, er kombiniert Precision und Recall [1].

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.3}$$

$$Accurancy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.4}$$

Die Accurancy gibt das Verhältnis von korrekten Treffern zu allen Treffern an.

2.4 Verwandte Arbeiten

Grundsätzlich ist die Domäne der juristischen Texte insbesondere der Datenschutzerklärungen in deutscher Sprache bisher wenig erforscht. Eine umfassende erste Erkundung stellt das Forschungsprojekt von Kettner et. al. dar [19]. Primäres Ziel des Projektes war die informationelle Selbstbestimmung der Nutzer zu verbessern, indem unter anderem Datenschutzerklärungen und AGBs verständlich dargestellt werden. In diesem Rahmen wurden auch die Verbraucherkenntnis und -motivation betreffend des Datenschutzes untersucht. Es wurde ein Nutzerführungskonzept erarbeitet, welches die informationelle Selbstbestimmung von Nutzern fördert. Eine automatisierte Suche nach Datenschutzerklärungen durchsuchte App-Stores und extrahierte diese. Daraus entstand in Zusammenarbeit mit Juristen ein annotierter Datensatz bestehend aus 175 deutschsprachigen Datenschutzerklärungen. Dabei wurden Versionen einer Datenschutzerklärung auf Unterschiede untersucht und mit einer technischen Prüfung gekoppelt, um zu verifizieren, dass die Datenverarbeitung so stattgefunden hat, wie es die Datenschutzerklärung ausführt. Als Ergebnis des Projektes sind drei Prototypen namens "PGuard Browser-Plugin", "DATENSCHUTZscanner App-Client", sowie der "Check-Your-APPS Datenschutzerklärungs-Analyzer" entstanden, die Nutzern den selbstbestimmten Umgang mit persönlichen Daten ermöglichen.

Eine weitere Arbeit auf diesem Gebiet ist das automatische Framework für die Datenschutzerklärungsanalyse namens Polisis [18]. Es ermöglicht skalierbare, dynamische und mehrdimensionale Abfragen von englischen Datenschutzerklärungen. Der Kern von Polisis ist ein Datenschutz zentriertes Sprachmodell, das mit 130.000 englischsprachigen Datenschutzerklärungen trainiert wurde. Die Funktionalität wird an zwei Anwendungsbeispielen demonstriert. Zum einen werden Informationen mit Piktogrammen zur verbesserten Verständlichkeit verknüpft, zum anderen können durch eine freie Abfragemöglichkeit gezielt Informationen aus den Datenschutzerklärungen extrahiert werden.

Eine weitere Option zur Verbesserung der Verständlichkeit von Datenschutzerklärungen ist ein Chatbot, der Rückfragen und Wahlmöglichkeiten interaktiv im Gespräch mit dem Nutzer ermöglicht [17].

Weiterhin wurde die Problematik der geringen Verständlichkeit der Datenschutzerklärungen adressiert, indem Methoden zur Identifikation intransparenter Ausdrucksweisen in englischsprachigen Texten entwickelt und evaluiert wurden [22].

Die Informationsextraktion könnte vereinfacht werden, wenn Datenschutzerklärungen bereits in einem maschinenlesbaren Format vorliegen würden. Untersuchungen in diesem Bereich erzielten eine Spezifikation der W3[1] [10], die allerdings nicht erfolgreich war und dementsprechend momentan als obsolet gilt.

Neben der Informationsextraktion aus Datenschutzerklärungen wurden kontextbasierte Methoden auch auf anderen Rechtstexten erforscht. Beispielsweise wurden Paragrafen aus unstrukturierten, italienischen Gesetzestexten nach ihrem Inhalt klassifiziert, um darauf aufbauend relevante Informationen zu extrahieren [6]. Im Gegensatz dazu wurden auch strukturierte italienische Gesetzestexte betrachtet, die als XML-Dateien vorlagen [8].

[1] `https://www.w3.org/TR/P3P/`

3 Dokumentenmodell und Textklassifikation

Zunächst wird in Kapitel 3.1 ein Überblick gegeben und die Zielsetzung dieses Kapitels beleuchtet. Anschließend wird in Kapitel 3.2 das Dokumenten- und Hintergrundmodell vorgestellt. Kapitel 3.3 beschäftigt sich mit Klassen und deren Zuordnung zu Rechts- und Hintergrunddokumenten. In Kapitel 3.4 werden die Entitäten besprochen. Abschließend wird in Kapitel 3.5 die Textanreicherung bei Beachtung der Klasse und den Entitäten gezeigt.

3.1 Überblick und Zielsetzung

Das Ziel des vorliegenden Kapitels ist die Nutzbarmachung der Datenschutzerklärungen zugrunde liegenden übertragbaren Struktur, um diese mithilfe von Anmerkungen der Hintergrundinformationen für juristische Laien verständlich zu machen. Dafür wird zunächst ein Modell der Rechtsdokumente sowie der Hintergrundinformationen ausgearbeitet. Auf dieser Grundlage werden die strukturellen Anteile des Texts inhaltlich anhand der zu erfüllenden Voraussetzungen der gesetzlichen Normen des jeweiligen Landes klassifiziert. Dabei bedingt die jeweilige Klasse das Vorkommen und die Beziehung der Entitäten, welche zur Erreichung des Ziels mit den Hintergrundinformationen verknüpft werden müssen. Daraus ergeben sich zwei Problemstellungen:

- Wie können Informationen in einer Datenschutzerklärung identifiziert und extrahiert werden?
- Wie kann eine passende Hintergrundinformation gefunden werden, um diese als Anmerkung für den Laien wieder ausgeben zu können?

Der Zweck des Rechtsdokumenten- und Hintergrundmodells ist die abstrakte Veranschaulichung juristischer Sachverhalte, um eine generalisierbare Problemlösung für die Datenextraktion und -zuordnung zu formulieren.

Die Rechtsdokumente sind auf die Datenschutzerklärungen beschränkt, da diese im Gegensatz zu anderen Dokumenten eine generalisierbare Struktur aufweisen. Hintergrunddokumente dienen als Blaupause und sind als Textbausteine zu verstehen. Mithilfe dieser Bausteine werden Anmerkungen generiert, indem extrahierte Informationen aus dem Rechtsdokument an vordefinierte Positionen statt "Platzhaltern" eingesetzt werden.

Das zentrale Element ist dabei die Bestimmung des Kontextes in welchem Rahmen Entitäten gefunden werden. Erst durch den Kontext kann eine gefundene Entität verstanden werden. Deswegen wird eine Klassifizierung eingeführt, die thematisch für sich stehende Bestandteile einer Datenschutzerklärung identifiziert und passenden Klassen zuordnet.

Dazu wird ein Überblick über Entitäten gegeben. Diese Entitäten gilt es in einer Datenschutzerklärung zu finden und zu extrahieren.

Nachdem der Kontext über die Klassifizierung ermittelt wurde und die zu erwartenden Entitäten eingeführt worden sind, wird eine Zuordnung zu den Hintergrunddokumenten hergestellt. Anschließend werden die gefundenen Entitäten in die passenden Hintergrunddokumente eingesetzt um so neuen Text auszugeben.

3.2 Dokumente- und Hintergrundmodell

Ein Rechtsdokument enthält einem Titel und zugehörigen Text und untergliedert sich in Abschnitte, welche wiederum einen Titel und zugehörigen Text beinhalten. Der Titel und der zugehörige Text sind Sätze. Sätze wiederum bestehen aus Wörtern, welche eine Sequenz von Zeichen darstellen. Die Kardinalität eines Abschnittes ist in Abbildung 3.2 dargestellt. Weiterhin lässt sich dieser Aufbau folgendermaßen in Abbildung 3.1 als Baumstruktur darstellen.

Abbildung 3.1: Darstellung der Baumstruktur, die sich durch die hierarchische Struktur ergibt $t, n \in \mathbb{N}$

Dabei ergibt sich für den Abschnitt nachfolgend dargestellte Kardinalität:

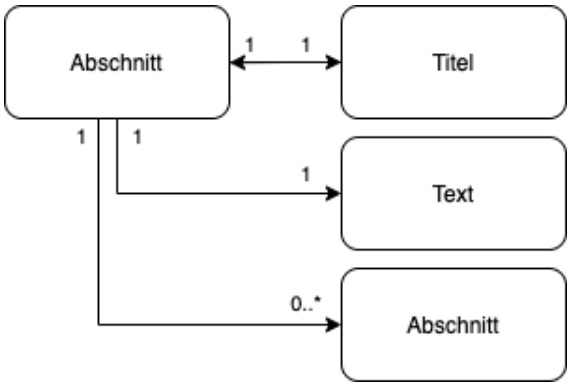


Abbildung 3.2: Abschnitt - Visuelle Darstellung der Kardinalität

Basierend auf der angeführten Struktur wird später in der Arbeit eine Klassifizierung eingeführt. Dabei wird jeder Abschnitt für sich betrachtet um so Rechtsdokument und deren Abschnitten jeweils Klassen zuzuordnen. Dies wird später dazu dienen, zugehörige Hintergrunddokumente zu identifizieren.

Daraus ergibt sich eine Zuordnung von Hintergrund und Abschnitt die sich folgendermaßen in Abbildung 3.3 abbilden lässt:

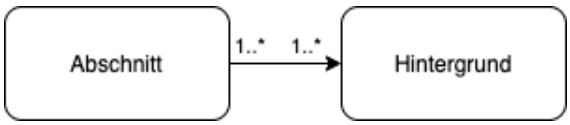


Abbildung 3.3: Hintergrund und Abschnitt - Visuelle Darstellung der Kardinalität

Ein Hintergrunddokument besteht aus strukturiertem Text. Struktur erhält der Text durch die Verwendung von Platzhaltern. Diese bestimmen die Position, an welche Informationen in die generierte Anmerkung aus dem Hintergrunddokument eingesetzt werden sollen.

3.3 Klassifizierung

Klassen stellen im Kontext dieser Arbeit eine Abstraktion der Konzeption juristischer Normen dar. So können Rechtsdokument und Abschnitte jeweils Klassen zugeordnet werden. Die juristische Norm stellt dabei das Hintergrundwissen dar. Dabei ergibt sich über die Zuordnung der Rechtsklasse die Textsorte. Diese Zuordnung ist strikt, so kann ein Rechtsdokument genau einer Rechtsklasse angehören. Beispielsweise wäre eine Rechtsklasse Datenschutzerklärung. Dies ist nachfolgend über die Kardinalität in Abbildung 3.4 dargestellt.

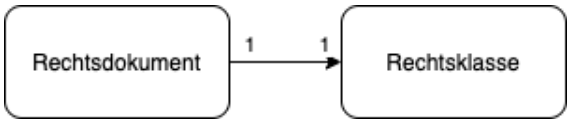


Abbildung 3.4: Rechtsdokument und Rechtsklasse - Visuelle Darstellung der Kardinalität

Durch Zuordnung einer Rechtsklasse ergibt sich eine Menge an geltenden gesetzlichen Normen. Die Zusammensetzung dieser Menge hängt maßgeblich mit dem betrachteten Rechtsgebiet zusammen. Dies sind in der Regel Staatsgebiete wie zum Beispiel "Baden-Württemberg". Dort geltende Datenschutzerklärungen werden auf Grundlage der Datenschutzgrundverordnung, der zugehörigen Erwägungsgründe, des Bundesdatenschutzgesetz und des Landesdatenschutzgesetz (BW) verfasst. Die Kenntnis der maßgeblich beeinflussenden Normen hat pragmatische Folgen. So bildet eine Rechtsklasse eine Menge an Abschnittsklassen ab, die es gilt, in einem solchen Dokument zu identifizieren.

Abschnittsklassen bilden Konzepte der gesetzlichen Norm ab. So wird beispielsweise in der Datenschutzgrundverordnung unter gewissen Voraussetzungen ein Datenschutzbeauftragter verlangt, den es im Rahmen einer Datenschutzerklärung zu nennen gilt. Demnach wird es einen Abschnitt geben, der sich dieser Aufgabe widmet. Daraus würde ein solcher Abschnitt der Klasse Datenschutzbeauftragter zugewiesen werden. Aus dieser Kenntnis ergibt sich eine Menge an zu erwartenden Entitäten und deren Beziehung zueinander. Dargestellt in Abbildung 3.5.

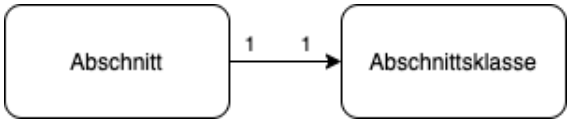


Abbildung 3.5: Abschnitt und Abschnittsklasse - Visuelle Darstellung der Kardinalität

Angelehnt an den Verantwortlichen der Universität Heidelberg [1] ergibt sich daraus folgendes Beispiel:

```
Zeile 0: I. Name und Anschrift des Verantwortlichen
Zeile 1: Max Musterfrau
Zeile 2: Datenstraße 1
Zeile 3: 12345 Datenhausen
Zeile 4: +49 1234 45-67839
```

dort gefundenen Entitäten:

```
Zeile 0: Keine Entitäten gefunden
Zeile 1: Vorname: Max, Nachname: Musterfrau
Zeile 2: Straße: Datenstraße, Hausnummer: 1
Zeile 3: Postleitzahl: 12345, Stadt: Datenhausen
Zeile 4: Telefonnummer: +49 1234 45-67839
```

Zunächst ist durch Kenntnis der Datenschutzgrundverordnung bekannt, dass es sich bei einem Verantwortlichen um eine natürliche Person oder juristische Person handeln kann. Damit begründet dies hinreichend die Suche nach Vor- und Nachnamen sowie einer Firmierung. Ebenso schreibt die Verordnung vor, dass eine Kontaktmöglichkeit bestehen muss, dies begründet eine Suche nach E-Mail-Adresse, Telefonnummer oder Faxnummer.

[1] `https://www.uni-heidelberg.de/de/datenschutzerklaerung`

Die Beziehung der genannten Entitäten zueinander ergibt sich damit wie folgt: Ist nur eine natürliche Person oder juristische Person gefunden worden, so stellt diese den Verantwortlichen dar. Straße, Hausnummer, Postleitzahl und Stadt lokalisieren diesen Verantwortlichen und die gefundene Telefonnummer stellt eine Möglichkeit dar, den Verantwortlichen zu kontaktieren.

Damit stellen die Klassen den Kontext dar, innerhalb dessen Informationen extrahiert werden.

Wir ordnen Hintergrunddokumente mindestens einer Abschnittsklasse zu. Dies dient der späteren Identifikation eines passenden Hintergrunddokuments für die Textanreicherung, welches näher in Kapitel 3.5 beschrieben wird.

3.4 Entitäten

In Form von Entitäten liegen Informationen innerhalb eines Rechtsdokuments vor. Eine Entität ist genau einem Typ und einem Wert zugeordnet. Dies ist dargestellt in Abbildung 3.6.

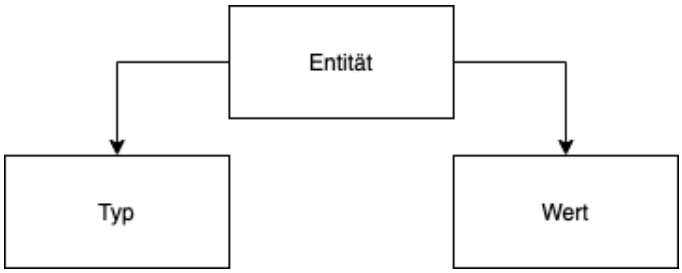


Abbildung 3.6: grafische Darstellung einer Entität

Beispielsweise: Telefonnummer und +49 6221 9876543.

Zunächst wird der Wert einer Entität betrachtet, den es zu extrahieren gilt:

Durch Kenntnis der gewählten Klasse erfolgt die Extraktion von Entitäten innerhalb eines Abschnitts. Dabei liegen bereits über die zuvor erfolgte Klassifizierung Informationen über die zu extrahierenden Entitäten, insbesondere des Typs, vor. Dies vereinfacht die Wahl von Methoden, die dazu geeignet sind, den Entitätswert zu extrahieren. Wenn E-Mail-Adressen extrahiert werden soll, kann das zum Beispiel durch reguläre Ausdrücke kostengünstig und effizient erfolgen.

Nachfolgend wird eine Übersicht der innerhalb einer Datenschutzerklärung zu erwartenden Entitäten gegeben.

Entität	Beispiel
Person	Max Mustermann
Unternehmen	Erika Musterstark GmbH
Straßen	Musterstraße
Hausnummer	1 - 2
Postleitzahl	12345
Stadt	Musterhausen
Faxnummer	+49 121 9876543
Telefonnummer	+49 6221 9876543
E-Mail Adresse	erika.mustermann@uni-heidelberg.de
Rechtszitat	Art. 6 Abs. 1 S.1 lit. a DSGVO
Drittdienst	Google Analytics
Datum	IP-Adresse, Browserversion, Adressdaten [...]

Tabelle 3.1: Übersicht über Entitäten

3.5 Textanreicherung

Durch die bisher beschriebenen Schritte und eingeführten Modelle liegen nun gefundene Entitäten mit Bezeichnung und Wert in Abhängigkeit von mindestens zwei Klassen vor. Dies wäre die zugeordnete Klasse des Rechtsdokuments, sowie des Abschnittes.

Wurde einem Abschnitt mehr als eine Klasse zugeordnet oder liegt der Abschnitt in Abhängigkeit eines anderen Abschnittes, liegen mehrere Klassen vor.

Durch die Kenntnis der Klassen kann nun das passende Hintergrunddokument identifiziert werden. Dabei wird ein Hintergrunddokument durch seine zugeordneten Rechts- und Abschnittsklasse identifiziert. Eine Abhängigkeit ergibt sich durch Betrachtung der Baumstruktur eines Rechtsdokuments, wie im Dokumentenmodell beschrieben wurde. Dies wird in Abbildung 3.7 dargestellt.

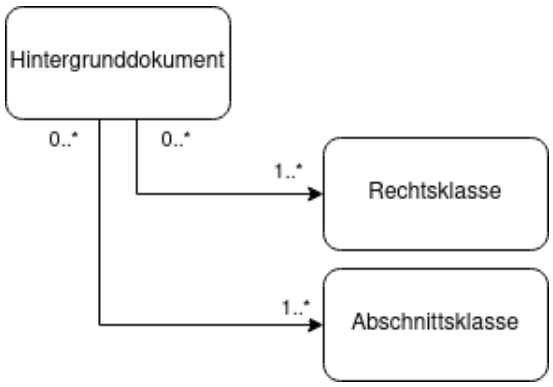


Abbildung 3.7: Hintergrundmodell - Visuelle Darstellung der Kardinalität

Hintergrunddokumente sehen durch die Nutzung von Platzhaltern eine entsprechende Textanreicherung vor. Hintergrunddokumente können dabei als Textbaustein verstanden werden. Dieser wird dann eingesetzt, falls eine Entität gefunden wurde.
Dabei kann beispielsweise ein Text aus zusammengefügten Textbausteinen folgendermaßen aussehen:

Hast du Fragen betreffend der Verarbeitung und Speicherung von personenbezogenen Daten auf dieser Webseite, so steht der Datenschutzbeauftragte PERSON_FIRST für Fragen zur Verfügung. Der Datenschutzbeauftragte kann über folgende Rufnummer: PHONE_FIRST erreicht werden. Der Datenschutzbeauftragte steht auch unter folgender E-Mail-Adresse zur Verfügung: EMAIL_FIRST.

Die erste Information des Platzhalters gibt den Typ einer Entität an, der in diesem Punkt eingesetzt werden kann. Die zweite Information gibt das Vorkommen der Entität an. So wird hier im Beispiel jeweils das erste Vorkommen einer Entität gefordert. Denkbar wäre dahingehend auch eine Liste von allen Entitäten. Dies kann im Kontext der Datenschutzerklärung beispielsweise erfolgen, wenn die Daten aufgelistet werden sollen, die im Rahmen der Nutzung einer Webseite verarbeitet werden. Wurden die Entitäten Vorname: Max, Nachname: Musterfrau und Telefonnummer 0123467890 gefunden, so ergibt dies:

Hast du Fragen betreffend der Verarbeitung und Speicherung von personenbezogenen Daten auf dieser Webseite, so steht der Datenschutzbeauftragte Max Musterfrau für Fragen zur Verfügung. Der Datenschutzbeauftragte kann über folgende Rufnummer: +49 6221 9876543 erreicht werden.

Der generierte Text kann so beispielsweise hilfreiche Erläuterungen für den Leser zur Verfügung stellen oder lediglich eine Zusammenfassung von Informationen sein, die weit über ein Dokument verstreut sind.

4 Experimentelle Evaluation

Dieser Abschnitt behandelt die experimentelle Evaluation und die Implementierung eines Prototyps. Zuerst wird in Kapitel 4.1 ein Überblick über den zugrunde liegenden Datensatz gegeben. In Kapitel 4.2 wird die Erstellung und Annotation des Datensatzes beleuchtet. Kapitel 4.3 behandelt die Implementierung eines Prototyps. Dies untergliedert sich in 4 Unterabschnitte. Zunächst wird in Kapitel 4.3.1 die Aufbereitung der Datenschutzerklärungen beleuchtet, anschließend beschreibt Kapitel 4.3.2 die Klassifikation der Abschnitte. Kapitel 4.3.3 beschreibt die Extraktion Abschnitts spezifischer Informationen. Abschließend demonstriert Kapitel 4.3.4 den Prototyp.

4.1 Zusammensetzung des Datensatzes

Der verwendete Datensatz umfasst 100 Datenschutzerklärungen von deutschsprachigen Webseiten, die in HTML als semistrukturierter Text vorliegen.

Die Ermittlung des Betreibers wurde in Handarbeit vorgenommen. Im Gegensatz dazu haben Färber et. al. gezeigt, wie eine Klassifikation erfolgreich automatisiert durchgeführt werden kann [14]. Diese Methode kann bei großen Datensätzen von Vorteil sein, wird hier aber nicht verwendet, weil der Aufwand der Implementierung den Nutzen für einen Datensatz der vorliegenden Größe übersteigt.

Die Beschreibung des Datensatzes kann anhand des Betreibers der Websites vorgenommen werden. Dabei waren privatwirtschaftliche Organisationen als Betreiber von 89 Websites als häufigstes vertreten. Gemeinnützige und öffentliche Organisation machen einen kleineren Anteil aus. Abbildung 4.1 stellt die genaue Zusammensetzung dar.

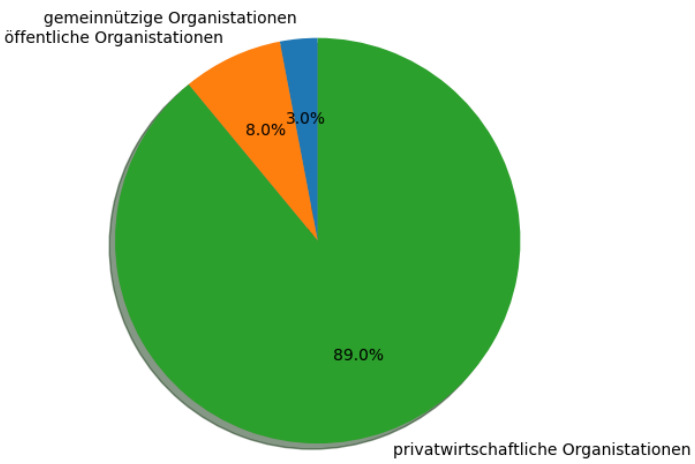


Abbildung 4.1: Zusammensetzung des Datensatzes nach Betreiber

Beispielsweise umfasst die Kategorie gemeinnützige Organisationen unter anderem: tierschutzstiftungen.de, oh-heilbronn.de und start-stiftung.de. Die Kategorie öffentlicher Organistationen umfasst unter anderem: uni-heidelberg.de, uni-erfurt.de und tagesschau.de. Privatwirtschaftliche Organisationen sind unter anderem: google.de, amazon.de und bonprix.de

4.2 Erstellung des Datensatzes

Der Datensatz wurde mithilfe eines Web Crawler erstellt. Die Vorgehensweise beinhaltet 4 Schritte:
Schritt 1: Zunächst wurde eine Internetsuche nach den 1000 häufig in der deutschen Sprache vorkommenden Worten durchgeführt[1]. Diese wurde mithilfe der Python Library google-search[2] ausgeführt. Durch Angabe des Parameters *lang=de* wurden die Ergebnisse nach deutschsprachigen Inhalten vorgefiltert. Als Suchmaschine im Hintergrund wurde Web Scraping Google verwendet.

Schritt 2: Aus Schritt 1 wurde ein Fully Qualified Domain Name (FQDN) übergeben, mithilfe dessen über die Python Library: request[3] die Stammseite der Internetseite heruntergeladen wird. Diese wird nach Links durchsucht, deren Text auf eine Datenschutzerklärung hinweist.

[1] `https://1000mostcommonwords.com/1000-most-common-german-words/`
[2] `https://pypi.org/project/google-search/`
[3] `https://requests.readthedocs.io/en/master/`

Als Stichworte wurden folgende gewählt: (1) Datenschutzerklärung, (2) Datenschutzrichtlinie, (3) Datenschutzinfo, (4) Datenschutzbestimmungen und (5) Datenschutz.

Schritt 3: Aus Schritt 2 wurde die Uniform Resource Locator (URL) einer potenziellen Datenschutzerklärung übergeben, die daraufhin als HTML heruntergeladen wird. Die Gliederung, die sich aus den deklarierten Überschriften innerhalb des semi-strukturierten Text ergibt, wurde anschließend visuell aufbereitet. Dafür wurde die Python Library beautifulsoup4[4] und HTMLParser[5] verwendet.

Schritt 4: Anhand der visuellen Darstellung der Gliederung, die mithilfe der Python Library treelib[6] erstellt wurde, wurde von Hand interaktiv entschieden, ob diese in den Datensatz aufgenommen wird. Gründe für die Ablehnung einer Website waren ein nichtdeutscher Inhalt, die fehlende Deklaration von Überschriften in HTML oder eine nicht abrufbare Datenschutzerklärung.

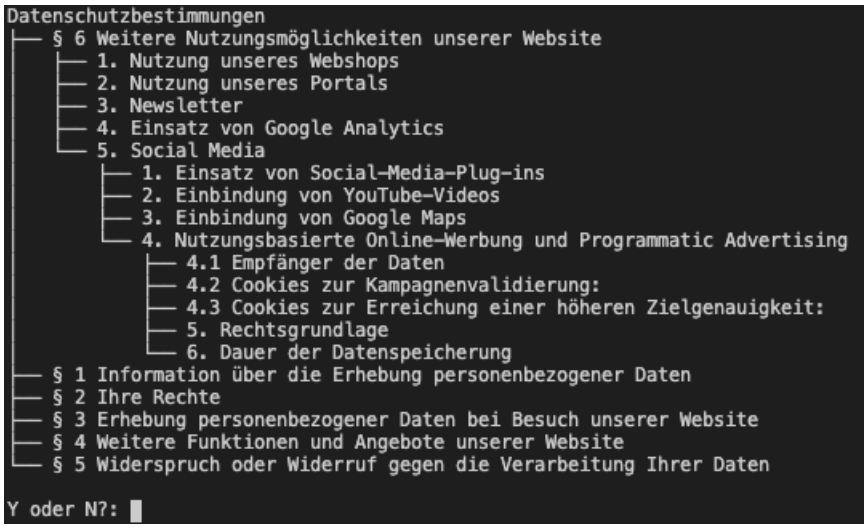


Abbildung 4.2: Ja oder Nein? Entscheidung anhand der Gliederung

Abbildung 4.2 zeigt den interaktiven Auswahlprozess, wobei nacheinander Gliederungen angezeigt werden, die jeweils durch Angabe von Ja (Y) oder Nein (N) in den Datensatz aufgenommen oder verworfen werden. Beispielhaft ist hier eine nutzbare Gliederung dargestellt.

[4] `https://pypi.org/project/beautifulsoup4/`
[5] `https://docs.python.org/3/library/html.parser.html`
[6] `https://treelib.readthedocs.io/en/latest/`

Die beschriebenen Schritte sind in Abbildung 4.3 dargestellt. Die geschilderte Vorgehensweise wurde so lange ausgeführt, bis ein Datensatz von 100 deutschsprachigen Datenschutzerklärungen aufgebaut wurde.

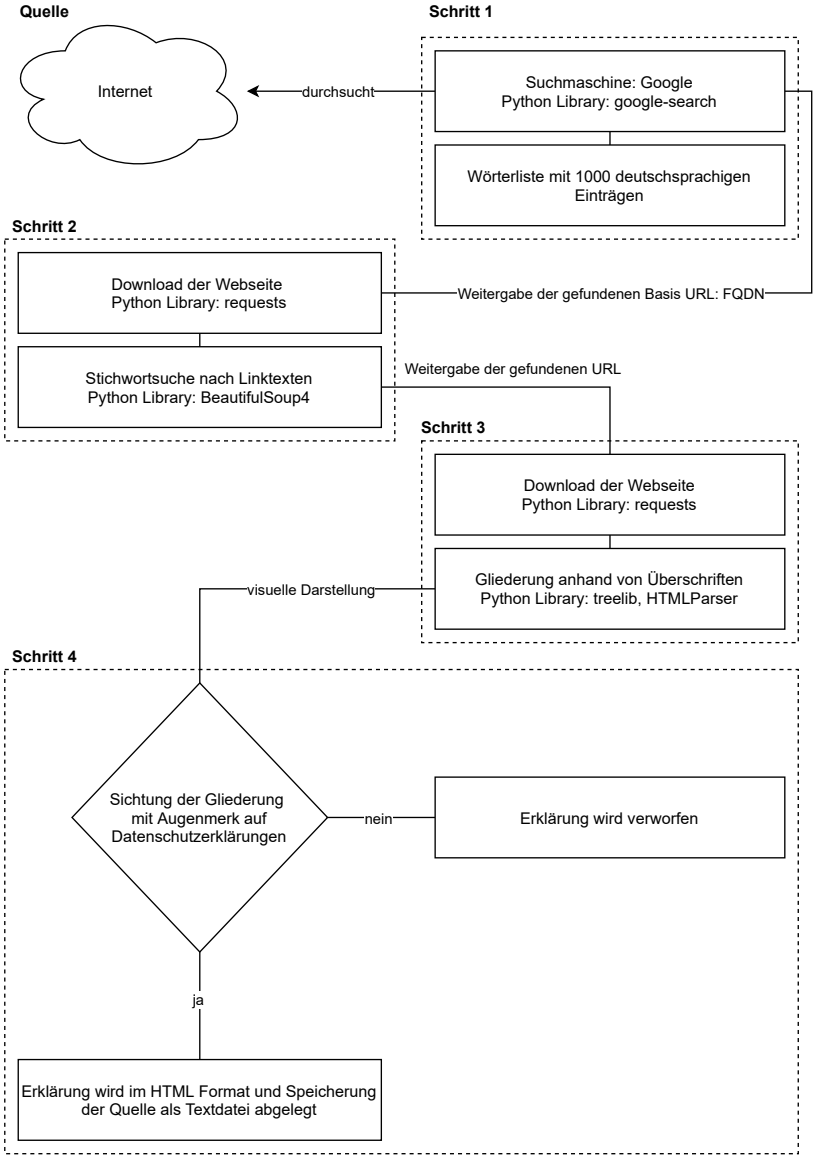


Abbildung 4.3: Visuelle Darstellung des Datensatzaufbaus

Nachdem der Datensatz aufgebaut worden war, galt es diesen zu annotieren. Dafür wurden verschiedene Methoden genutzt.

Zunächst wurde die Klassen eines Abschnittes bestimmt, dadurch wurden die 10.053 vorliegenden Abschnitte der 100 deutschsprachigen Datenschutzerklärungen mithilfe der Software Doccano[7] den betrachtenden Klassen zugeordnet. Dabei wurden auch einige Abschnitte gefunden, die zwar durch Verwendung von Überschriften auf der Webseite entstanden sind, sich jedoch keiner der Klassen und folglich damit nicht Bestandteil einer Datenschutzerklärung waren. Aus den 10.049 Abschnitten konnten 1099 Abschnitte einer der folgenden Klasse zugeordnet werden: (1) Datenerhebung, (2) Rechtsgrundlage, (3) Drittdienst und (4) Verantwortlichkeit. Abbildung 4.4 zeigt das Interface der Software Doccano, welches zur Klassifizierung genutzt wurde.

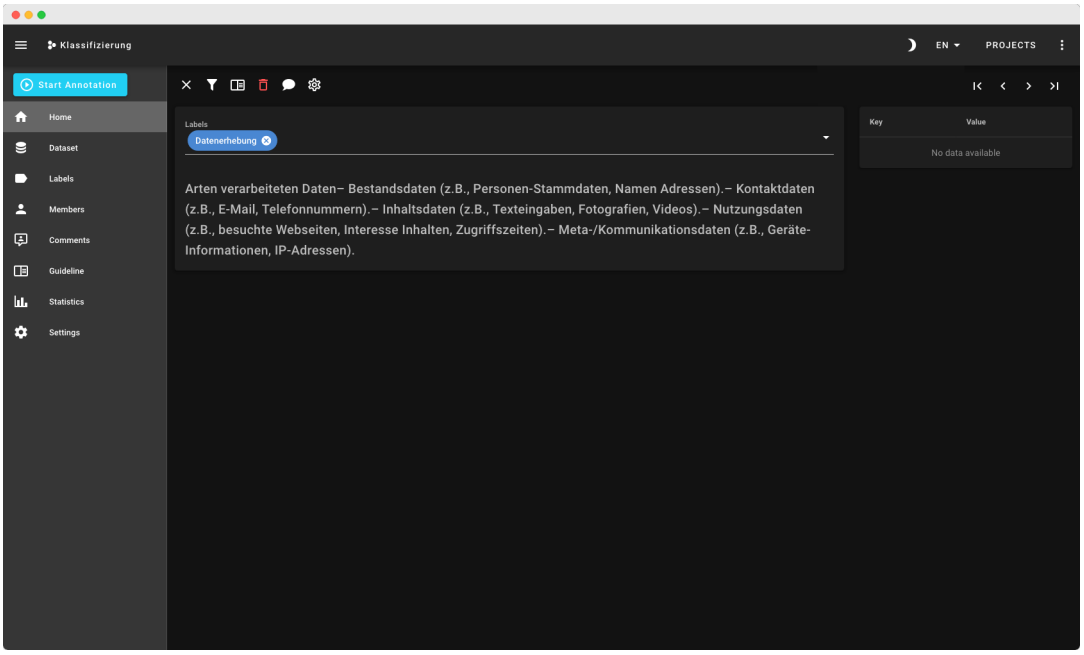


Abbildung 4.4: Klassifizierung der Abschnitte mithilfe von Doccano

Nachdem die Klassen bestimmt wurden galt es die abschnittsspezifischen Entitäten zu annotieren.

(1) Datenerhebung: Bei der Datenerhebung galt es Begrifflichkeiten zu annotieren die bezeichnend für Daten sind die im Rahmen der Datenschutzerklärung zur Verarbeitung genannt worden. Dafür wurde ebenfalls Doccano verwendet bei der in 518 Abschnitten insgesamt 2069 Begrifflichkeiten annotiert. Abbildung 4.5 zeigt die Markierung von Begrifflichkeiten die ein Datum referenzieren mithilfe der Software Doccano.

[7] `https://github.com/doccano/doccano`

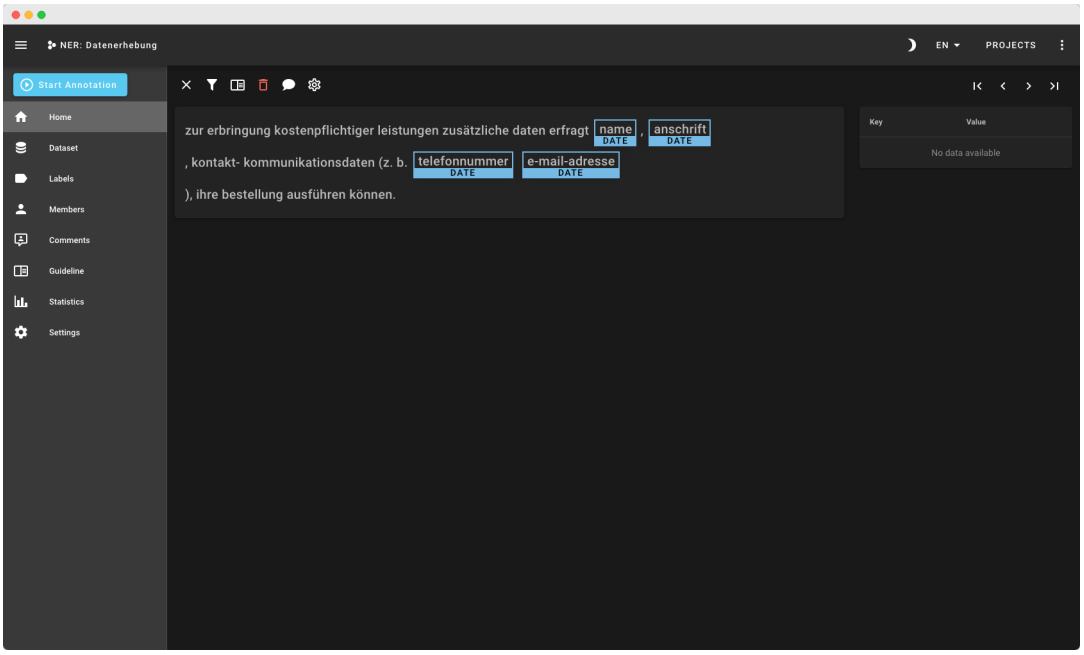


Abbildung 4.5: Markierung der Entitäten mithilfe von Doccano

(2) Rechtsgrundlage: Bei Rechtsgrundlagen galt es Rechtszitate, die auf eine Gesetzesgrundlage verweisen zu extrahieren. Dafür wurden in 565 Abschnitten insgesamt 535 Rechtszitate annotiert. Die kleinere Zahl ergibt sich daraus, dass einige Webseiten keine Rechtszitate genutzt haben, um auf die Rechtsgrundlage zu verweisen. Zur Annotierung wurde eine selbst entwickelte Software verwendet die bereits das vortrainierte Modell von Leitner [21] nutzt, um Rechtszitate zu erkennen[8]. Dabei bestand innerhalb der Software die Möglichkeiten gefundene Rechtszitate vollständig, teilweise oder gar nicht zu akzeptieren und bei beiden letzteren eine Korrektur vorzunehmen. Dies wird in Abbildung 4.6 dargestellt.

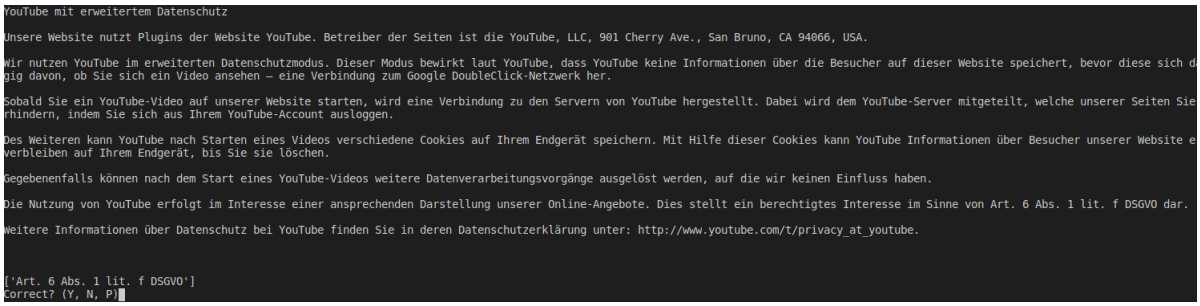


Abbildung 4.6: Annahme (Y), Ablehnung (N) oder Teilweise (P) - Annotierung der Rechtsgrundlagen

[8] https://github.com/elenanereiss/Legal-Entity-Recognition

Das Resultat wurde als JSON formatierte Datei abgespeichert. Diese Datei enthielt als Information die Referenz in Form des Dateinamen sowie die Rechtsgrundlagen die gefunden wurden. Dies verdeutlicht an einem nachfolgenden Abschnitt der Datenschutzerklärung von der Zeit Verlagsgruppe.

```
{"reference": "zeit.de.18", "result": ["Art. 6 lit. a) DSGVO"]}
```

(3) Drittdienst: Bei den Drittdiensten galt es Bezeichnungen von Drittdiensten die im Rahmen der Datenschutzerklärung genannt wurden zu annotieren. Dafür wurden in 506 Abschnitten insgesamt 2919 Drittdienste annotiert. Beispiel eines Drittdienstes wäre "Google Analytics".

(4) Verantwortlichkeit: Bei der Verantwortlichkeit galt es Kontaktinformationen zu extrahieren. Folgende Kontaktinformationen wurden beachtet: (1) Person, (2) Unternehmen / Firmierung, (3) Straße, (4) Hausnummer, (5) Postleitzahl, (6) Stadt, (7) Telefonnummer, (8) Faxnummer und (9) E-Mail-Adressen. Dabei wurden in 98 Abschnitten insgesamt 621 Entitäten annotiert. Es ergibt sich folgende Aufschlüsselung:

Entität	Anzahl
Person	39
Unternehmen	82
Straße	85
Hausnummer	85
Postleitzahl	84
Stadt	85
Fax	22
E-Mail	68
Telefon	61
Summe	**611**

Tabelle 4.1: Überblick über annotierte Entitäten

4.3 Implementierung

Der Weg von einer Datenschutzerklärung zu der Generierung von Anmerkungen gliedert sich in (4.3.1) Aufbereitung, (4.3.2) Klassifikation, (4.4.3) Extraktion von Entitäten und abschließend (4.3.4) der Generierung von neuem Text, der nachfolgend als Anmerkung bezeichnet wird. Diese Vorgehensweise ist in Abbildung 4.7 übersichtlich dargestellt.

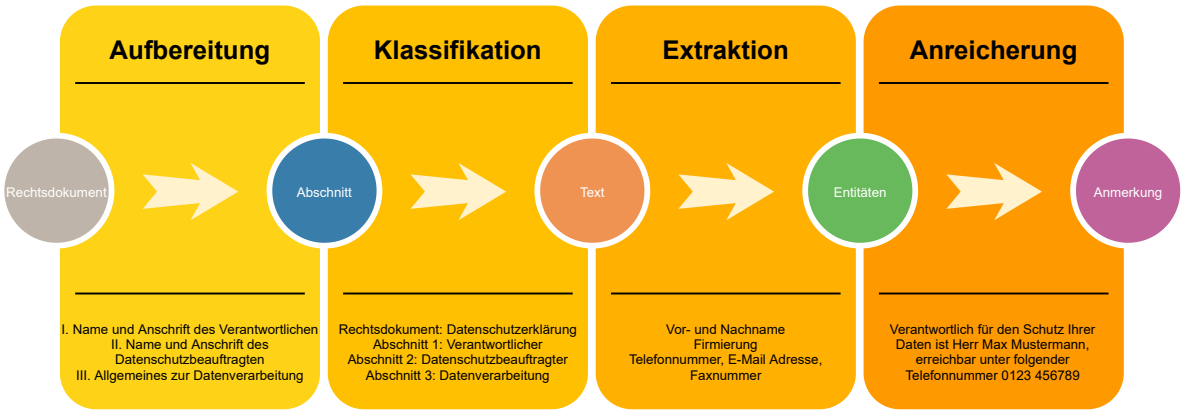


Abbildung 4.7: Vom Rechtsdokument zur Anmerkung

4.3.1 Aufbereitung

Der Datensatz beinhaltet semistrukturierten Text im HTML-Format. Ziel der Aufbereitung ist die Extraktion der Struktur, die sich durch Verwendung von Header-Tags ergibt. Der HTML-Standard gibt Überschriften von h1 bis h6 vor. Es werden in der Regel nicht alle 6 Abstufungen verwendet. Teilweise wird die Hierarchie nicht eingehalten und ein Level übersprungen. So gibt es beispielsweise Webseiten, die h2-Tag und h4-Tag verwenden, aber keinen h3-Tag. Daraus ergibt sich als Anforderung an die Implementierung eine gewisse Fehlertoleranz.

Nachfolgend die Datenschutzerklärung der Universität Heidelberg als Baum in Abbildung 4.8 dargestellt.

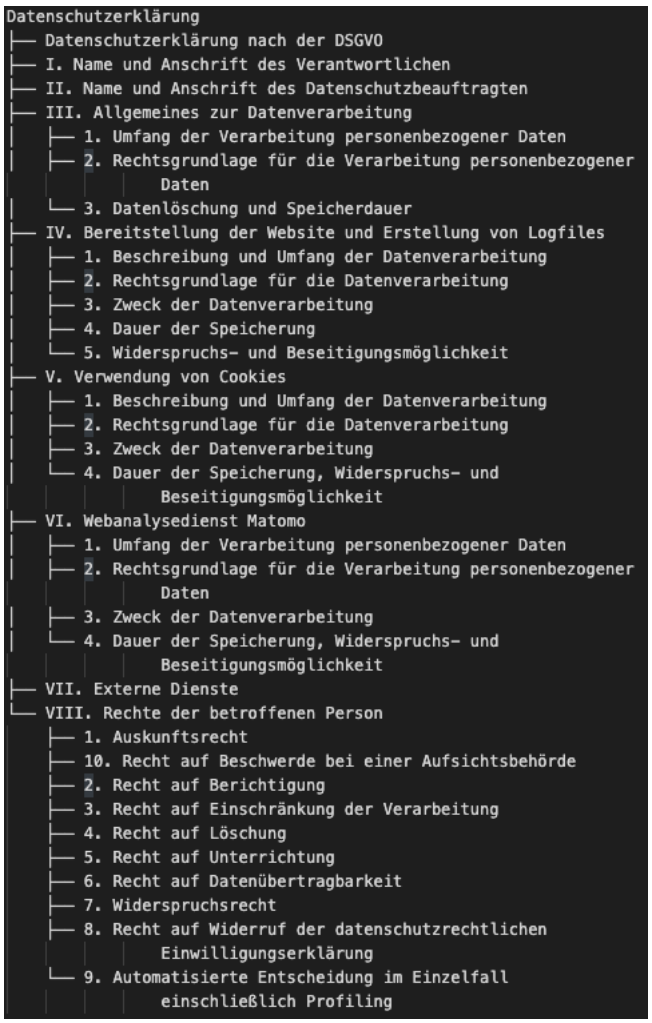


Abbildung 4.8: Gliederung der Datenschutzerklärung der Universität Heidelberg

Die so erhaltenen Abschnitte werden abgetrennt und separat in JSONL[9] in Kleinschreibung gespeichert. Dabei werden Stoppwörter, Absätze und zusätzliche Leerzeichen entfernt. Dazu wurde die Python Library nltk[10] verwendet. Für die Evaluation der Klassifikation erfolgt eine Separierung in (1) Titel, (2) Titel und Abschnittstext und (3) Abschnittstext.

4.3.2 Klassifikation

Ziel der Klassifikation ist, den Abschnitten eine Klasse zuzuordnen. Im Rahmen der Evaluation werden folgende Klassen betrachtet: (1) Drittdienste, (2) Verantwortlichkeit der Datenerhebung, (3) Umfang der Datenerhebung und (4) Rechtsgrundlage der Datenerhebung.

[9] `https://jsonlines.org/`
[10] `https://www.nltk.org/`

Die Abschnitte aus der Aufbereitung wurden manuell diesen Klassen zugeordnet. Daraus ergibt sich folgende Verteilung dargestellt in Tabelle 4.2 und Abbildung 4.9.

Klasse	**Anzahl an Textabschnitten**
Datenerhebung und Rechtsgrundlage	248
Drittdienste	232
Datenerhebung	181
Drittdienste und Rechtsgrundlage	163
Verantwortlichkeit	97
Rechtsgrundlage, Drittdienste und Datenerhebung	87
Rechtsgrundlage	67
Datenerhebung und Drittdienste	24
Summe	**1099**

Tabelle 4.2: Übersicht über den annotierten Datensatz nach Klasse

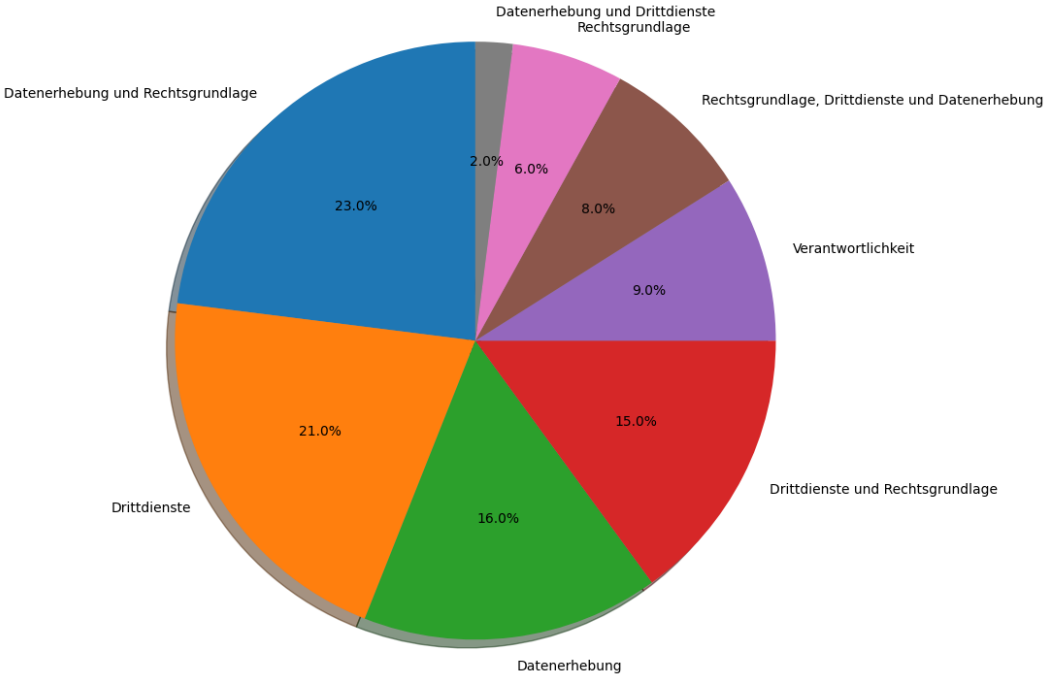


Abbildung 4.9: Zusammensetzung Datensatz nach festgelegten Klassen

Kombinationen von Klassen, die in der Tabelle und Abbildung 4.9 nicht aufgeführt sind, wurden im Rahmen der Annotation nicht gefunden.

Folglich sind diese Kombinationen nicht im Datensatz vorhanden und der Übersicht halber nicht in der Tabelle und in der Grafik dargestellt.

Evaluiert werden zwei Ansätze, der erste Ansatz besteht aus einem regelbasierten Ansatz. Dabei werden Begrifflichkeiten aus der Datenschutzgrundverordnung entnommen, die in einem solchen Abschnitt erwartet werden. Beispielsweise soll die Klasse Rechtsgrundlage durch Dokumente gekennzeichnet sein, bei denen die Rechtsgrundlage angegeben ist.

4.3.2.1 Regelbasierter Ansatz

Zunächst wird ein regelbasierter Ansatz geprüft. Dabei wird über alle Abschnittsklassen zunächst die Worthäufigkeit angeschaut und die Worte ausgewählt die besonders in einer Abschnittsklasse vorkommen ohne, in den anderen übermäßig aufzutreten. Dabei ergeben sich folgende Worte:

- (1) Datenerhebung: keine
- (2) Rechtsgrundlage: 'rechtsgrundlage', 'rechtsgrundlagen', 'grundlage', 'grundlagen'
- (3) Drittdienste: 'übermitteln', 'übermittelt', 'übermittelte', 'übermittelten', 'übertragen', 'übertragene', 'übertragenen', 'übertragung'
- (4) Verantwortlichkeit: 'verantwortlich', 'verantwortliche', 'verantwortlichen', 'verantwortlicher', 'verantwortlichkeit', 'Verantwortlichkeit'

Dabei zeichnet sich ab, dass für Rechtsgrundlage, Drittdienste und Verantwortlichkeit sich Wörter finden lassen die bezeichnend für eine der Abschnittsklassen sind. Bei der Klasse Datenerhebung stechen keine Wörter hervor, die auffällig nur in dieser Abschnittsklasse auftreten. Daraus werden folgende Regeln aufgestellt die es zu evaluieren gilt:

(2) Rechtsgrundlage: Enthält ein Abschnitt folgende Wörter: 'rechtsgrundlage' oder 'rechtsgrundlagen' oder 'grundlage' oder 'grundlagen'

(3) Drittdienste: Enthält ein Abschnitt folgende Wörter: 'übermitteln' oder 'übermittelt' oder 'übermittelte' oder 'übermittelten' oder 'übertragen' oder 'übertragene' oder 'übertragenen' oder 'übertragung'

(4) Verantwortlichkeit: Enthält ein Abschnitt folgende Wörter: 'verantwortlich' oder 'verantwortliche' oder 'verantwortlichen' oder 'verantwortlicher' oder 'verantwortlichkeit' oder 'Verantwortlichkeit'

Um nun die Klassifizierung anhand des regelbasierten Ansatzes zu prüfen, gilt festzuhalten, in welchem Fall ein Erfolg vorliegt. Dies gilt dann, sobald ein Abschnitt identifiziert wurde, in dem Informationen der Klasse zu erwarten sind. Beispielsweise sollen auch Abschnitte erkannt werden, denen mehrere Klassen zugeordnet wurden. Dabei wurden drei Durchgänge ausgeführt, bei denen jeweils die Klassifizierung anhand des Titels, des Abschnittstextes und Titel und Abschnittstext zu erfolgen hat.

(2) Rechtsgrundlage:

Regelanwendung auf	Precision	Recall	F1-score	Accurancy
Titel	98,39%	10,80%	19,46%	54,13%
Abschnittstext	84,91%	80,71%	82,74%	82,75%
Titel und Abschnittstext	84,92%	81,78%	82,76%	82,75%

Tabelle 4.3: Ergebnisse regelbasierter Ansatz zur Klassifizierung von Rechtsgrundlagen

Rechtsgrundlagen werden selten eigene Abschnitte gewidmet, diese finden sich meistens zusätzlich in Abschnitten die zusätzlich auch anderen Klassen angehören. Wird ein Abschnitt jedoch den Rechtsgrundlagen alleine gewidmet, lautet der Titel meistens Rechtsgrundlage. Daraus lässt sich das schlechte Ergebnis des Titels ableiten.

(3) Drittdienste:

Regelanwendung auf	Precision	Recall	F1-score	Accurancy
Titel	-	0%	-	54,04%
Abschnittstext	46,59%	32,41%	38,22%	51,86%
Titel und Abschnittstext	59,64%	54,18%	56,78%	63,22%

Tabelle 4.4: Ergebnisse regelbasierter Ansatz zur Klassifizierung von Drittdiensten

(4) Verantwortlichkeit:

Regelanwendung auf	Precision	Recall	F1-score	Accurancy
Titel	95,83%	69,70%	80,70%	97,00%
Abschnittstext	35,47%	83,84%	49,85%	84,83%
Titel und Abschnittstext	39,13%	100%	56,25%	86,01%

Tabelle 4.5: Ergebnisse regelbasierter Ansatz zur Klassifizierung von Verantwortlichkeit

4.3.2.2 Support Vector Machine (SVM)

Nachfolgend wurde eine Evaluation mithilfe einer SVM ausgeführt. Dazu wurde die Implementierung von Sklearn[11] genutzt, die mit folgenden Parametern aufgerufen wurde: loss=hinge (Festlegung der Verlustfunktion), penalty=l2 ("Slack-Variable"). Dabei wurden folgende Ergebnisse erzielt:

(1) Datenerhebung:

SVM auf	Precision	Recall	F1-score
Titel	46,88%	46,76%	46,82%
Abschnittstext	45,77%	54,05%	34,84%
Titel und Abschnittstext	46,29%	48,92%	46,57%

Tabelle 4.6: Ergebnisse einer SVM zur Klassifizierung von Datenerhebung

(2) Rechtsgrundlage:

SVM auf	Precision	Recall	F1-score
Titel	35,06%	34,62%	34,84%
Abschnittstext	34,83%	25,90%	29,71%
Titel und Abschnittstext	36,02%	34,36%	35,17%

Tabelle 4.7: Ergebnisse einer SVM zur Klassifizierung von Rechtsgrundlage

[11] `https://scikit-learn.org/`

(3) Drittdienste:

SVM auf	Precision	Recall	F1-score
Titel	56,44%	59,51%	57,94%
Abschnittstext	55,47%	59,24%	57,29%
Titel und Abschnittstext	58,71%	59,51%	59,11%

Tabelle 4.8: Ergebnisse einer SVM zur Klassifizierung von Drittdienste

(4) Verantwortlichkeit:

SVM auf	Precision	Recall	F1-score
Titel	98,25%	78,87%	87,50%
Abschnittstext	77,22%	85,92%	81,33%
Titel und Abschnittstext	95,24%	84,51%	89,55%

Tabelle 4.9: Ergebnisse einer SVM zur Klassifizierung von Verantwortlichkeit

4.3.3 Extraktion von Informationen

Nachdem im vorherigen Kapitel die Klassifizierung evaluiert wurde, gilt es nun die Abschnitts spezifische Informationsextraktion zu evaluieren.

Dabei wurde in Datenerhebung und Drittdienste der annotierte Datensatz in Training- (70%) und Testdatensatz (30%) aufgeteilt. Mit Verwendung von Spacy[12] wurde daraufhin ein Modell trainiert und evaluiert.

4.3.3.1 Datenerhebung

Hier gilt es, Entitäten zu identifizieren, die ein Datum beschreiben, welches durch den Datenverarbeiter verarbeitet wird. Dazu wurde Named Entity Recognition genutzt.

Dabei konnte für die Extraktion der Datenerhebung eine Precision von 97,22%, einen Recall von 98,59% und einen F1-score von 97,90% erzielt werden.

[12] `https://spacy.io`

4.3.3.2 Rechtsgrundlage

Hier gilt es, primär Rechtszitate zu erkennen, die auf eine Rechtsgrundlage verweisen. Dazu wurde Legal Entity Recognition (LER) genutzt, welches eine spezialisierte Form von Named Entity Recognition (NER) darstellt. Es wurde von Leitner [21] auf der Basis der Arbeit von Reimers und Gurevych [26] entwickelt. Dabei wurde auf einem Bidirectional Long short-term memory (BLSTM) trainiertem Modell[13] folgende Ergebnisse erzielt. Eine Precision 88,73%, einem Recall von 81,35%, einer Accurancy von 99,89% und einem F1-score von 84,65%.

4.3.3.3 Drittdienste

Hier gilt es, Entitäten zu identifizieren, die einen Drittdienst bezeichnen. Dazu wurde Named Entity Recognition genutzt. Hosseini et. al. haben gezeigt, dass NER mit vortrainierten Modellen schlechte Ergebnisse liefern und sich die Ergebnisse verbessern lassen, indem ein eigenes Modell dafür trainiert wird [7].

Dabei konnte für die Extraktion von Drittdiensten eine Precision von 92,86% einen Recall von 100% und einen F1-score von 96,30% erzielt werden.

4.3.3.4 Verantwortlichkeit

Hier gilt es, Kontaktinformationen zu extrahieren. Dazu wurden je nach Entität verschiedene Ansätze genutzt. Dabei wurde eine vergleichbare Vorgehensweise im PGuard Projekt gewählt [19, S. 90-91].

Dabei ergab sich folgende Übersicht:

Entität	Methode	Anmerkung	Precision	Recall	Accuracy	F1-score
Person	spacy NER	Modell: de_core_news_lg	53.06%	72,22%	68,27%	61.18%
Unternehmen	spacy NER	Modell: de_core_news_lg	83,78%	75,61%	41,24%	79,49%
Straße	statistisches NLP [5]	libpostal	100%	91,76%	92,78%	95,71%
Hausnummer	statistisches NLP [5]	libpostal	100%	92,94%	93,81%	96,34%
Postleitzahl	statistisches NLP [5]	libpostal	100%	100%	100%	100%
Stadt	statistisches NLP [5]	libpostal	100%	95,29%	95,87%	97,59%
Telefon	regelbasiert	phonenumbers	98,31%	61,05%	97,94%	75,32%
Fax	regelbasiert	phonenumbers	81,48%	100%	94,85%	89,80%
E-Mail Adresse	regelbasiert	regex	100%	69,47%	58,28%	81,99%
Zusammengefasst	**-**	**-**	**93,75%**	**86,63%**	**86,82%**	**90,05%**

Tabelle 4.10: Ergebnisse Extraktion von Kontaktinformationen

[13] `https://github.com/UKPLab/emnlp2017-bilstm-cnn-crf`

Die hohe Performance zeigt, dass der statistische NLP-Ansatz, den die C Library libpostal[14] und deren Python Wrapper pypostal[15] verwendet, für die Extraktion von Adressen geeignet ist. Auch die regelbasierten Ansätze, die Python Library phonenumbers[16] und der reguläre Ausdruck der benutzt wurde[17], erzielte gute Ergebnisse für die Extraktion von Telefonnummern und E-Mail-Adressen. Der vergleichsweise niedrige Recall kann durch die Erkennung nicht-annotierter E-Mail-Adressen erklärt werden. Das vortrainierte Spacy-Modell[18] eignet sich für die Extraktion von Personen und Unternehmen, wobei die Erkennung im Vergleich zu den anderen Ansätzen niedriger ausfällt.

4.3.4 Textanreicherung

Mithilfe der im vorherigen Kapitel erzielten Ergebnisse der Evaluation gilt es nun einen Prototyp zur Demonstration zu implementieren.

Zunächst wird eine Anforderung an den Prototyp in Form einer User-Story formuliert:

> Als Anwender möchte ich eine HTML Datei einer Datenschutzerklärung als Eingabe verwenden und erwarte dabei als Ausgabe Informationen über die Datenerhebung, Rechtsgrundlage, Drittdienste und Verantwortlichkeit, damit ich die Datenschutzerklärung nicht lesen muss.

Protoyp: Zunächst gilt es, die HTML Datei in Abschnitte zu zerlegen, um diese anschließend zu klassifizieren. Dabei werden folgende Methoden angewendet, um die jeweilige Klasse zu identifizieren:

Klasse	Methode	Anwendung auf
(1) Datenerhebung	SVM	Titel
(2) Rechtsgrundlage	regelbasiert	Titel und Abschnittstext
(3) Drittdienste	regelbasiert	Titel und Abschnittstext
(4) Verantwortlichkeit	SVM	Titel und Abschnittstext

Tabelle 4.11: Protoyp: Übersicht gewählter Methoden zur Klassifizierung

Nach der Erkennung der einzelnen Abschnitte wurden die abschnittsspezifischen Informationen extrahiert.

[14] https://github.com/openvenues/libpostal
[15] https://github.com/openvenues/pypostal
[16] https://pypi.org/project/phonenumbers/
[17] https://stackoverflow.com/questions/201323/201378#201378
[18] https://spacy.io/models/de/

Dabei ergab sich folgende Übersicht:

Klasse und Information	Methode
(1) Datenerhebung	
1.1 Datum	NER
(2) Rechtsgrundlage	
1.2 Rechtszitat	LER
(3) Drittdienste	
3.1 Bezeichnung von Drittdiensten	NER
(4) Verantwortlichkeit	
4.1 Person	NER
4.2 Unternehmen / Firmierung	NER
4.3 Straße	Statistik
4.4 Hausnummer	Statistik
4.5 Postleitzahl	Statistik
4.6 Stadt	Statistik
4.7 Telefon	regelbasiert
4.8 Fax	regelbasiert
4.9 E-Mail Adresse	regelbasiert

Tabelle 4.12: Übersicht der verwendeten Methoden

Folgende Textbausteine wurden verwendet:

Klasse und Information	Textbaustein
(1) Datenerhebung	
1.1 Datum	Es werden folgende Daten erhoben: DATUM_LIST
(2) Rechtsgrundlage	
1.2 Rechtszitat	Folgende Rechtsgrundlagen werden zur Verarbeitung von Daten angeführt: LEGAL_LIST
(3) Drittdienste	
3.1 Bezeichnung von Drittdiensten	Es werden folgende Drittdienste verwendet: THIRDPARTY_LIST
(4) Verantwortlichkeit	
4.1 Person	Verantwortlich für die Datenverarbeitung auf dieser Webseite ist die natürliche Person: PERSON_FIRST
4.2 Unternehmen / Firmierung	Verantwortlich für die Datenverarbeitung auf dieser Webseite ist die juristische Person: COMPANY_FIRST
4.3 Straße	Als Anschrift sind folgende Informationen hinterlegt: STREET_FIRST
4.4 Hausnummer	HOUSENUMBER_FIRST
4.5 Postleitzahl	POSTCODE_FIST
4.6 Stadt	CITY_FIRST
4.7 Telefon	Folgende Telefonnummer ist angegeben: PHONE_FIRST
4.8 Fax	Folgende Faxnummer ist angegeben: FAX_FIRST
4.9 E-Mail Adresse	Folgende E-Mail Adresse ist angegeben: EMAIL_FIRST

Tabelle 4.13: Übersicht Textbausteine

Umgesetzt und am Beispiel der Datenschutzerklärung der Universität Heidelberg ergibt dies folgendes Ergebnis:

Datenschutzerklärung
Generiert aus der Datenschutzerklärung von: uni-heidelberg.de

---

Es werden folgende Daten erhoben: Name, Anschrift, Betriebssystem, Datum, Uhrzeit, Cookie.
Es werden folgende Drittdienste verwendet: Matomo, Youtube.
Folgende Rechtsgrundlagen werden zur Verarbeitung von Daten angeführt: Art. 6 Abs. 1 lit. b DSGVO, Art. 6 Abs. 1 lit. c DSGVO, Art. 6 Abs. 1 lit. d DSGVO, Art. 6 Abs. 1 lit. f DSGVO, DSGVO, Art. 21 Abs. 1 DSGVO, Art. 6 Abs. 1 lit. a oder Art . 9 Abs. 2 lit. a DSGVO, Art. 21 Abs. 2 DSGVO, Art. 8 Abs. 1 DSGVO, Art. 17 Abs. 1 DSGVO, Art. 9 Abs. 2 lit. h und i sowie Art . 9 Abs. 3 DSGVO, Art. 89 Abs. 1 DSGVO
Verantwortlich für die Datenverarbeitung auf dieser Webseite ist die juristische Person: Universität Heidelberg.
Als Anschrift sind folgende Informationen hinterlegt:
Grabengasse 1
69117 Heidelberg

Folgende E-Mail-Adresse ist angegeben: rektor@rektorat.uni-heidelberg.de

Folgende Telefonnummer ist angegeben: 1 69117

Die vollständige Ausführung der in diesem beschriebenen Schritte benötigt für diese Datenschutzerklärung im implementierten Datensatz 1 Minute und 46 Sekunden.

Dabei wurden nicht alle Informationen korrekt erkannt. Zum Zeitpunkt des Abrufs der Datenschutzerklärung am 26.03.2021 ergibt sich folgende nach Abschnittsklassen gegliederte Differenz.

Datenerhebung: Browsertyp, IP-Adresse, Links (eingehend und ausgehend besuchte Webseiten).

Drittdienste: Openstreet Maps und Google wurde als Drittdienste nicht erkannt.

Telefonnummer: Es ist im Ursprungstext keine Telefonnummer angegeben, die hier angegebene Werte entstammen der Hausnummer und der Postleitzahl.

5 Zusammenfassung und Ausblick

Abschließend werden die Ergebnisse dieser Arbeit zusammengefasst und ein Ausblick auf weitere Aspekte gegeben, die in dieser Arbeit nicht Gegenstand war.

5.1 Zusammenfassung

Die Hauptthese dieser Arbeit beinhaltet, dass der Einbezug der Struktur einer Datenschutzerklärung die Informationsextraktion aus dem Text vereinfacht, weil so für jede Abschnittsklasse isoliert eine geeignete Methode evaluiert werden kann.

Um dies zu untersuchen, wurde ein Datensatz von 100 deutschsprachigen Datenschutzerklärungen angelegt und nach vier Abschnittsklassen, also Datenerhebung, Rechtsgrundlagen, Drittdienste und Verantwortlichkeit, annotiert. Es wurde gezeigt, dass der Ressourcen-günstigere regelbasierte Ansatz zur Klassifikation der Abschnitte vergleichbare Ergebnisse zu einer Support Vector Machine für die Klassen Drittdienste und Verantwortung erzielen kann und sogar bessere für Rechtsgrundlagen. Beide Ansätze konnten die Klassen zum Teil sogar nur am Titel des Abschnitts erkennen, ohne den eigentlichen Abschnittstext miteinzubeziehen. Innerhalb der Abschnittsklassen wurden 12 klassenspezifische Entitäten annotiert, anhand derer eine auf die jeweilige Abschnittsklasse passende Methode zur Informationsextraktion trainiert und getestet wurde. Für die Klassen der Drittdienste und der Datenerhebung konnte mit einem F1-score von jeweils 96,30% und 97,90% ein auf dem Datensatz selbst trainiertes Spacy-NER-Modell erfolgreich evaluiert werden. Für die Klasse der Verantwortlichkeit wurde eine Mischung aus einem regelbasierten und vortrainierten NER-Modell genutzt. Dabei konnte im Durchschnitt ein F1-score von 90,05% erzielt werden. Die Extraktion der Klasse Rechtsgrundlagen erfolgte mithilfe von einem NER-Modell auf Basis der Arbeit von Leitner [21], welches einen F1-score von 84,65% erreichte.

Der Vorteil der These wurde anhand des praktischen Beispiels der Textanreicherung zur Anmerkungsgenerierung demonstriert. Dabei wurden selbst erstellte Dokumente um die Informationen angereichert. So konnten die verwendeten Drittdienste zentral an einer

Stelle ausgegeben werden. Im Ursprungstext waren diese über verschiedene Stellen in der Datenschutzerklärung verstreut.

5.2 Ausblick

Der Datensatz von 100 deutschsprachigen Datenschutzerklärungen, der im Rahmen dieser Arbeit genutzt wurde, ist nur ein Bruchteil der existierenden, frei abrufbaren Datenschutzerklärungen. Deshalb ist es von größter Wichtigkeit, einen umfangreichen, repräsentativen Datensatz anzulegen und zu annotieren, um das Training der Algorithmen und damit die Extraktion der Informationen zu verbessern. Die Annotationen können durch weitere Abschnittsklassen und Entitäten ergänzt werden, damit ein größerer Anteil der Informationen extrahiert werden kann. Weiterhin könnten im Text verlinkte Daten durch die Zusammenführung mit dem Hintergrundwissen abgerufen werden. Beispielsweise könnten verlinkte Informationen aus anderen Dokumenten, wie dem Impressum, extrahiert werden. Zudem könnten mithilfe des Hintergrundwissens erweiterte nichtverlinkte Informationen, zum Beispiel die zuständige Aufsichtsbehörde in Abhängigkeit des Benutzerstandortes, angezeigt werden. Eine praktische Anwendung könnte eine Webseite sein, die sich an Laien richtet und nach Eingabe einer URL die dahinterliegende Datenschutzerklärung abruft, die Informationen extrahiert und in einer verständlichen Weise aufarbeitet. Danach könnten die relevanten Informationen dem Leser in Form von erklärenden Anmerkungen präsentiert werden.

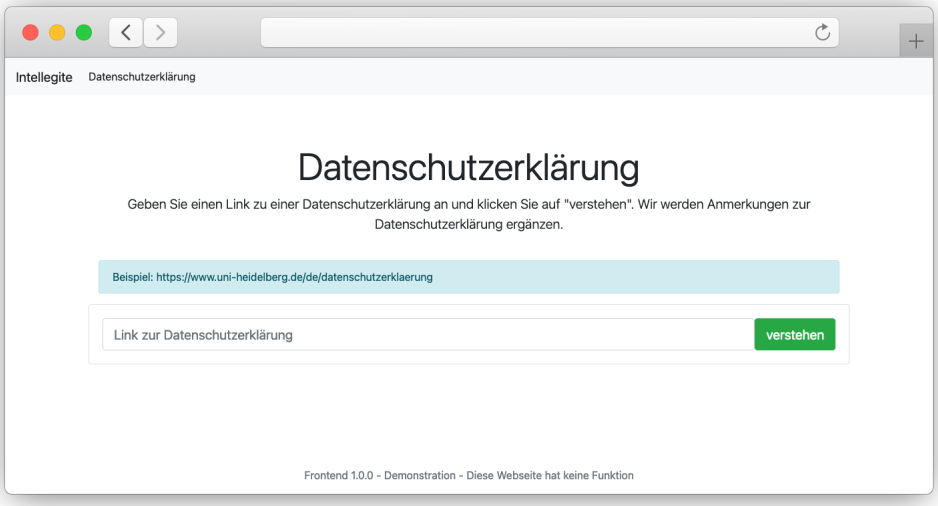


Abbildung 5.1: Darstellung einer Webseite, welche auf Grundlage der Arbeit erstellt werden könnte

Literaturverzeichnis

[1] *MUC4 '92: Proceedings of the 4th Conference on Message Understanding*, USA, 1992. Association for Computational Linguistics. dl.acm.org: doi:10.3115/1072064.1072067.

[2] M. Adedjouma, M. Sabetzadeh, and L. C. Briand. Automated detection and resolution of legal cross references: Approach and a study of Luxembourg's legislation. In *2014 IEEE 22nd International Requirements Engineering Conference (RE)*. IEEE, 2014. ieeexplore.ieee.org: doi:10.1109/re.2014.6912248.

[3] C. C. Aggarwal. *Machine Learning for Text*. Springer-Verlag GmbH, 2018. URL `https://www.ebook.de/de/product/33420020/charu_c_aggarwal_machine_learning_for_text.html`.

[4] I. Badji. Legal Entity Extraction with NER Systems. Master's thesis, Universidad Politécnica de Madrid, 2018. URL `http://oa.upm.es/51740/1/TFM_INES_BADJI.pdf`.

[5] A. Barrentine. Statistical NLP on OpenStreetMap. *machinelearnings.co*, 2016. URL `https://machinelearnings.co/statistical-nlp-on-openstreetmap-b9d573e6cc86`.

[6] C. Biagioli, E. Francesconi, A. Passerini, S. Montemagni, and C. Soria. Automatic semantics extraction in law documents. In *Proceedings of the 10th international conference on Artificial intelligence and law - ICAIL '05*. ACM Press, 2005. dl.acm.org: doi:10.1145/1165485.1165506.

[7] M. Bokaie Hosseini, P. K C, I. Reyes, and S. Egelman. Identifying and Classifying Third-party Entities in Natural Language Privacy Policies. In *Proceedings of the Second Workshop on Privacy in NLP*. Association for Computational Linguistics, 2020. aclweb.org: doi:10.18653/v1/2020.privatenlp-1.3.

[8] A. Bolioli, L. Dini, P. Mercatali, and F. Romano. For the Automated Mark-Up of Italian Legislative Texts in XML. *Knowledge and Information Systems - KAIS*,

2002. URL https://www.researchgate.net/publication/240926739_For_the_Automated_Mark-Up_of_Italian_Legislative_Texts_in_XML.

[9] F. H. Cate. The Limits of Notice and Choice. *IEEE Security & Privacy Magazine*, 8(2):59–62, 2010. ieeexplore.ieee.org: doi:10.1109/msp.2010.84.

[10] L. F. Cranor. *Web Privacy with P3p: The Platform for Privacy Preferences*. OREILLY MEDIA, 2002. URL https://www.ebook.de/de/product/3793894/lorrie_faith_cranor_web_privacy_with_p3p_the_platform_for_privacy_preferences.html.

[11] C. Dozier, R. Kondadadi, M. Light, A. Vachher, S. Veeramachaneni, and R. Wudali. Named Entity Recognition and Resolution in Legal Text. In *Semantic Processing of Legal Texts*, pages 27–43. Springer Berlin Heidelberg, 2010. link.springer.com: doi:10.1007/978-3-642-12837-0_2.

[12] N. Elssied, A. P. D. O. Ibrahim, and A. Hamza Osman. A Novel Feature Selection Based on One-Way ANOVA F-Test for E-Mail Spam Classification. *Research Journal of Applied Sciences, Engineering and Technology*, 7:625–638, 2014. maxwellsci.com: doi:10.19026/rjaset.7.299.

[13] L. Fu and G. Gali. *Classification Algorithm for Filtering E-mail Spams*, volume 157, pages 149–154. 2012. doi:10.1007/978-3-642-28798-5_21.

[14] M. Färber, B. Scheer, and F. Bartscherer. Who's Behind That Website? Classifying Websites by the Degree of Commercial Intent. In *Lecture Notes in Computer Science*, pages 130–145. Springer International Publishing, 2020. link.springer.com: doi:10.1007/978-3-030-50578-3_10.

[15] J. Gluck, F. Schaub, A. Friedman, H. Habib, N. Sadeh, L. Cranor, and Y. Agarwal. How Short Is Too Short? Implications of Length and Framing on the Effectiveness of Privacy Notices. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 321–340, Denver, CO, 2016. USENIX Association. URL https://www.usenix.org/conference/soups2016/technical-sessions/presentation/gluck.

[16] R. Grishman and B. Sundheim. Message Understanding Conference-6. In *Proceedings of the 16th conference on Computational linguistics* -. Association for Computational Linguistics, 1996. dl.acm.org: doi:10.3115/992628.992709.

[17] H. Harkous, K. Fawaz, K. G. Shin, and K. Aberer. PriBots: Conversational Privacy with Chatbots. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, Denver, CO, 2016. USENIX Association. URL https://www.usenix.org/conference/soups2016/workshop-program/wfpn/presentation/harkous.

[18] H. Harkous, K. Fawaz, R. Lebret, F. Schaub, K. G. Shin, and K. Aberer. Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. *CoRR*, abs/1802.02561, 2018, 1802.02561. URL http://arxiv.org/abs/1802.02561.

[19] S. E. Kettner, S. Ludwig, W. Bolte, F. Ingenrieth, C. Thorun, G. Heyer, S. Wolters, C. Rost, and J. Wittmann. PGuard Gemeinsamer Abschlussbericht. *datenschutz-scanner.de*, 2019. URL https://datenschutz-scanner.de/fileadmin/pguard/files/Abschlussbericht_PGuard_publikation.pdf.

[20] R. Kumar, M. Reddy, and P. Pappula. Text Classification Performance Analysis on Machine Learning. *MATTER: International Journal of Science and Technology*, 28:691–697, 2019. URL http://sersc.org/journals/index.php/IJAST/article/view/2900/2039.

[21] E. Leitner. Eigennamen- und Zitaterkennung in Rechtstexten. Master's thesis, Universität Potsdam, Potsdam, 2019. URL https://raw.githubusercontent.com/elenanereiss/Legal-Entity-Recognition/master/docs/Leitner_LER_BA.pdf.

[22] F. Liu, N. L. Fella, and K. Liao. Modeling Language Vagueness in Privacy Policies using Deep Neural Networks. *CoRR*, abs/1805.10393, 2018, 1805.10393. URL http://arxiv.org/abs/1805.10393.

[23] A. M. McDonald. The Cost of Reading Privacy Policies. *I/S: A Journal of Law and Policy for the Information Society, vol. 4, no. 3 (2008), 543-568.*, 2008. URL https://kb.osu.edu/bitstream/handle/1811/72839/ISJLP_V4N3_543.pdf?sequence=1&isAllowed=y.

[24] A. Niedermann. Freely-Given and Informed Consent? The User's Perspective. *Institut für Demoskopie Allensbach*, 2019.

[25] President's Council of Advisors on Science and Technology. Big data and privacy: A technological perspective. Report to the President. *Executive Office of the President*, 2014. URL https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf.

[26] N. Reimers and I. Gurevych. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 338–348, Copenhagen, Denmark, 2017. URL `http://aclweb.org/anthology/D17-1035`.

[27] J. VanderPlas. *Python Data Science Handbook*. O'Reilly UK Ltd., Oct. 2016. URL `https://www.ebook.de/de/product/24777490/jake_vanderplas_python_data_science_handbook.html`.

[28] V. Vinitha and K. R. Dhanaraj. Performance Analysis of E-Mail Spam Classification using different Machine Learning Techniques. pages 1–5, 2019. ieeexplore.ieee.org: doi:10.1109/ICACCE46606.2019.9080000.

[29] P. Świtalski and M. Kopówka. Machine Learning Methods in E-mail Spam Classification. *Studia Informatica*, pages 57–76, 2020. czasopisma.uph.edu.pl: doi:10.34739/si.2019.23.04.