Gottfried Wilhelm
Leibniz Universität Hannover
Fakultät für Elektrotechnik und Informatik
Institut für Praktische Informatik
Fachgebiet Software Engineering

Entwicklung eines Tools zur Erklärung von Datenschutzrichtlinien

Development of a tool to explain privacy policies

Bachelorarbeit

im Studiengang Informatik

von

Jonathan Thomczik

Prüfer: Prof. Dr. rer. nat. Kurt Schneider
Zweitprüfer: Dr. rer. nat. Jil Ann-Christin Klünder
Betreuer: M. Sc. Larissa Chazette und M. Sc. Wasja Brunotte

Hannover, 26.02.2021

Erklärung der Selbstständigkeit

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbständig und ohne fremde Hilfe verfasst und keine anderen als die in der Arbeit angegebenen Quellen und Hilfsmittel verwendet habe. Die Arbeit hat in gleicher oder ähnlicher Form noch keinem anderen Prüfungsamt vorgelegen.

Hannover, den 26.02.2021

Jonathan Thomczik

Zusammenfassung

Datenschutzerklärungen werden selten gelesen. Die Gründe hierfür lassen sich auf zwei Probleme zurückführen. Einerseits enthalten Datenschutzerklärungen viele Informationen, was einen Nutzer ohne technische Vorkenntnis überfordern kann (Information Overkill). Andererseits sind die Informationen für den Nutzer abstrakt und die negativen Folgen technischer Art kaum spürbar (mangelnde Fühlbarkeit). In der Folge ignoriert der Nutzer die Datenschutzerklärung oftmals und nimmt Schäden an seiner Privatsphäre in Kauf. Zur Lösung dieser Problematik wird in dieser Arbeit ein Tool entwickelt und implementiert, welches eine Datenschutzerklärung prägnant und anschaulich mittels Piktogrammen benutzerfreundlich zusammenfasst.

Als besonderer Vorteil wird dem Nutzer eine Hilfestellung angeboten, welche bei Bedarf den Nutzer durch einen einfachen Klick an die entsprechende Stelle der Datenschutzerklärung leitet, die weitere Informationen enthält. Das Tool trägt den Namen **Privacy Check**. Der Privacy Check benutzt Techniken aus dem Natural Language Processing und Information Retrieval, um eine Zusammenfassung aus einer Datenschutzerklärung zu generieren. Diese Techniken werden mit Python und spaCy umgesetzt. Da Python und spaCy nicht auf jedem Rechner vorhanden sind, wird der Privacy Check von einem Server unterstützt. Der Privacy Check selbst ist eine Webextension und kann in einen Browser geladen werden. Während der Entwicklung des Privacy Checks zeigen sich Probleme, welche am Ende der Arbeit diskutiert werden.

Abstract

Development of a tool to explain privacy policies

A user rarely reads privacy statements. The reasons for this can be traced back to two problems. On the one hand, data privacy statements contain much information, which can be overwhelming for users without prior technical knowledge (information overkill). On the other hand, the information is abstract for the user. The negative consequences of a technical nature are hardly perceptible (lack of perceptibility). As a result, the user often ignores the privacy statement and accepts damage to his privacy. A tool is developed and implemented to solve this problem, which summarizes a privacy statement concisely and picturesque through pictograms.

As a unique advantage, the user is offered a help function, which guides the user by a simple click to the corresponding part of the privacy statement, which contains further information. This tool is called **Privacy Check**. The Privacy Check uses Natural Language Processing and Information Retrieval techniques to generate a summary from a privacy statement. These techniques are implemented using Python and spaCy. Since Python and spaCy are not available on every computer, the Privacy Check is supported by a server. The Privacy Check itself is a web extension and can be loaded into a browser. During the privacy check development, some problems appear, which will be discussed at the end of the thesis.

Inhaltsverzeichnis

Kapitel 1

Einleitung

Nach Artikel (Art.) 12 Datenschutz-Grundverordnung[1] (DSGVO) soll die Kommunikation von Datenschutzrichtlinien zwischen einem Anbieter und einem Nutzer vereinfacht werden. Ein Unternehmen soll in prägnanter Form und in einfacher Sprache die Verarbeitung der personenbezogenen Daten kommunizieren. Die Kommunikation findet in der Regel über Datenschutzrichtlinien statt. Diese tauchen in Form einer sogenannten Datenschutzerklärung auf[2]. Die Datenschutzerklärung wird vom Nutzer selten gelesen. Auch kann der Nutzer den Datenschutz als lästig erachten, sofern der Nutzer kein Bewusstsein für seinen persönlichen Datenschutz hat. [24, S.11]. Dabei ist der Schutz der Daten eines Nutzers ein Grundrecht nach der Charta der Europäischen Union[3].

1.1 Problemstellung

Die Stiftung Datenschutz hat zwei zentrale Gründe für die zum Teil fehlende Kenntnisnahme durch Nutzer festgehalten [24, S.11]:

- **Information Overkill**
  Als Information Overkill beschreibt man die Tatsache, dass Nutzer durch Datenschutz überfordert werden. Sie benötigen einerseits Know-how, um die technischen Aspekte rund um den Datenschutz zu verstehen. Andererseits müssen sie über nötige Kenntnisse zum

---

[1] Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung).

[2] Nach Art. 13 und 14 der DSGVO.

[3] Charta: Charta der Grundrechter der Europäischen Union 2012/C 326/02, Amtsblatt der Europäischen Union, 26.10.2012, C 326/391; siehe auch Art. 16 abs.1 des Vertrags über die Arbeitsweise der Euroäischen Union (AEUV), Amtsblatt der europäischen Union, 26.10.2012, C 326/47.

Datenschutzrecht verfügen, um den Sachverhalt richtig einordnen zu können.

- **Fehlende Fühlbarkeit**
  Datenschutz ist für den Nutzer abstrakt. Diesem fehlt hierfür ein intuitives Verständnis. Die nachteiligen Folgen sind für den Nutzer kaum spürbar. Erschwerend kommt hinzu, dass die Konsequenzen von Datenschutzverstößen oft mit langer Verzögerung eintreten. Dieses Phänomen wird als fehlende Fühlbarkeit bezeichnet.

1.2 Motivation

Datenschutz fühlbarer zu machen und den Information Overkill zu verringern, soll Gegenstand dieser Arbeit werden. Dazu wird ein Tool entwickelt, das Datenschutzerklärungen auswerten kann. Der Nutzer soll da erreicht werden, wo er Datenschutzerklärungen häufig antrifft. Dies ist unter anderem im Internet der Fall. Dafür soll ein Tool entwickelt werden, das den Information Overkill verringert und den Datenschutz fühlbarer macht. Dieses Tool soll dem Information Overkill durch eine prägnante Zusammenfassung entgegensteuern. Zusätzlich soll der Zusammenfassung eine Fühlbarkeit gegeben werden. Als Anknüpfungspunkt in der Kommunikation zwischen Anbieter und Nutzer bietet sich der Sehsinn an, womit eine visuelle Unterstützung des Tools anzustreben ist. Konkret kann durch den Einsatz von Bildern eine Anschaulichkeit hergestellt werden, die sich der Fühlbarkeit zumindest annähert.

1.3 Lösungsansatz

Dieses Tool wird als Webextension für den Firefox entworfen. Die Webextension soll eine Datenschutzerklärung erfassen und auswerten. Der Nutzer bekommt eine Zusammenfassung durch die Webextension auf der Webseite mit der Datenschutzerklärung angezeigt. Zusätzlich soll die Möglichkeit geboten werden, dem Nutzer anzuzeigen, aus welchen Textabschnitten eine Information aus der Zusammenfassung stammt. Dazu soll eine Hilfestellung gegeben werden, wo die wichtigsten Fragen und Begriffe aus dem Datenschutz erklärt sind. Das Tool bekommt folgenden Namen: **Privacy Check**.

1.4 Struktur dieser Arbeit

In Kapitel 2 werden die Grundlagen für das Verständnis dieser Arbeit vermittelt und erläutert. Danach werden in Kapitel 3 ähnliche Arbeiten beleuchtet. Hiernach werden in Kapitel 4 die Anforderungen an den Privacy Checks erarbeitet und darauf aufbauend in Kapitel 5 das Konzept für diesen entwickelt. In Kapitel 6 wird die Implementation des Privacy Checks beschrieben. Anschließend wird sich in Kapitel 7 kritisch mit der Implementation auseinandersetzen. Zum Schluss wird in Kapitel 8 die Arbeit zusammengefasst und ein Ausblick gegeben.

Kapitel 2

Grundlagen

Dieses Kapitel erklärt die Grundlagen für das Verständnis dieser Arbeit. Anzuknüpfen ist dabei zunächst an der Datenschutzgrundverordnung (DSGVO). Da die zugrunde liegende Thematik sprachlicher Art ist, ist auf die Verarbeitung natürlicher Sprachen im Forschungsgebiet des Natural Language Processing einzugehen. Darüber hinaus zeigt sich in der Einleitung bereits die Notwendigkeit aus einer Fülle an Informationen eine Auswahl zu treffen, um den Information Overkill zu verhindern. Hierfür ist auf Pipes and Filters einzugehen. Abschließend ist eine wesentliche Anforderung in der Kommunikation an den Nutzer zu betrachten, die auch für diese Arbeit maßgeblich ist: die Explainability.

2.1 Einführung in den Datenschutz

Die DSGVO entfaltet als Rechtsverordnung unmittelbare Wirkungen in den Rechtsordnungen der Mitgliedsstaaten der Europäischen Union (EU). Nach Art. 2 DSGVO tritt diese in Kraft, wenn persönliche Daten automatisiert oder halbautomatisiert gespeichert werden. Daten, die eine Person identifizieren können sind persönliche Daten [23, S.13]. Automatisiert bedeutet, dass ein Computer ohne direktes menschliches Zutun Daten erhebt und speichert [23, S.11-12]. Halbautomatisiert bedeutet, dass ein Mensch die Daten in ein Computersystem eingibt [23, S.11-12]. Zusätzlich gilt die DSGVO nicht nur in der EU, sondern auch wenn jemand außerhalb der EU von einem EU-Bürger Daten erhebt [23, S.28-29].

2.1.1 Die Datenschutzerklärung

Unternehmen sind nach Art. 12, 13 und 14 DSGVO dazu verpflichtet, ihrer Transparenzpflicht nachzukommen. Dies geschieht unter anderem in Form einer Datenschutzerklärung. Diese Erklärung sollte einfach, klar und lesbar für einen Nutzer sein. Art. 13 legt für den Anbieter fest, dass dieser viele

Informationen über die Verarbeitung der persönlichen Daten kommunizieren soll.[1]. Ebenfalls ist ein Anbieter angehalten, dem Nutzer zu kommunizieren, welche Information dieser aus anderen Quellen bezieht.

2.1.2 Rechte eines Nutzers

Jedem Bürger der EU, von dem Daten erhoben werden, stehen Rechte nach der DSGVO zu. Diese werden im Folgenden kurz erläutert:
Nach den Art. 15, 16 und 17 DSGVO wird Betroffenen das Recht zugesichert, bei einem datenerhebenden Anbieter die persönlichen Daten zu erfragen. Zusätzlich darf der Betroffene jederzeit das Unternehmen auffordern, seine Daten zu korrigieren oder zu löschen.

Nach den Art. 18, 19 und 20 DSGVO wird dem Nutzer das Recht eingeräumt, die Verarbeitung seiner Daten einschränken zu lassen. Zudem ist ein Anbieter verpflichtet, einen Nutzer zu informieren, wenn er die Daten des Nutzers berichtigt oder löscht. Hinzukommend hat ein Nutzer das Recht, seine Daten von einem Anbieter zu bekommen und an einen anderen Anbieter weiterzuleiten.

Nach den Art. 21, 22 und 23 DSGVO hat ein Nutzer das Recht, seine personenbezogenen Daten vom Anbieter in einem maschinenlesbaren Format zu bekommen. Dem Nutzer soll somit zugesichert werden, dass er seine Daten an Andere übertragen kann. Außerdem darf der Nutzer unter gewissen Umständen die Verarbeitung seiner Daten komplett widerrufen. Zusätzlich hat der Nutzer das Recht, nicht von automatisierten Entscheidungen unterworfen zu werden. Dies bedeutet für den Nutzer, dass ein Computer nicht anhand der persönlichen Daten über einen Menschen entscheiden darf, wenn sich dies für den Nutzer nachteilig auswirken würde.

Nach Art. 77 DSGVO hat jeder Betroffene das Recht, bei einer Behörde Hilfe einzuholen, wenn der Nutzer vermutet, dass seine persönlichen Daten vom Anbieter missbraucht werden.

2.2 Einführung in Natural Language Processing

Das Natural Language Processing (NLP) ist ein Forschungsgebiet, das sich mit der Verarbeitung natürlicher Sprachen wie Deutsch oder Englisch befasst [12]. Damit eine Datenschutzerklärung dem Nutzer kommuniziert werden kann, muss diese Sprache verarbeitet werden. Dafür finden ein paar

---

[1] Eine mögliche Zusammenfassung des Art. 13 würde über den Rahmen dieses Kapitels und der Arbeit hinausgehen.

Methoden aus dem NLP in dieser Arbeit Anwendung. Diese werden im Folgenden erläutert:

2.2.1 Tokenization

Damit ein Satz maschinell betrachtet werden kann, muss er zunächst in Wörter aufgeteilt werden. Das Verfahren wird Tokenization genannt [25].

Beispiel: *In der KFZ-Zulassungstelle muss man lange warten.*
Die Token wären dann:
(in) (der) (KFZ-Zulassungstelle) (muss) (man) (lange) (warten) (.)

2.2.2 Lemmatization

Wörter können je nach Kontext und Zeitform eine unterschiedliche Schreibweise haben. Die Aufgabe der Lemmatization ist es, ein Wort in seine Ursprungsform zu bringen [13].

Beispiel: *Ich laufe die Treppe hoch. Du liefst die Treppe hoch.*

Würde man beispielsweise das Wort „liefst“ betrachten, so wäre die Lemmatization „laufen“.

2.2.3 Part of Speech Tagging

Für die Verarbeitung einer natürlichen Sprache ist die Benennung der Wortart, das Part of Speech Tagging (POS), hilfreich. [1]
Beispiel: *Das Auto ist grün.*

- Das: Artikel
- Auto: Nomen
- ist: Hilfsverb
- grün: Adjektiv

2.2.4 Namend Entity Recognition

Ein weiterer Bestandteil der Verarbeitung der natürlichen Sprache ist die Named Entity Recognition (NER). Sie beschreibt die Erkennenung und die Einordnung von Eigennamen in einem Textabschnitt [15]. Jedem Eigennamen wird eine Beschreibung zugeordnet. Das NER gibt dem Computer die Möglichkeit, Texte besser zu klassifizieren oder nach bestimmten Begriffen zu suchen, die zu einer Gruppe gehören [15].
Beispiel: *Volkswagen baut Autos in Wolfsburg.*

- Volkswagen: Organisation
- Auto: Produkt
- Wolfsburg: Ort

2.3 Term Frequency and Inverse Document Frequency Maßstab

Für die Auswertung einer Datenschutzerklärung muss nach Begriffen gesucht werden. Das Gebiet des Information Retrival befasst sich mit der Suche von Begriffen in Texten. Dieses Gebiet benutzt dafür unter anderem zwei Maßstäbe, den Term Frequency (TF) und Inverse Document Frequency (IDF). TF und IDF messen die Wichtigkeit eines Wortes in einem Text [26]. Beide Maßstäbe gehen dabei gegensätzlich vor. Während der TF die Häufigkeit misst, um die Wichtigkeit zu bestimmen, ermittelt der IDF konträr dazu, wie selten ein Wort vorkommt und schließt daraus auf die Wichtigkeit [26]. Die Kombination beider Maßstäbe in Form einer Multiplikation, bekannt als TF*IDF-Maßstab, liefert nach Robertson [19] robuste Ergebnisse.

2.4 Pipes and Filters

Datenschutzerklärungen sind umfangreich. Um diese prägnant zusammenzufassen, müssen bestimmte Informationen gefiltert werden. Dafür bietet sich die Architektur Pipes and Filters an. Filters reduzieren Informationen und Pipes leiten diese an andere Filter weiter [21, S.44-45].
In der Grundstruktur zeigen sich in der Architektur bereits Grenzen. Durch Vorgaben der Pipes-and-Filters-Architektur, kennen Filter nur eine Eingabe und Ausgabe [21, S.44-45]. Zusätzlich ist einem Filter die Arbeitsweise des vorherigen Filters und des nachfolgenden Filters nicht bekannt. Dies ist nicht zufriedenstellend, da die Ergebnisse zweier Filter nicht zusammengeführt werden können [21, S.44-45].

Diesen Schwierigkeiten begegnend ermöglicht eine modifizierte Version der Pipes and Filters, die sogenannte Tee-and-join-Pipe, eine Architektur umzusetzen, bei der ein Filter mehrere Ein- und Ausgaben benutzt [5, S.67].

2.5 Explainability

Nach Chazette et al. [6] beschreibt Explainability die Fähigkeit einer Software Erklärungen zu geben. Darunter fällt einerseits die Fähigkeit, Informationen an einen Menschen zu vermitteln. Anderseits beschreibt

Explainability die Fähigkeit der Software ihre Entscheidungsprozesse an einen Nutzer transparent zu vermitteln. Der Nutzer verliert Vertrauen zur Software, wenn dieser die Entscheidungsprozesse des Programms nicht versteht. Werden die Entscheidungsprozesse jedoch kommuniziert, so kann das Vertrauen des Nutzers wieder angehoben werden.

2.6 Privacy Bots

Nach Kettner et al. [11, S.71-73] sind Privacy Bots Programme, die bestehende Datenschutzerklärungen analysieren und für einen Nutzer auswerten. Diese weisen Ähnlichkeiten zu einem Werbeblocker auf. Privacy Bots treten entweder als unterstützendes Programm oder als Addon für einen Browser auf.

Kapitel 3

Verwandte Arbeiten

Die Untersuchung der vereinfachten Kommunikation von Datenschutzerklärungen ist Gegenstand verwandter Arbeiten. In der Recherche zeigen sich zwei Ausprägungen:

1. Arbeiten, die auf der Anbieterseite ansetzen
2. Arbeiten, die stattdessen auf der Nutzerseite ansetzen

Für den Ansatz auf der Anbieterseite erklären beispielsweise Kettner et al. [11, S.18-20] in ihrer Arbeit, *Wege zur besseren Informiertheit*, den One Pager Matrix Ansatz. Die One Pager Matrix dient dazu, Datenschutzerklärungen zu vereinfachen. Die Vereinfachung soll auf eine DIN-A4 Seite passen. Dabei sollen folgende Fragen beantwortet werden:

1. Welche Daten werden erfasst?
2. Auf welche Weise werden Daten erhoben?
3. Wofür werden die persönliche Daten genutzt?
4. Welche Rechte hat der Betroffene?
5. Welche Personen können bei Fragen angesprochen werden?

Der Onlinedienst Zalando hat diesbezüglich ein Tool entwickelt, das eine solche One Pager Matrix erstellt. Ob dieser Dienst weiterentwickelt wird, kann in dieser Arbeit nicht ermittelt werden, da die Webpräsenz zum Erstellen einer One Pager Matrix nicht erreichbar ist[1].

Kettner et al. [11, S.40-41] untersuchen zudem, wie hoch die Lesewahrscheinlichkeit für die One Pager Matrix gegenüber einer normalen Datenschutzerklärung ist und ob der Nutzer den Inhalt der One Pager Matrix versteht. Es wird festgestellt, dass zwar die

---

[1] https://geta1pager.de/ (zuletzt abgerufen am 15.02.2021).

Lesewahrscheinlichkeit einer One Pager Matrix höher ist, jedoch der Nutzer trotzdem Verständnisschwierigkeiten hat. Zudem kann nicht festgestellt werden, ob es auch einen Vertrauensgewinn in den Anbieter gibt.

Einen anderen Ansatz Datenschutzerklärungen und deren Kommunikation auf Anbieterseite zu vereinfachen, versucht das Platform for Privacy Projekt (P3P). Dieses ist ein freiwilliger Standard für Privacy Policies. Dieser Standard wurde von dem World Wide Web Consortium (W3C) veröffentlicht[2]. Nach Cranor [9] spezifiziert der P3P Standard die Erfassung von Daten in der Extensible Markup Language (XML). Ein Betreiber gibt in dieser XML-Datei nach dem P3P-Data Schema die von ihm erhobenen Daten an. Das P3P Data Schema wiederum spezifiziert die Angaben von erhobenen Daten. Zusätzlich gibt das Schema den Speicherort der XML-Datei vor. Dieser ist unter dem Pfad /w3c/p3p.xml zu finden. Ziel dieser Spezifikation ist es, dass ein Nutzer die Privacy Policies nicht mehr lesen muss. Ergänzend dazu soll ein Internetbrowser die Privacy Policy auswerten können und dem Nutzer aufbereitet vorlegen. Der Internet Explorer 7 und der Netscape Navigator 7.0 konnten diesen Standard interpretieren[3]. Der Standard wurde im April 2002 veröffentlicht und im August 2018 als veraltet erklärt[4].

Die zweite Ausprägung setzt stattdessen auf der Nutzerseite an und ist als Privacy Bot ausgestaltet. Tomuro et al. [22] haben beispielweise in ihrer Arbeit *Automatic Summarization of Privacy Policies using Ensemble Learning* einen Privacy Bot entwickelt, der Datenschutzerklärung auswerten kann. Dieser Privacy Bot ist über einen Webservice implementiert und kann durch einen Link abgerufen werden. In dieser Arbeit konnte der Privacy Bot nicht getestet werden, da der angegebenen Link[5] nicht zu erreichen war. Tomuro et al. [22] kategorisieren Informationen, die für einen Nutzer relevant sein könnten. Die Informationen sind nach folgenden Fragen kategorisiert:

- Für welche Zwecke benutzt ein Anbieter die persönlichen Daten?
- Teilt der Anbieter seine Daten mit Dritten?
- Holt sich der Anbieter Daten über den Nutzer aus Drittquellen?
- Wird der Anbieter die persönlichen Daten verkaufen?
- Kann der Anbieter Daten dauerhaft speichern?

Die Zusammenfassung beinhaltet die Kategorien mit den jeweiligen Informationen, die ein Anbieter in seiner Datenschutzerklärung preisgibt.

---

[2] https://www.w3.org/

[3] https://www.w3.org/P3P/implementations.html (zuletzt abgerufen am 13.02.2021).

[4] https://www.w3.org/TR/P3P11/ (zuletzt abgerufen am 13.02.2021).

[5] http://slytinen-ntomuro.rhcloud.com/index.jsp (zuletzt abgerufen am 11.02.2021).

Tomuro et al. [22] benutzten Machine Learning Ansätze, um eine Datenschutzerklärung auszuwerten.

Ein weiterer Ansatz aufseiten des Nutzers, ist der sogenannte Privacy Bird. Dieser ist ein Projekt der W3C Gruppe [6]. Nach Cranor et al. [10] hat sich der Privacy Bird zur Aufgabe gemacht, in P3P-Standard verfasste Privacy Policies auszuwerten und eine Zusammenfassung zu liefern. Der Privacy Bird ist in der Lage, bei Webseitenaufruf eine Datenschuterklärung auszuwerten. Dafür lädt der Privacy Bird die nach P3P standardisierte XML-Datei des Anbieters. Der Privacy Bird kann Einstellungen nach den persönlichen Präferenzen vornehmen. Ist dem Nutzer zum Beispiel wichtig, dass seine Daten nicht an Dritte weitergeleitet werden, so kann der Privacy Bird beim Aufrufen der Webseite den Nutzer warnen. Der Privacy Bird wurde in der Betaversion 1.3 veröffentlicht. Er kann für den Internet Explorer 5.0, 5.6 und 6.0 benutzten werden.

Eine Weiterentwicklung eines Privacy Bots konzipieren Nüske et al.[7] [17]. In der Konzeption findet sich eine umfassende Diskussion der Auswertung einer Datenschutzerklärung. Da die Auswertung einen wesentlichen Teilschritt dieser Arbeit darstellt, werden die verschiedenen Möglichkeiten der Auswertung einer Datenschutzerklärung im Folgenden ausführlicher behandelt. Dieses Konzept sieht zunächst einen Privacy Bot vor, der vier Fragen für einen Nutzer beantworten soll, ähnlich zu der One Pager Matrix:

1. Welche Daten werden vom Anbieter gesammelt?
2. Welche Technologien benutzt der Anbieter?
3. Wofür benutzt der Anbieter diese Daten?
4. Wohin leitet der Anbieter Daten weiter?

Zusätzlich sieht das Konzept ein sogenanntes Datenschutzprofil vor. In diesem Profil kann der Nutzer festlegen, welche Daten dieser bereit ist, Preiszugeben und welche nicht. Der Privacy Bot kann dann anhand des Datenschutzprofiles eine Datenschutzerklärung auswerten und dem Nutzer Auskunft darüber geben.

Nach dem Konzept ergeben sich zwei Optionen, eine Datenschutzerklärung in den Privacy Bot zu übertragen. Entweder soll der Nutzer den Link zur Datenschutzerklärung in einem Webservice eingeben oder die

---

[6] http://www.privacyfinder.org/

[7] Nüske et al. gewannen 2017 ein Preis für einen Privacy Bot von der Deutschen Telekom AG. Ob der beschriebene Privacy Bot der Privacy Bot ist, mit denen Nüske et al. einen Preisge wannenn, ist unklar. https://www.telekom.com/de/verantwortung/datenschutz-und-datensicherheit/archiv-datenschutznews/news/privacy-bots-telekom-praemiert-ideen-499988 abgerufen am 16.02.2021

Datenschutzerklärung wird automatisiert durch ein Addon ausgewertet.
In dem Konzept wird diskutiert, wie die Datenschutzerklärungen ausgewertet werden können. Dazu können die Datenschutzerklärung manuell über Experten oder durch einen Crowd-Ansatz ausgewertet werden. Die Auswertung durch Experten bietet den Vorteil einer hohen Qualität in der Auswertung. Der Nachteil ist, dass es sehr aufwendig und teuer wäre, dies umzusetzen. Alternativ könnte ein Crowd-Ansatz gewählt werden, bei dem Nutzer dem Privacy Bot die nötigen Informationen liefern. Bevor der Nutzer Funktionen des Privacy Bots nutzen kann, wird dieser gebeten, eine Datenschutzerklärung auszuwerten. Die Qualität des Ansatzes soll durch eine große Anzahl von Nutzern sichergestellt sein und die Inhalte durch eine Mehrheitsmeinung festgelegt werden. Durch dieses Vorgehen kann in der Anfangsphase des Bots nur für wenige Webseiten eine Auswertung geboten werden.
Eine andere Möglichkeit der Auswertung könnte durch Schnittstellen zwischen Anbieter und dem Privacy Bot gegeben sein. Der Vorteil dieser Vorgehensweise ist, dass der Privacy Bot automatisch Einstellungen der Nutzerpräferenzen über die Schnittstelle beim Anbieter vornehmen kann.
Alternativ könnte auf die algorithmische Verarbeitung mittels Textmining zurückgegriffen werden. Das Konzept sieht vor, einen großen Trainingssatz mit gängigen Formulierungen aus Datenschutzerklärungen zu erstellen. Anhand dessen sollen Algorithmen trainiert werden. Als Beispiele für Algorithmen werden das Vektorenverfahren oder die hierarchische Cluster Analyse genannt. Vorteil algorithmischer Auswertung ist die schnelle Verarbeitung. Der Nachteil liegt darin, dass hohe Implementierungskosten und ein ausgeprägtes Know-how für die Implementierung notwendig sind.
In dieser Arbeit konnte nicht festgestellt werden, ob das Konzept umgesetzt wird.

Die vorliegenden Ausführungen zeigen, dass die Zielsetzung dieser Arbeit, Datenschutzerklärungen prägnant und anschaulich zusammenzufassen, bisher nicht zufriedenstellend erreicht wird. Insbesondere werden folgende Probleme nicht gelöst: Die Kommunikation, die sich auf die Anbieterseite bezieht, bleibt immer abhängig vom Anbieter. Der Nutzer bleibt darauf angewiesen, dass der Anbieter verständlich kommuniziert. Die Privacy Bots setzten demgegenüber beim Nutzer an und lösen diese Abhängigkeit. Die betrachteten Privacy Bots sind im Kern mit verschiedenen Problemen behaftet. Der hier entwickelte Privacy Check versucht diese Probleme zu lösen. Im Einzelnen:

- **Benutzerunfreundlichkeit des Privacy Bots**
  Der Privacy Bot nach Tumoro et al. [22] wurde als Webservice umgesetzt. Will der Nutzer eine Datenschutzerklärung zusammengefasst haben, so muss er bei dem Privacy Bot den Link kopieren und

auf der Webseite des Privacy Bots einfügen. Das kann für einen Nutzer zu umständlich sein. Hingegen stellt der Privacy Check eine benutzerfreundliche Alternative dar. Der Nutzer muss lediglich die Webextension öffnen und die Zusammenfassung durch einen Knopfdruck anfordern.

- **Mangelnde Anschaulichkeit der Zusammenfassung**
  Ein Privacy Bot sollte dem Information Overkill entgegen wirken und die Auswertung einer Datenschutzerklärung ansprechend gestalten. Wird dies nicht umgesetzt, so kann der Nutzer sein Interesse am Schutz seiner Daten verlieren. Sowohl der Privacy Bot nach Tumoro et al. [22] als auch der Privacy Bird geben eine Auswertung in reiner Textform an. Durch diese reine Textform wird die Explainability der beiden Bots verringert, da der Nutzer durch viel Text überfordert wird. Der Privacy Check hingegen benutzt in seiner Auswertung Piktogramme als Unterstützung und ist somit für einen Nutzer ansprechender.

- **Mangelnde Prägnanz der Zusammenfassung**
  Bei den genannten Privacy Bots kann auch ein Information Overkill auf andere Weise stattfinden. Die Auswertung erfordet ein technisches Know-how. Im Falle des Privacy Bots nach Nüske et al. [17] wird der Nutzer informiert, welche Technologien der Anbieter verwendet, um Daten zu erheben. Hat der Nutzer kein ausgeprägtes Verständnis für Technologien zur Erhebung von Daten, kann ihn dies unter Umständen überfordern. Der Privacy Check verfolgt einen anderen Ansatz. Dem Nutzer wird in der Zusammenfassung nur vermittelt, welche persönlichen Daten erhoben werden und welche weiteren Drittparteien persönliche Daten erheben. Bei weiterem Klärungsbedarf kann der Nutzer die Hilfestellung des Privacy Checks nutzen, um in die entsprechenden Passagen einer Datenschutzerklärung zu gelangen. Der Privacy Check kann somit den Information Overkill reduzieren, aber bei Bedarf eine bessere Explainbility durch die Hilfestellung beim Lesen einer Datenschutzerklärung gewährleisten. Dies ist ein Alleinstellungsmerkal des Privacy Checks. Alle betrachteten Privacy Bots vereinfachen hingegen die Datenschutzerklärungen.

Kapitel 4

Anforderungen

In diesem Kapitel werden die Anforderungen an den Privacy Check festgelegt. Diese Anforderungen werden mit den Betreuern dieser Arbeit abgestimmt.

Die erste Anforderung bezieht sich auf den Zugang zum Datenschutz.
Da Nutzer über unterschiedliche technische Fähigkeiten und oftmals geringen Datenschutzkenntnissen verfügen, sollte der Zugang zum Privacy Check erleichtert werden. Dies kann erreicht werden, wenn der Information Overkill verringert und dem Datenschutz eine ansprechende Form gegeben wird Abschnitt 1.1). Dazu soll dem Nutzer eine prägnante und anschauliche Zusammenfassung der Datenschutzerklärung geboten werden. Diese Zusammenfassung soll aus einer Liste der erhobenen Daten und den Drittparteien, die persönliche Daten eines Nutzers vom Anbieter bekommen, bestehen. Zudem wird jeder Eintrag in der Liste durch ein Piktogramm unterstützt. Nach Barr et al. [3] sind Piktogramme ein Schlüsselelement für benutzerfreundliche Oberflächen. Zudem kann nach Makini et al. [14] die Verbindung von Wörtern mit Bildern einen Lerneffekt verursachen. Dieses Vorgehen kann den Information Overkill verringern und den Datenschutz für den Nutzer fühlbarer machen

Hinzukommend soll dem Nutzer eine Hilfestellung zum Lesen einer Datenschutzerklärung gegeben werden. Diese Hilfestellung soll dem Nutzer gängige Begriffe und Rechte zum Datenschutz erklären. Dadurch wird die Explainability einer Datenschutzerklärung und des Privacy Checks verbessert. Hinzukommend soll die Hilfestellung Abschnitte in der Datenschutzerklärung markieren. Die Markierungen der Abschnitte in der Datenschutzerklärung sollen aus der Zusammenfassung generiert werden. Dem Nutzer soll hiermit ermöglicht werden, nachzuvollziehen, aus welchen Abschnitten der Datenschutzerklärung die Zusammenfassung die Informationen bezieht. Damit wird die Transparenz des Privacy Checks gesteigert. Durch diese kann das Vertrauen des Nutzers und somit auch die Explainability des Privacy Checks gesteigert werden (vgl. Abschnitt 2.5). Hinzukommend kann beim Nutzer ein Lerneffekt durch die Erklärung der Begriffe, Fragen und Rechte zum Datenschutz eintreten.

Der Privacy Check sollte leicht bedienbar sein, da ansonsten für Nutzer mit unter anderem wenig ausgeprägten technischen Kenntnissen, der Zugang zum Datenschutz erschwert wird. Der Nutzer soll zudem nicht an seinem gewohnten Surfverhalten gehindert werden. Würde der Privacy Check in das Surfverhalten des Nutzers eingreifen, so könnte dieser abgeschreckt werden. Zusammenfassend soll das Programm folgendes können:

1. **Prägnante und anschauliche Zusammenfassung**
    - Zusammenfassung erhobener Daten
    - Zusammenfassung der an Dritte weitergeleiteten Daten
2. **Hilfestellung zur Erklärung**
    - Erklärung gängiger Begriffe aus dem Datenschutz
    - Erklärung von Datenschutzrechten eines Nutzers
    - Hinweis auf den Abschnitt, auf den sich die Zusammenfassung bezieht
3. **Benutzerfreundlichkeit in der Bedienung**
    - Leichte Bedienung des Privacy Checks
    - Keine Behinderung des Nutzers in seinem Surfverhalten

Kapitel 5

Konzeptentwicklung

Dieses Kapitel entwickelt das Konzept des Privacy Checks anhand der Anforderungen aus Kapitel 4. Dafür ist zunächst ausführlich auf die Entwicklung einer prägnanten Zusammenfassung und daran anschließend auf die Entwicklung einer anschaulichen Zusammenfassung einzugehen. Des Weiteren wird erläutert, wie die Hilfestellung zu entwickeln ist und wie Benutzerfreundlichkeit in der Bedienung erreicht wird.

5.1 Entwicklung der prägnanten Zusammenfassung

Zu Beginn der Erläuterung der Entwicklung einer prägnanten Zusammenfassung ist zunächst kurz zu skizzieren, warum die Prägnanz notwendig ist und welche Informationen kommuniziert werden sollen. Anschließend wird erläutert wie der Zusammenfassungsalgorithmus entwickelt wird. Dafür werden zunächst nötige Voarbeiten durchgeführt und daraufhin der Algorithmus erklärt.

5.1.1 Notwendigkeit und Ausgestaltung der Zusammenfassung

Breite und Tiefe der zu übermittelnden Informationen nach DSGVO über die zu erhebenden Daten und der Übermittlung an Drittparteien (vgl. Unterabschnitt 2.1.1 können zu einem Information Overkill führen. Aus diesem Grund müssen die zu kommunizierenden Informationen im Privacy Check zusammengefasst werden. Zwingend notwendig für den Nutzer ist nur zu wissen, welche Daten gespeichert werden und an wen die Übermittlung erfolgt, da alle weiteren Informationen dem Nutzer ein ausgeprägtes Know-how in Datenschutz und Technik voraussetzen.

5.1.2 Vorarbeit zur Entwicklung des Zusammenfassungsalgorithmus

Datenschutzerklärungen sollen vom Privacy Check kurz und prägnant zusammengefasst werden (vgl. Kapitel 4). Der Privacy Check soll dazu mittels eines Algorithmus die relevanten Informationen für die Datenschutzerklärung filtern. Damit dies gelingt, werden Datenschutzerklärungen untersucht, um Muster festzustellen, die eine Filterung vereinfachen. Daher werden für diese Arbeit 50 Datenschutzerklärungen betrachtet (vgl. Anhang A). Die Stichprobe der Datenschutzerklärungen wird wie folgt erhoben: Es wird der Tor Browser[1] mit der Suchmaschiene DuckDuckGo [2] benutzt. Dies soll verhindern, dass das Suchverhalten Einfluss auf die Ergebnisse hat. Assoziativ werden Suchbegriffe eingegeben und Webseiten geöffnet. Sinn dieser Vorgehensweise ist es, die Zufälligkeit der Stichprobe zu erhöhen.
Zum Erkennen von Mustern sind folgende Untersuchungsfragen leitend:

1. Führen die Betreiber eine Liste von erhobenen Daten?
2. Kommunizieren die Betreiber eine Erfassung der Daten auch oder ausschließlich über einen Fließtext?
3. Wird eine Widerrufsmöglichkeit für Cookies angegeben?
4. Gibt es Hinweise auf eine automatisierte Wiederufsmöglichkeit der verarbeiteten Daten?

Dabei wird vernachlässigt, ob die Betreiber korrekt ihrer Transparenzpflicht nachkommen (vgl. Unterabschnitt 2.1.1). Zusätzlich wird vernachlässigt, ob die Datenschutzerklärung DSGVO-konform ist.
Folgende Muster werden beobachtet:

1. 74% der Betreiber führen eine Liste der erfassten Daten.
2. 64% der Betreiber kommunizieren über einen Fließtext die Erfassung von Daten.
3. 4% der Betreiber haben eine Wiederufsmöglichkeit für Cookies angegeben.
4. 10% der Betreiber bieten dem Nutzer eine technische Möglichkeit, seine Daten zum Teil oder komplett zu wiederrufen.

In dieser Untersuchung wird festgestellt, dass Datenschutzerklärungen folgende Muster aufweisen: Datenschutzerklärungen unterteilen sich in Sinnabschnitte.

---

[1] Der Tor Browser anonymisiert seine Nutzer und deren Surfverhalten.

[2] Diese Suchmaschiene wird nicht vom Suchverhalten eines Nutzers beeinflusst.

Zudem wurde beobachtet, dass die Auflistung auch Fließtexte beinhalten können.
Ebenfalls kann in der Untersuchung keine semantische Regel festgestellt werden, die auf die Speicherung von Daten hinweist. Zusätzlich wurde beobachtet, dass nicht nur erfasste Daten aufgelistet werden, sondern auch andere datenschutzrechtliche Sachverhalte, wie zum Beispiel eine Aufzählung der Rechte bei einer Rechtsbelehrung.

Folgerungen aus der Untersuchung: Infolge des Zeitrahmens wird auf einen maschinellen Lernansatz verzichtet. Für die Suche nach relevanten Begriffen wird auf eine Technik des Information Retrieval zurückgegriffen (vgl. Abschnitt 2.3). Mittels NLP wird die Datenschutzerklärung ausgewertet.

5.1.3 Entwicklung des Zusammenfassungsalgorithmus

Nach den Anforderungen aus Kapitel 4 erstellt der Privacy Check eine Zusammenfassung einer Datenschutzerklärung. Daher müssen Begriffe und Sätze aus einer Datenschutzerklärung gefiltert werden, die von einer Erhebung der Daten sprechen. Diese Begriffe und Sätze werden im Folgenden **Phrasen** genannt. Dazukommend sollen in der Zusammenfassung alle Drittparteien genannt werden, die Daten durch den Anbieter erheben. Damit zählen auch erwähnte Drittparteien zu Phrasen.

Jeder Sinnabschnitt kann Phrasen enthalten, nach denen gesucht wird. Um die Wichtigkeit des Sinnabschnittes zu bemessen wird der TF*IDF-Maßstab verwendet (vgl. Abschnitt 2.3). Die gesuchten Phrasen sind Nomen oder Eigennamen. Dafür wird aus dem NLP das Part of Speach Tagging verwendet, um die richtigen Phrasen zu klassifizieren (vgl. Unterabschnitt 2.2.3). Bevor jedes Wort den richtigen Tag erhalten kann, müssen zuerst die einzelnen Phrasen aus dem Sinnabschnitt extrahiert werden. Dazu wird die Technik der Tokenization aus dem NLP benutzt (vgl. Unterabschnitt 2.2.1). Sind alle Wörter in Tokens umgewandelt und haben einen Tag erhalten, kann die Relevanz eines Sinnabschnittes bestimmt werden.

Für jeden Sinnabschnitt wird der durchschnittliche TF*IDF-Maßstab über alle Eigennamen oder Nomen bestimmt. Dies dient der Vorfilterung.

Während der Untersuchung von Datenschutzrichtlinien werden Phrasen in einer Liste gespeichert. Diese Liste unterstützt die weitere Filterung. Die Phrasen sind einzelne Sätze, die die Erhebung von Daten erwähnen. Wenn eine Phrase in einem Sinnabschnitt vorkommt, soll der TF*IDF-Wert des Sinnabschnittes erhöht werden.
Jeder Sinnabschnitt, der über dem Durchschnitt liegt, wird als wichtig

bezeichnet. Die Sinnabschnitte, die unter dem Durchschnitt liegen, werden vernachlässigt.

Aus den Untersuchungen aus Unterabschnitt 5.1.2 ergibt sich, dass 74% der Betreiber eine Liste der erhobenen Daten führen. Diese Tatsache soll der Algorithmus nutzen, indem ein Sinnabschnitt, der eine Liste enthält, in die Zusammenfassung aufgenommen wird.

Das Zustandekommen des Algorithmus und dessen Probleme werden in Unterabschnitt 7.1.3 erläutert.

5.2 Entwicklung der anschaulichen Zusammenfassung

Datenschutzrichtlinen sind abstrakt und nicht fühlbar (vgl. Abschnitt 1.1). Um dem Nutzer einen leichteren Zugang zum Datenschutz zu geben, muss die Zusammenfassung ansprechend gestaltet werden. Nach Kettner et al. [11, S.64] bieten Piktogramme die Möglichkeit in vereinfachter Form über die Erhebung von Daten zu kommunizieren. Daher soll der Privacy Check Piktogramme benutzen, um die Anschaulichkeit zu gewährleisten. Diese Piktogramme werden unter anderem in der Zusammenfassung genutzt. In der Zusammenfassung wird jedem Element ein Piktogramm zu geordnet. Die Piktogramme sollen möglichst kommunizieren, welche persönliche Daten erhoben werden. Dies soll wie im folgenden Beispiel geschehen: Die Abbildung 5.1[3] zeigt eine Weltkugel, auf der ein Ort markiert ist. In Verbindung mit dem Begriff Internet Protokoll Adresse (IP-Adresse) kann der Nutzer eine Assoziation zwischen dieser und seinem Standort schaffen. Dies steigert die Explainability.

Abbildung 5.1: Beispiel eines Piktogramms

---

[3] Piktogramm erstellt von Anu Rocks und entnommen aus https://freeicons.io/social-media-icons/location-icon-12974 (zuletzt abgerufen am 24.02.2021).

5.3 Entwicklung einer Hilfestellung zur Erklärung

Nach den Anforderungen aus Kapitel 4 soll der Privacy Check eine Hilfestellung zum Lesen einer Datenschutzerklärung geben. Diese soll in Form eines Frequently Asked Questions (FAQ) umgesetzt werden. Dieses FAQ wird neben der Datenschutzerklärung platziert. Dem Nutzer sollen durch das FAQ gängige Begriffe und Rechte aus dem Datenschutz erklärt werden. Durch diese Informationen wird die Explainablity des Privacy Checks verbessert.

Darüber hinaus gibt es im FAQ die Möglichkeit, die Zusammenfassung der Datenschutzerklärung anzuzeigen (vgl. Kapitel 4). Der Privacy Check soll Transparent anzeigen, aus welchen Abschnitten der Datenschutzerklärung die Informationen für die Zusammenfassung erhoben wurden. Dazu soll jedes Element der Liste vom Nutzer angesprochen werden, sodass der Privacy Check automatisch die entsprechende Stelle der Datenschutzerklärung markiert und anzeigt.

5.4 Entwicklung einer benutzerfreundlichen Bedienung

Nach Kapitel 4 soll der Nutzer den Privacy Check leicht bedienen können. Zusätzlich wird der Nutzer an seinem Surfverhalten nicht gehindert. Damit dies gelingen kann, soll der Privacy Check als Webextension für einen Browser entwickelt werden. Diese Webextension wird den Prozess der Zusammenfassung aus Abschnitt 5.1 erst auf Anfrage des Nutzers einleiten. Hinzukommend soll für den Privacy Check eine Oberfläche gewählt werden, in der ein Nutzer sich zurecht finden kann.

Die Entwicklung des Privacy Checks als Erweiterung für den Browser besitzt weitere Vorteile. Diese Erweiterung kann Informationen aus einer Webseite entnehmen oder eine Webseite erweitern. Somit kann für die Zusammenfassung aus Abschnitt 5.1 die Datenschutzerklärung aus einer Webseite durch die Erweiterung entommen werden. Zusätzlich kann die Hilfestellung aus Abschnitt 5.3 in die Webseite der Datenschutzerklärung eingebettet werden.

Kapitel 6

Implementierung des Privacy Checks

In diesem Kapitel wird das Konzept des Privacy Checks technisch umgesetzt. Zunächst wird auf die Architektur des Privacy Checks eingegangen. Dabei wird erklärt, warum die Auswertung der Datenschutzerklärung auf einen Server ausgelagert wird. Danach wird die Architektur des Privacy Checks anhand der grundlegenden Struktur einer Webextension erklärt. Anschließend wird die Oberfläche des Privacy Checks anhand von Screenshots erläutert.

6.1 Architektur des Privacy Checks

Nach Konzept soll der Privacy Check als Webextension entwickelt werden (vgl. Abschnitt 5.4). Die Auswertung soll mittels NLP folgen, was innerhalb einer Webextension schwer umzusetzen ist. Als Lösung bietet es sich an, die Auswertung auf einen Server auszulagern. Dies wird im Folgenden erläutert. Daneben ist für die Architektur zu klären, wie eine Webextension grundlegend aufgebaut ist und wie sich der Privacy Check dieser Strukturen bedient.

6.1.1 Auslagerung der Auswertung auf einen Server

Nach dem Konzept aus Kapitel 4 werden Datenschutzerklärungen für den Nutzer zusammengefasst. Das Konzept sieht vor, die Auswertung der Datenschutzerklärungen mitels NLP umzusetzen (vgl. Unterabschnitt 5.1.3). Für Javascript kann für diese Arbeit keine geeignete NLP-Bibliothek ermittelt werden. Dieses Teilproblem wird stattdessen mit Python umgesetzt. Python besitzt trainierte Bibliotheken für die Verarbeitung natürlicher Sprachen. Die Programmiersprache Python ist aber nicht auf jedem Rechner installiert. Daher wird die Verarbeitung einer Datenschutzerklärung auf einen Server ausgelagert.

Somit ist der Privacy Check Teil eines Client-Server-Systems. Dieser Server wird mit Python Flask implementiert[1], das für die Umsetzung eines Servers konzipiert ist.

Der Server ist mit einer Tee-and-Join Filter Architektur implementiert (vgl. Abschnitt 2.4). In Abbildung 6.1 ist die Architektur des Servers zu sehen. Der Server empfängt vom Privacy Check eine Datenschutzerklärung[2]. Flask sendet diese an den sogenannten „Summary Creator“ weiter. Der Summary Creator macht die TF*IDF-Bewertung der einzelnen Sinnabschnitte (vgl. Unterabschnitt 5.1.3). Danach sendet dieser das Ergebniss an die Filter weiter (Tee). Die Filter geben ihre Auswertung wiederum an den Summary Creator zurück. Der Summary Creator vereint (Join) die beiden Ergebnisse und leitet diese an Flask weiter. Flask wiederum leitet das Ergebnis an den Privacy Check zurück. Die genaue Arbeitsweise der Filter wird in Unterabschnitt 6.2.2 und Unterabschnitt 6.2.3 beschrieben.

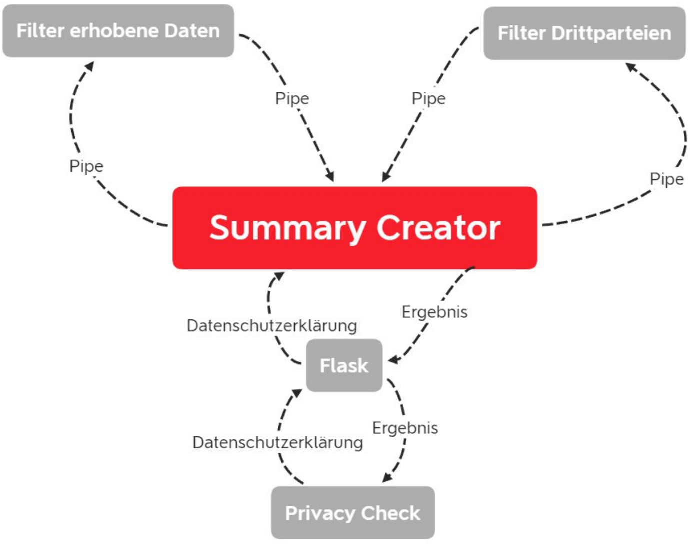


Abbildung 6.1: Architektur des Servers

Die Filter sind mit der Python Bibliothek spaCy implementiert. Diese wird verwendet, da nach Colic et al. [8] spaCy den schnellsten verfügbaren Parser für natürliche Sprachen hat und zudem eine hohe Genauigkt aufweist.

---

[1] https://flask.palletsprojects.com/

[2] Die Verbindung zwischen Privacy Check und Server ist verschlüsselt.

Die Bibliothek spaCy hat jedoch einen Nachteil:

Der Filter muss nach den Anforderungen des Privacy Checks private oder öffentliche Organisationen erkennen (vgl. Unterabschnitt 5.1.3). Nach der Arbeit von Schmitt et al.[20] gibt es keine vollständige Erkennung von Organisationen in Texten durch die NER (vgl. Unterabschnitt 2.2.4). Schmitt et al. [20] zeigen, dass bekannte NLP-Bibliotheken weniger als die Hälfte aller Organisationen erkennen. Zudem weisen Schmitt et al. [20] nach, dass spaCy in 40% aller Fälle eine Organisation erkennt. Jedoch erzielt eine Bibliothek bei der Erkennung von Organisationen bessere Ergebnisse: die Bibliothek CoreNLP. CoreNLP ist eine Bibliothek für die Sprache Java, die in 80% aller Fälle eine Organisation erkennt. Die Erkennung der erhobenen Daten beruht bei dieser auf dem beschrieben Algorithmus aus Unterabschnitt 5.1.3. Der Algorithmus benutzt die NLP-Techniken Tokenization und POS (vgl. Unterabschnitt 2.2.3). Das POS und die Tokenization sind Aufgaben eines Parsers. Da spaCy den schnellsten Parser nach Colic et al. [8] besitzt und zudem treffsicher arbeitet, verwendet diese Arbeit dennoch spaCy. Die Schnelligkeit dient der Benutzerfreundlichkeit.

Sowohl Flask als auch spaCy sind über den Python Paketmanager Anaconda[3] installiert. Eine Anleitung zum Starten des Servers befindet sich in Anhang B.

6.1.2 Strukturen einer Webextension

Der Privacy Check soll als Erweiterung in einem Browser benutzt werden (vgl. Abschnitt 5.4). Damit ist der Privacy Check eine Webextension. Webextensions werden in Javascript entwickelt und sollen die Funktionalitäten eines Browsers erweitern. Für Webextensions sind feste Strukturen definiert, die im Folgenden näher erklärt werden:

Backgroundscripte

Je nach Aufgabenstellung braucht eine Webextension einen Service, der im Hintergrund dauerhaft läuft. Diese Aufgabe wird den sogenannten Backgroundscripten zugedacht. Backgroundscripte können nur über Nachrichten kommunizieren. Für alles Weitere wird auf die Mozilla Developer Network (MDN) Webdocs verwiesen[4].

---

[3] https://anaconda.org/

[4] https://developer.mozilla.org/en-US/docs/Mozilla/Add-ons/WebExtensions/Anatomy_of_a_WebExtension#background_scripts

Contentscripte

Um eigene Javascripte auf einer fremden Webseite auszuführen, bedarf es sogenannter Contentscripte. Beim Aufruf können Contentscripte in eine Webseite geladen werden. Sie sind in der Lage, das Document Object Model der Webseite zu manipulieren. Contentscripte können mit den Backgroundscripten nur über Nachrichten kommunizieren. Für alles Weitere wird auch hier auf die MDN-Webdocs verwiesen[5].

Popup und Scripte

Das Popup-Window ist für die Kommunikation des Nutzers mit der Extension gedacht. Dieses Popup-Window benutzt die Hypertext Markup Language (HTML) um die Oberfläche zu beschreiben. Das Fenster kann bei Klick des Icons in der Statusleiste eines Browsers aktiviert werden (ähnlich Abbildung 6.2). Die Scripte werden im HTML-Dokument des Popup-Windows definiert und können mit dem Backgroundscripten nur über Nachrichten kommunizieren. Für alles weitere wird auf die MDN-Webdocs verwiesen[6].

Das Manifest

Jede Webextension hat ein Manifest. Dort werden generelle Informationen und Strukturen der Webextension festgehalten. Darunter fallen auch die Content- und Backgroundscripte und das Popup-Window. Eine Webextension kann über das Manifstest in den Browser geladen werden. Eine Anleitung zum Laden einer Webextension befindet sich in Anhang B. Für alles Weitere wird auf die MDN-Webdocs verwiesen[7].

Aufgaben der Strukturen

Da Webextensions eine feste Struktur haben, müssen die Aufgaben, die sich aus Kapitel 5 ergeben, auf diese verteilt werden (vgl. Kapitel 4). Zusätzlich werden den Scripten weitere Funktionen zugeordnet, die in dieser Arbeit als sinnvoll erachtet wurden und die Usability und Explainbility verbessern. Aufgabe der Contentscripte:

1. Filtern der Datenschutzerklärung

2. Filtern von E-Mail-Adressen

---

[5] https://developer.mozilla.org/en-US/docs/Mozilla/Add-ons/WebExtensions/Content_scripts

[6] https://developer.mozilla.org/en-US/docs/Mozilla/Add-ons/WebExtensions/user_interface/Popups

[7] https://developer.mozilla.org/en-US/docs/Mozilla/Add-ons/WebExtensions/manifest.json

3. Filtern von Drittpartei-Scripten
4. Erstellen der Hilfestellung
5. Userinteraktion mit der Hilfestellung
6. Erkennen, ob die vorliegende Webseite eine Datenschutzerklärung ist[8]

Aufgaben der Popup Scripte:

1. Einleitung des Filterungsprozesses
2. Erstellen einer erweiterten Ansicht
3. Userinteraktion mit der erweiterten Ansicht
4. Cookielisten erstellen
5. Listen von Drittanbieter-Scripten
6. Kommunikation mit dem Server

Aufgaben des Backgroudscriptes:

1. Festhalten, ob eine Webseite eine Datenschutzerklärung ist
2. Weiterleitung der Zusammenfassung

Aufgrund der festen Strukturen einer Webextension wird kein gewöhnliches Softwarearchitekturmuster gewählt. Es wird jedoch darauf geachtet, dass alle Scripte des Privacy Checks möglichst modular programmiert sind. Diese Vorgehensweise kann bei der Weiterentwicklung des Privacy Checks helfen, einzelne Scripte bei Bedarf auszutauschen.

Für diese Arbeit wird zusätzlich die Javascript Bibliothek Jquery[9] benutzt. Jquery besitzt Funktionen, die eine Manipulation einer Webseite und die Anfragen zum Server vereinfachen. Zusätzlich hilft Jquery dabei, die Datenschutzerklärung aus einer Webseite zu filtern (vgl. Unterabschnitt 6.2.1).

6.2 Filterung der relevanten Informationen aus einer Datenschutzerklärung

In diesem Abschnitt wird erklärt, wie die Filterung der relevanten Informationen aus Unterabschnitt 5.1.3 implementiert ist.

---

[8] Das Erkennen einer Datenschutzerklärung wird in Unterabschnitt 7.1.1 beschrieben.

[9] https://jquery.com/

6.2.1 Filterung der Datenschutzerklärung

Die Auswertung der Datenschutzerklärung findet auf einem Server statt (vgl. Unterabschnitt 6.1.1). Damit der Server eine Datenschutzerklärung zusammenfassen kann, muss im Vorfeld die eigentliche Datenschutzerklärung aus der Webseite gefiltert werden.
Die Filterung der Datenschutzerklärung aus einer Webseite geschieht durch die Contentscripte (vgl. Abschnitt 6.1.2).
Nach dem Konzept des Privacy Checks, müssen Sinnabschnitte erfasst werden (vgl. Unterabschnitt 5.1.3). Diese Aufgabe fällt ebenfalls den Contentscripten zu.
Damit das Contentscript Sinabschnitte erfassen kann, wird nach Überschriften in der Datenschutzerklärung gesucht. Jeder Sachverhalt, der nach einer Überschrift folgt, wird als Sinnabschnitt gespeichert. Ein Sachverhalt kann eine Liste oder ein Fließtext sein (vgl. Unterabschnitt 5.1.2).

Jedem Sinnabschnitt wird eine ID zugewiesen. Diese ID erfüllt eine Aufgabe: Die Sinnabschnitte werden von dem Server verarbeitet. Aus den Sinnabschnitten werden die Begriffe für die Zusammenfassung gefiltert. Die Begriffe für die Zusammenfassung erhalten diese ID. Damit ist es für die Hilfestellung rückverfolgbar, welche Begriffe aus der Zusammenfassung zu welchem Sinnabschnitt gehören. Somit macht die ID es möglich, Begriffe aus der Zusammenfassung einem Sinnabschnitt zuzuordnen (vgl. Abschnitt 5.3 und Unterabschnitt 6.3.3). Hinzukommend kann damit die Scrollfunktion für die Hilfestellung implementiert werden. Durch diese ID weiß die Scrollfunktion, zu welchem Abschnitt sich diese bewegen soll (vgl. Abschnitt 5.3).

Sofern eine Liste in dem Sinnabschnitt vorhanden ist, soll nach Konzept diese Liste in die Zusammenfassung aufgenommen werden (vgl. Unterabschnitt 5.1.3). Damit dies gelingt, wird jeder Sachverhalt eines Sinnabschnitts markiert. Diese Markierung gibt an, ob der Sachverhalt eine Liste ist.

Das beschriebene Vorgehen zur Filterung einer Datenschutzerklärung hat Probleme, die in Unterabschnitt 7.1.2 genauer erörtert werden.

6.2.2 Filterung der erhobenen Daten

Nach Konzept soll der Nutzer eine Zusammenfassung der Datenschutzerklärung in Form einer Liste erhalten (vgl. Kapitel 4). Das Konzept des Privacy Checks sieht in Unterabschnitt 5.1.3 einen Algorithmus vor, der eine Filterung der erhobenen Daten vornehmen kann.

Zur Umsetzung des TF*IDF-Maßstabes werden Techniken aus dem NLP verwendet, die mit der Python-Bibliothek spaCy umgesetzt sind (vgl. Unterabschnitt 6.1.1 und Abschnitt 2.3). Nach Konzept des Privacy Checks wird vorgesehen, nur Nomen und Pronomen im Text zu suchen und mit dem TF*IDF Maßstab zu bewerten (vgl. Abschnitt 2.3). Mittels spaCy wird die Technik der Tokenization verwendet, um einen Text in Wörter aufzuteilen (vgl. Unterabschnitt 2.2.1). Jedes Wort kann dann mit dem POS betrachtet werden (vgl. Unterabschnitt 2.2.3). Ist das Wort ein Pronomen oder Nomen, so wird darauf der TF*IDF-Maßstab angewendet. Damit alle Wörter richtig erfasst sind, werden zusätzlich die Techniken der Lemmatization benutzt (vgl. Unterabschnitt 2.2.2).

Nach Konzept soll eine Liste gängiger Phrasen aus einer Datenschutzerklärung erstellt werden. Diese Phrasen werden bei der Analyse der Datenschutzerklärungen gesammelt (vgl. Unterabschnitt 5.1.2).

Das Konzept schreibt ebenfalls vor den TF*IDF-Maßstab zu manipulieren, wenn eine bekannte Phrase gefunden wird (vgl. Unterabschnitt 5.1.3). Dafür wurde spaCys Pattern Matcher verwendet[10]. Mit spaCys Pattern Matcher kann man in einem Text nach Mustern suchen. Der Privacy Check nutzt dies für die Suche nach Phrasen aus der bekannten Phrasenliste (vgl. Unterabschnitt 5.1.3).

6.2.3 Filterung der Drittparteien

Nach den Anforderungen aus Kapitel 4 sollen Drittparteien gefiltert werden. Die Filterung der Drittparteien geschieht, wie die Filterung der erhobenen Daten, auf dem Server (vgl. Unterabschnitt 6.1.1) Die Filterung der Drittparteien geschieht jedoch nicht durch den beschrieben Algorithmus aus Unterabschnitt 5.1.3. Es wird festgestellt, dass der Algorithmus nur bedingt die gesuchten Drittparteien liefert. Um dieses Problem zu lösen, wird ein anderer Ansatz gewählt:

Es wird eine Liste von möglichen Drittparteien aus den Datenschutzerklärungen aus Anhang A erstellt. Taucht eine der Drittparteien aus der Liste in der Datenschutzerklärung auf, wird diese in die Zusammenfassung aufgenommen.

Zusätzlich wird die NER von spaCy verwendet (vgl. Unterabschnitt 2.2.4). Findet spaCy mittels NER eine Drittpartei, so wird diese in die Zusammenfassung aufgenommen.

---

[10] https://spacy.io/usage/rule-based-matching

6.3 Gestaltung der Nutzeroberfläche

In den folgenden Abschnitten wird die Umsetzung der Oberfläche erläutert und mit Screenshots aus dem Privacy Check unterstützt.

6.3.1 Material Design

Damit die Oberfläche des Privacy Checks ansprechend und konsistent wirkt, muss ein benutzerfreundliches Design gewählt werden. In dieser Arbeit wird das Material Design gewählt. Material Design ist eine Designvorgabe von Google und wird für Oberflächen in Android verwendet[16]. Das Material Design wird unter anderem auch im Internet genutzt [18].

Ein Nutzer kann mit Material Design durch die Bedienung eines Smartphones vertraut sein. Somit fördert Material Design die Usability des Privacy Checks.

Die Material Design Bibliotheken stehen zur freien Verfügung. Außerdem ist Material Design Open Source und kann über Git oder Node[11] installiert werden[12]. Damit Material Design für Webanwendungen benutzt werden kann, muss die Sprache SCSS oder SASS verwendet werden[13]. In dieser Arbeit wird dafür SCSS verwendet. SCSS muss jedoch in CSS umgewandelt werden, damit eine Webanwendung das Material Design benutzen kann. Um die Umwandlung zu bewerkstelligen, wird in dieser Arbeit das Programm gulp verwendet[14].

6.3.2 Popup-Window und die Zusammenfassung

Nach Konzept soll der Nutzer eine Zusammenfassung der Datenschutzerklärung bekommen (vgl. Kapitel 4). Diese Zusammenfassung ist eine Liste von erhobenen Daten eines Anbieters. Zusätzlich beinhaltet die Zusammenfassung eine Liste von Drittparteien, die Daten erheben können. Diese Zusammenfassung wird in das Popup-Window eingebaut. Der Nutzer muss nur auf den Knopf in der Toolbarleiste seines Browsers klicken und kann dann eine Zusammenfassung vom Server anfordern.
Das Bild für den Knopf der Toolbarleiste benutzt das Logo der Universität Hannover und ist in Abbildung 6.2 zu sehen.

Hat der Nutzer eine Datenschutzerklärung offen, zeigt das Popup-Window eine Meldung, dass der Nutzer die offene Datenschutzerklärung auswerten kann. Dies ist in Abbildung 6.3 zu sehen.

---

[11] Dies ist der Javascript-Paketmanager.

[12] https://material.io/

[13] https://sass-lang.com/

[14] https://gulpjs.com/

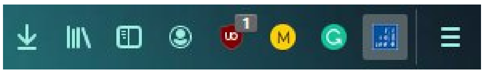
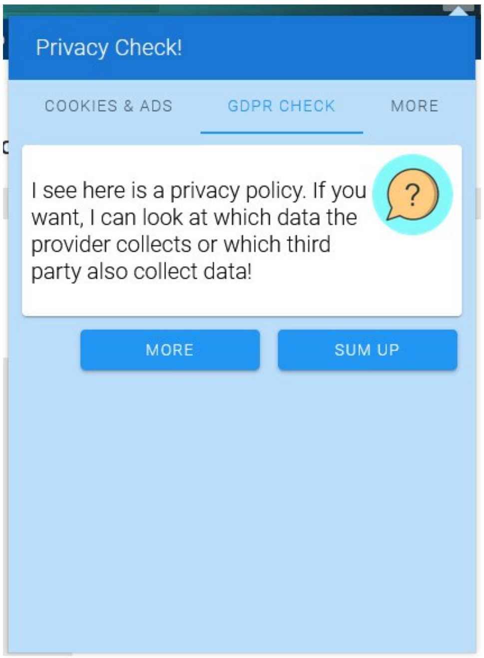


Abbildung 6.2: Privacy Check Button in der Toolbar



Privacy Check!

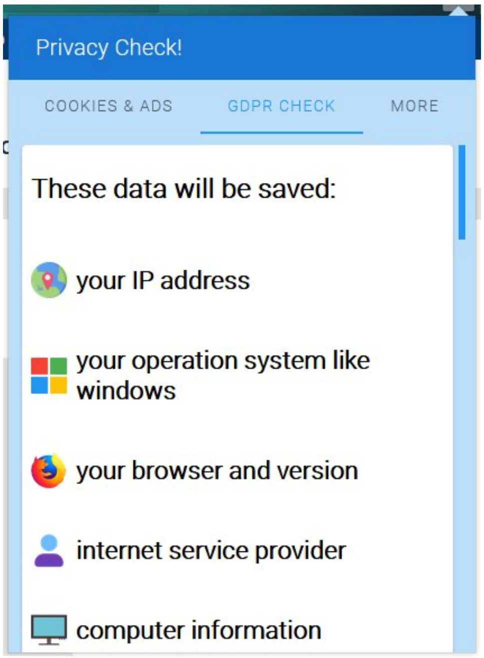
I see here is a privacy policy. If you want, I can look at which data the provider collects or which third party also collect data!

Abbildung 6.3: Hinweis für den Nutzer, dass er eine Auswertung der Datenschutzerklärung starten kann

Klickt der Nutzer auf den „Sum Up“-Button, wird das Popup-Window dem Nutzer eine Zusammenfassung der Datenschutzerklärung nach Konzept anzeigen (vgl. Kapitel 4). Dies ist in Abbildung 6.4 und Abbildung 6.5 zu sehen. Hier kann der Nutzer die Hilfestellung nach Konzept anfordern, indem er auf den „Advanced View“-Button drückt (Abschnitt 5.3).

Hat der Nutzer hingegen keine Datenschutzerklärung offen, so wird das Cookies & Ads-Fenster geöffnet (vgl. Abbildung 6.6). In diesem Fenster kann der Nutzer sehen, welche fremden Scripte der Anbieter auf die aufgerufene Webseite lädt. Diese fremden Scripte können genutzt werden, um Third Party Cookies oder Werbung zu platzieren, um den Nutzer zu tracken. Klickt der Nutzer auf einen dieser Links, wird die datenschutzfreundliche Suchmaschiene DuckDuckGo geöffnet, mit der ein Nutzer mehr über diese Scripte erfahren kann. Der Nutzer kann so angeregt werden, sich mehr mit der Materie des Trackings auseinander zu setzen. Somit kann ein Lerneffekt erreicht werden, was sich positiv auf die Explainability des Privacy Checks auswirkt.

Zusätzlich kann der Nutzer in diesem Fenster die aktiven Cookies der



These data will be saved:

- your IP address
- your operation system like windows
- your browser and version
- internet service provider
- computer information

Abbildung 6.4: Zusammenfassung der erhobenen Daten

Webseite nach Bedarf betrachten. Die First Party Cookies sind mit einem grünen Piktogramm versehen (Vgl Abbildung 6.6). Ist hingegen ein Third Party Cookie auf der Webseite aktiv, so zeigt der Privacy Check ein Warnsymbol an. Das Cookies & Ads-Tab kann auch geöffnet werden, obwohl keine Datenschutzerklärung offen ist.

6.3.3 Umsetzung der Hilfestellung

Der Nutzer soll nach Konzept eine Hilfestellung zum Lesen einer Datenschutzerklärung bekommen (vgl. Kapitel 4). Diese Hilfestellung soll nach Konzept ein FAQ zum Lesen einer Datenschutzerklärung beinhalten (vgl. Abschnitt 5.3). Zusätzlich soll für den Nutzer markiert werden, aus welchen Abschnitten der Datenschutzerklärung die Zusammenfassung die Informationen bezieht.

In Abbildung 6.7 sieht man die Hilfestellung zum Lesen einer Datenschutzerklärung. Diese Hilfestellung wird mittels eines Contentscripts in die Webseite geladen (vgl. Abschnitt 6.1.2). Wenn der Nutzer eine Frage hat, so kann er durch Klicken auf einen der Buttons eine Antwort erhalten.

Die Abbildung 6.8 zeigt eine Antwort auf die Frage, was persönliche Daten sind. Jeder Antwort zu jeder Frage aus dem FAQ wird ein Piktogramm zugeordnet. Dies hilft dem Nutzer Assoziationen zu knüpfen. Dieser Effekt

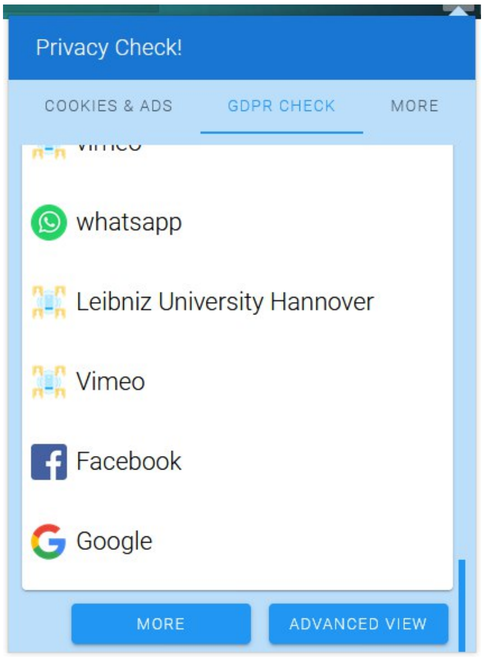


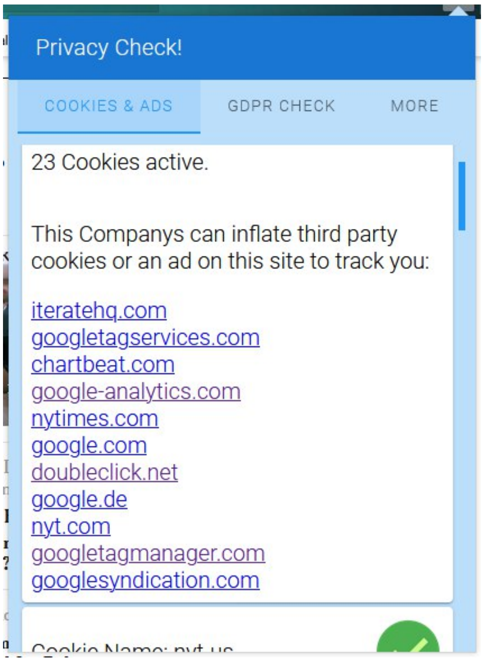
Abbildung 6.5: Zusammenfassung mit Drittparteien und dem Advanced View Knopf

der Assoziation steigert die Explainability des Privacy Checks.

Nach Konzept soll der Nutzer ein Element aus der Zusammenfassung nehmen können und in der Datenschutzerklärung markiert bekommen, woher die Zusammenfassung diese Information bezieht (vgl. Abschnitt 5.3). Diesbezüglich wird in dieser Arbeit der Ansatz gewählt, die Zusammenfassung in eine der Fragen des FAQ aufzunehmen. In Abbildung 6.9 sieht man, dass die Frage, welche Daten vom Anbieter erhoben werden, geöffnet ist. Der Nutzer kann auf „mark in text“ klicken. Der Privacy Check markiert dann in der originalen Datenschutzerklärung die Textpassage, aus der die Zusammenfassung eine Information bezieht. Klickt der Nutzer auf ein anderes Element der Zusammenfassung, so springt der Privacy Check an eine andere Stelle im Text.

Ähnlich verhält es sich mit den gesammelten Drittparteien in einer Zusammenfassung (vgl. Kapitel 4). Auch hier kann der Nutzer durch einen Klick auf „mark in text“ Textstellen in der Datenschutzerklärung markieren.

Der Nutzer bekommt weitere Hilfestellungen, die in Kapitel 4 nicht behandelt sind. Sofern ein Anbieter eine E-Mail-Adresse in der Datenschutzerklärung erwähnt, kann der Privacy Check diese aufnehmen. In Abbildung 6.10 ist zu sehen, dass die Frage im Bezug auf den Wiederruf der persönlichen Daten geöffnet ist. Der Nutzer wird hier informiert, dass er die Benutzung seiner Daten jederzeit widerrufen kann, indem er den Anbieter



23 Cookies active.

This Companys can inflate third party cookies or an ad on this site to track you:

iteratehq.com
googletagservices.com
chartbeat.com
google-analytics.com
nytimes.com
google.com
doubleclick.net
google.de
nyt.com
googletagmanager.com
googlesyndication.com

Cookie Name: nyt-us

Abbildung 6.6: Liste an Drittparteien, die ein Script auf die offene Seite laden können

kontaktiert. Dabei werden die E-Mail-Adressen des Anbieters angezeigt.

Das Filtern der E-Mail-Adresse geschieht durch ein Contentscript. Dieses Script benutzt reguläre Ausdrücke, welche in Texten nach Mustern suchen können[4]. Während der Arbeit stellt sich heraus, dass die Suche nach E-Mail-Adressen über Contentscripte effizienter ist, als diese Aufgabe dem Server zu überlassen.

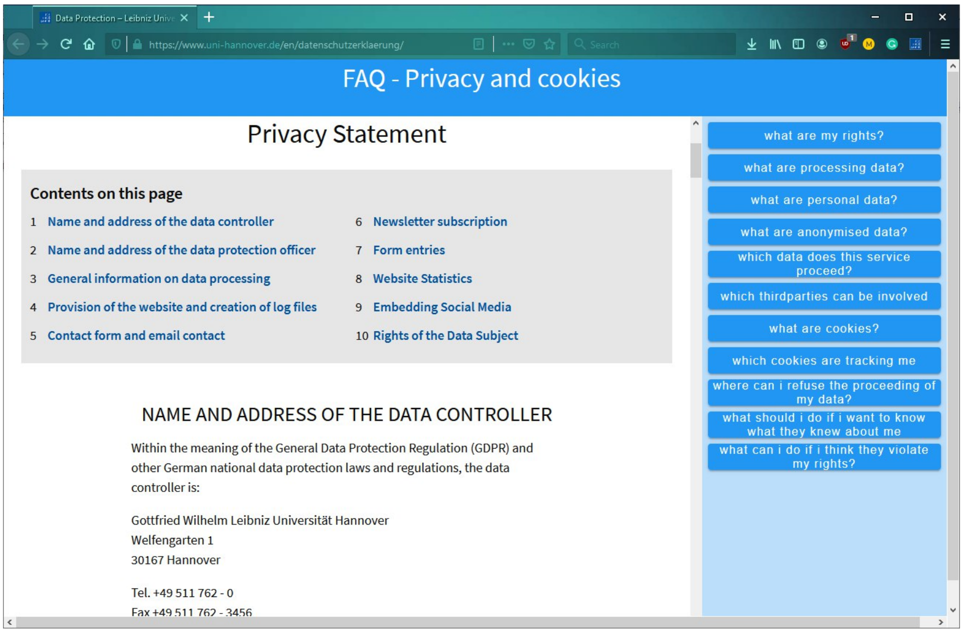
Bei dem Verdacht auf Missbrauch der eigenen Daten durch den Anbieter kann der Nutzer auf einen Link klicken. Dieser Link leitet den Nutzer an eine Liste der Landesbeauftragten für Datenschutz weiter[15].

6.3.4 Erweiterte Ansicht

In der erweiterten Ansicht kann der Nutzer mehr über das Programm erfahren. Zusätzlich bekommt dieser dort eine Anleitung, wie er das Tracking mit seinem Browser erschweren kann. In dieser Arbeit wird ebenfalls untersucht, ob die Anleitung auch automatisiert auf Knopfdruck umgesetzt werden kann. Es wird festgestellt, dass dies nicht möglich ist, da es

---

[15] https://www.bfdi.bund.de/DE/Infothek/Anschriften_Links /AufsBehoerdFuerDenNichtOeffBereich/AufsichtsbehoerdenNichtOeffBereich_liste.html



Privacy Statement

Contents on this page

1 Name and address of the data controller
2 Name and address of the data protection officer
3 General information on data processing
4 Provision of the website and creation of log files
5 Contact form and email contact
6 Newsletter subscription
7 Form entries
8 Website Statistics
9 Embedding Social Media
10 Rights of the Data Subject

NAME AND ADDRESS OF THE DATA CONTROLLER

Within the meaning of the General Data Protection Regulation (GDPR) and other German national data protection laws and regulations, the data controller is:

Gottfried Wilhelm Leibniz Universität Hannover
Welfengarten 1
30167 Hannover

Tel. +49 511 762 - 0
Fax +49 511 762 - 3456

Abbildung 6.7: Die Hilfestellung des Privacy Checks mit dem FAQ

Webextensions nicht erlaubt ist, Browsereinstellungen direkt zu ändern.[16] Zusätzlich befindet sich in der erweiterten Ansicht eine Liste an benutzten Piktogrammen und die Verweise auf die Ersteller.

6.3.5 Implementation und Nutzung der Piktogramme

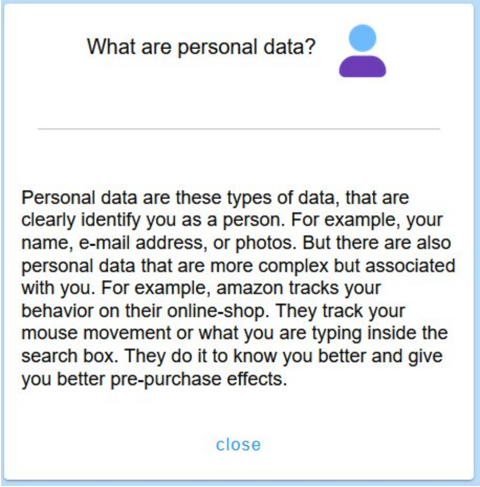
Aus Abschnitt 5.2 geht hervor, dass Piktogramme zur Unterstützung einer Zusammenfassung gewählt sind. Damit die Piktogramme in einer Zusammenfassung eingesetzt werden können, müssen diese angesprochen werden. Die Wahl der Piktogramme erfolgt über eine Liste. In dem Privacy Check gibt es eine Liste, die ein Wort einem Piktogramm zuordnet. Im Falle einer Zusammenfassung gleicht der Privacy Check jedes Wort mit der Liste ab. Ist dieses Wort in der Liste, so wird das zugehörige Piktogramm gewählt.

Alle benutzen Piktogramme stammen von der Webseite freeicons.io[17]. Diese stehen frei zur Verfügung unter der Bedingung, dass angegeben werden muss, dass die Piktogramme von dieser Webseite stammen und welcher Autor diese erstellt hat.

---

[16] https://wiki.mozilla.org/WebExtensions/FAQ#Will_I_have_access_to_about:config_or_the_preferences.3F (zuletzt abgerufen am 09.02.2021).

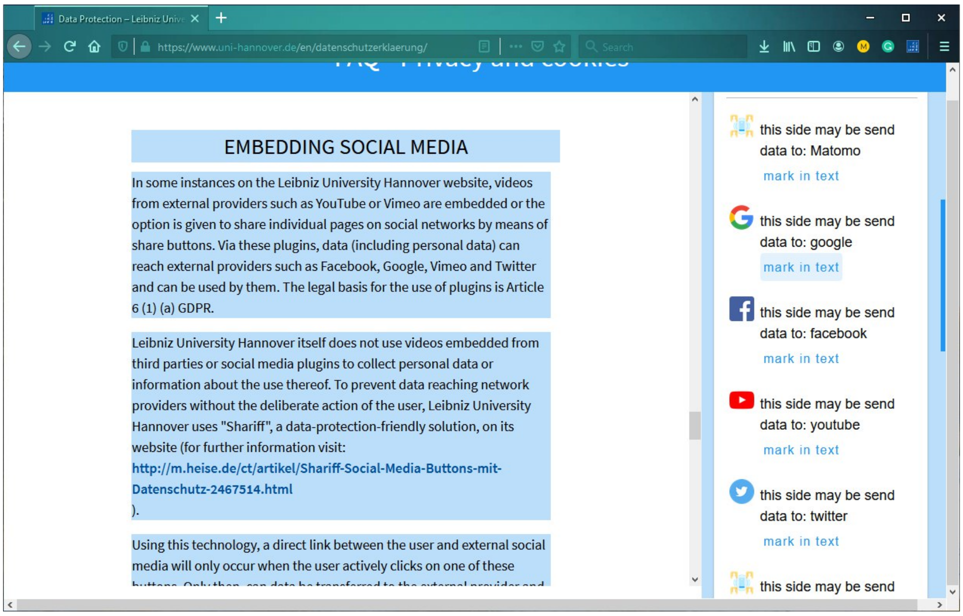
[17] https://freeicons.io/



What are personal data?

Personal data are these types of data, that are clearly identify you as a person. For example, your name, e-mail address, or photos. But there are also personal data that are more complex but associated with you. For example, amazon tracks your behavior on their online-shop. They track your mouse movement or what you are typing inside the search box. They do it to know you better and give you better pre-purchase effects.

Abbildung 6.8: Antwort auf die Frage was persönliche Daten sind



EMBEDDING SOCIAL MEDIA

In some instances on the Leibniz University Hannover website, videos from external providers such as YouTube or Vimeo are embedded or the option is given to share individual pages on social networks by means of share buttons. Via these plugins, data (including personal data) can reach external providers such as Facebook, Google, Vimeo and Twitter and can be used by them. The legal basis for the use of plugins is Article 6 (1) (a) GDPR.

Leibniz University Hannover itself does not use videos embedded from third parties or social media plugins to collect personal data or information about the use thereof. To prevent data reaching network providers without the deliberate action of the user, Leibniz University Hannover uses "Shariff", a data-protection-friendly solution, on its website (for further information visit: http://m.heise.de/ct/artikel/Shariff-Social-Media-Buttons-mit-Datenschutz-2467514.html
).

Using this technology, a direct link between the user and external social media will only occur when the user actively clicks on one of these

this side may be send data to: Matomo

this side may be send data to: google

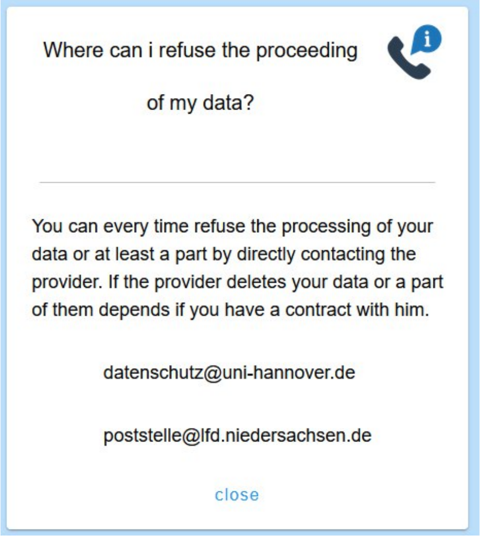
this side may be send data to: facebook

this side may be send data to: youtube

this side may be send data to: twitter

this side may be send

Abbildung 6.9: Markierter Textabschnitt aus dem eine Zusammenfassung seine Informationen bezieht



Where can i refuse the proceeding

of my data?

You can every time refuse the processing of your data or at least a part by directly contacting the provider. If the provider deletes your data or a part of them depends if you have a contract with him.

datenschutz@uni-hannover.de

poststelle@lfd.niedersachsen.de

Abbildung 6.10: Antwort auf die Frage, wie ein Nutzer die Verarbeitung seiner Daten wiederrufen kann.

Kapitel 7

Grenzen der Implementierung

Dieses Kapitel setzt sich kritisch mit den Grenzen der Implementierung auseinander. Dabei wird auf die Probleme hingewiesen, die während der Bearbeitungszeit nicht gelöst werden konnten.

7.1 Probleme der Filterung relevanter Informationen

Bei der Filterung relevanter Informationen kann es zu Fehlern kommen. Warum diese Fehler auftreten, wird im Folgenden beschrieben.

7.1.1 Erkennen einer Datenschutzerklärung

Das Contentscript ist in der Lage, zu erkennen, ob eine Seite eine Datenschutzerklärung enthält. Dies tut es mit einem naiven Ansatz. Es sucht nach festgelegten Begriffen. Diese Begriffe sind: Privacy Policy, GDPR oder Legal. Das Contentscript sucht nach diesen Begriffen und zählt diese. Wird einer dieser Begriffe häufiger als eine Zahl X gezählt, so wird die Seite als Datenschutzerklärung angesehen. Dieser naive Ansatz kann fehlschlagen, wenn eine Webseite keine Datenschutzerklärung enthält, diese Begriffe jedoch öfter als X auftauchen. Dann wird die Seite falsch als Datenschutzerklärung erkannt. Ein anderer Ansatz konnte aufgrund der zeitlichen Begrenzung dieser Arbeit nicht entwickelt werden. Der Nutzer könnte versehentlich eine Seite auswerten lassen, die keine Datenschutzerklärung beinhaltet.

7.1.2 Filterung der Datenschutzerklärung

Wie in Abschnitt 6.1.2 beschrieben, sucht das Contentscript nach HTML-Überschriften. In HTML werden die Überschriften mit <h1> bis <h6> Tags versehen und im Folgenden als H-Tags bezeichnet.

Bei manchen Datenschutzerklärungen kann es vorkommen, dass sie wenige H-Tags beinhalten. Dies kann der Fall sein, wenn eine Datenschutzerklärung durch ein sogenanntes Web Content Management System (Web-CMS) erstellt wurde. Ein Web-CMS ist eine Software, die automatisch den Inhalt auf einer Webseite verwaltet [2, S.36]. Diese Technologie erlaubt es den Betreibern ohne programmiertechnische Kenntnisse eine Webseite zu verwalten. Der Betreiber kann mittels eines Web-CMS neue Inhalte erstellen. Das Web-CMS setzt diese um. Wie genau das Web-CMS die Inhalte des Anbieters übersetzt ist je nach Implementierung unterschiedlich. In dem Fall der Erstellung einer Datenschutzerklärung wird vom Web-CMS manchmal für Überschriften der H-Tag verwendet. Wird dieser nicht verwendet, benutzt das Web-CMS den Paragraph-Tag von HTML. Wird der Paragraph-Tag verwendet, kann das Contentscript die Sinnabschnitte (vgl. Unterabschnitt 5.1.3) nicht wie gewünscht filtern. Problematisch ist, dass das Contentscript die komplette Datenschutzerklärung ohne die Sinnabschnitte sendet. Daraus resultiert, dass der Server die Benutzung des TF*IDF-Maßstabes nicht anwenden kann (vgl. Abschnitt 2.3 und Unterabschnitt 5.1.3). Damit kommt ein zentraler Baustein des Algorithmus zur Erfassung der erhobenen Daten nicht zur Anwendung.

Als Folge nimmt der Server unter Umständen alle Informationen einer Liste (vgl. Unterabschnitt 5.1.3) in die Zusammenfassung auf. Bei diesen Informationen handelt es sich nicht notwendigerweise um erhobene Daten, vielmehr können andere Sachverhalte ebenfalls in diesen Listen erfasst sein. Dadurch werden andere Sachverhalte, die nichts mit der Erhebung von Daten zu tun haben, mit in die Zusammenfassung aufgenommen.
Damit kann der Server keine exakte Zusammenfassung liefern, was die Explainability dieser Software vermindert.

7.1.3 Grenzen bei der Filterung erfasster Daten

In dieser Arbeit wurde der Zusammenfassungsalgorithmus mittels des TF*IDF-Maßstabs nach Suhartono et al. [7] verwendet.

Der Algorithmus kann grundsätzlich Zusammenfassungen liefern. Für die Zielsetzung wäre es notwendig gewesen, dass der Algorithmus Begriffe, die mit der Erhebung von Daten zu tun haben, höher gewichtet. In der Anwendung des Algorithmus zeigt sich jedoch, dass die Begriffe nicht höher eingestuft werden, obwohl der TF*IDF-Maßstab die Relevanz eines Begriffs in einem Text beschreibt.

Zur Lösung dieses Problems wird eine Liste von gängigen Phrasen aus mehreren Datenschutzerklärungen erstellt (vgl. Unterabschnitt 5.1.3). Taucht eine solche Phrase in einem Sinnabschnitt auf, wird der TF*IDF-

Wert des Sinnabschnitts automatisch erhöht. So ist gewährleistet, dass relevante Sinnabschnitte erfasst werden. Zudem kann gewährleistet werden, dass Sinnabschnitte, die eine Liste von erhobenen Daten beinhalten, in die Zusammenfassung miteinbezogen werden. Dies erhöht die Explainability.

In der Anwendung zeigt sich indessen, dass längere Texte vollständig in die Zusammenfassung aufgenommen werden. Dies führt dazu, dass die Zusammenfassung nicht mehr prägnant ist. Damit ist die Explainability des Programms nicht mehr gegeben.

Es wird versucht, die Prägnanz wieder zu erhöhen, indem der Zusammenfassungsalgorithmus von Suhartono et al. [7] auf die langen Texte angewendet wird. Diese Vorgehensweise erweist sich nicht als zielführend, da der Algorithmus nur selten die passenden Sätze auswählt. Um eine gewisse Prägnanz dennoch zu erreichen, werden nur Texte mit weniger als 15 Wörtern einbezogen. Textabschnitte mit mehr als 15 Wörtern werden vernachlässigt. Diese Art der Zusammenfassung ist nicht optimal.

7.1.4 Grenzen der Filterung von Drittparteien

Für die Zusammenfassung muss festgehalten werden, welche Drittparteien persönliche Daten vom Anbieter erhalten. Unter Drittparteien sind private oder öffentliche Organisationen zu verstehen. Dazu müssen die Namen der Drittparteien herausgefiltert werden. Diese Filterung geschieht auf zwei Wegen:

1. Durch das Erkennen von spaCys NER

2. Durch eine Liste gesammelter Organisationen

Named Entity Recognition beschreibt das Erkennen und Einordnen von Eigennamen (vgl. Unterabschnitt 2.2.4). Die genannte Liste beinhaltet eine Reihe von Organisationen, die während der Testphase des Privacy Check gesammelt werden. Die Filterung der Drittparteien wird nicht zufriedenstellend gelöst. Dies ist darauf zurückzuführen, dass spaCy bereits im Ansatz nur 40% der relevanten Organisationen erkennt (vgl. Unterabschnitt 6.1.1). Auch der Einbezug der Liste führt nicht zur Vollständigkeit der Zusammenfassung, da eine manuell gepflegte Liste nur unter Einsatz vieler Ressourcen vervollständigt werden kann.

Es wird eine graduelle Verbesserung mittels des TF*IDF - Maßstabes auf Eigennamen untersucht. Beim Testen wird jedoch festgestellt, dass sich die Fehlerquote nicht signifikant verbessert. Aufgrund des Zeitrahmens wird keine weitere Optimierung der Filterung von Drittparteien vorgenommen.

7.1.5 Das Duplikat Problem

Eine vermeintliche Problematik auf der Seite des Servers ergibt sich bei der Anzeige der erhobenen Daten: Die gleiche Information kann mehrfach erfasst werden, wenn sie in der Datenschutzerklärung an verschieden Stellen vorkommt. Wenn beispielsweise die IP-Adresse mehrfach in der Datenschutzerklärung genannt wird, kann dies dazu führen, dass in der Übersicht der gespeicherten Daten die IP-Adresse dementsprechend häufig vorkommt.

Bei näherer Betrachtung zeigt sich jedoch, dass diese Duplikate Potenzial haben, die Explainability zu erhöhen. Obwohl die Webextention immer denselben Begriff in der Zusammenfassung darstellt, liegen dahinter verschiedene Informationen. Die Begriffe müssten demnach weiter differenziert werden, um die Informationen vollständiger abbilden zu können. Diese Differenzierung wird aufgrund des Zeitrahmens nicht erreicht. In diesem Kontext kann zumindest die Hilfestellung Abhilfe schaffen. Durch ein Anklicken des Begriffs führt die Hilfestellung den Nutzer zu der betreffenden Stelle im Dokument.

7.2 Darstellungsfehler der Nutzeroberfläche

Dieser Abschnitt befasst sich mit den Gründen für Darstellungsfehler, die beim Testen des Programmes aufgetreten sind.

7.2.1 Implementationsformen der Hilfestellung

In der Anwendung zeigen sich gelegentlich Abweichung in der Darstellung der Hilfestellung. Hierbei handelt sich um Veränderungen in der Formatierung, wie z.B. der Größe der Schriftart. Gründe für diese Veränderungen finden sich im bereits geschilderten problematischen Zusammenwirken des Pivacy Checks mit dem Web-CMS (vgl. Unterabschnitt 7.1.2).
Zur Vermeidung der geschilderten Formatierungsabweichungen werden zwei alternative Implementationen erprobt.

1. **Extrahierung der Datenschutzerklärung**
   Bei diesem Lösungsansatz wird die Datenschutzerklärung aus einer Webseite extrahiert und in eine neue Webseite geladen. Mit dieser Lösung wäre eine einheitliche Formatierung gewährleistet, nicht tragbar erweist sich jedoch, dass die Formatierung der Datenschutzerklärung gebrochen wird. Dies ist darauf zurückzuführen, dass nicht die komplette HTML-Seite geladen wird, sondern nur der die Datenschutzerklärung betreffende Teil. Dabei werden das CSS und die Skripte nicht übernommen. Alternativ könnte die komplette HTML-

Seite übernommen werden, womit die Skripte des Web-CMS wieder nachteilig Einfluss nehmen (vgl. Unterabschnitt 7.1.2).

2. **Einbindung eines IFrames**
   IFrame ist eine HTML-Technik, die ermöglicht, andere Webseiten in die eigene Webseite einzubinden. Dabei werden das CSS und die Scripte der anderen Webseite nicht in die Eigene geladen. In dieser Implementation werden die Contentscripte in das IFrame geladen. Problematisch erweist sich dabei die Kommunikation mit den Content-Scripten des IFrames. Die Contentscripte konnten nicht angesprochen werden. In dem gegebenen Rahmen erscheint eine weitere Befassung mit dem IFrame Ansatz nicht aussichtsreich.

7.2.2 Probleme bei der Scrollanimation

Es kann gelegentlich zu Fehlern in der Scrollanimation kommen. Die Animation trifft nicht den markierten Absatz. Der Grund dafür ist, dass die Scripte des Web-CMS Einfluss auf die Scrollanimation nehmen. Zum Erreichen der Explainability besteht hier Nachbesserungsbedarf. Aufgrund des Zeitrahmens dieser Arbeit wird keine Verbesserung vorgenommen.

7.2.3 Probleme bei der Makierung der Textabschnitte

Für die Hilfestellung zum Lesen einer Datenschutzerklärung ist vorgesehen, dass manche Textabschnitte in der Datenschutzerklärung markiert werden. Das Web-CMS kann dies verhindern. Zum Erreichen der Explainability besteht hier ebenfalls Nachbesserungsbedarf. Aufgrund des Zeitrahmens dieser Arbeit wird auch hier keine Verbesserung vorgenommen.

7.2.4 Probleme bei der Auswahl von Piktogrammen

Für die Zusammenfassung werden Piktogramme verwendet, da die Visualisierung der Explainablity dienlich ist (vgl. Abschnitt 5.2). Die Zuordnung der Piktogramme erfolgt nicht in jedem Fall treffsicher.

Die Wahl des Piktogrammes ist im Privacy Check davon abhängig, welche Wörter im Antwortsatz der Zusammenfassung stehen. Wenn am Anfang eines Satzes ein Wort der Liste auftaucht, ordnet das Programm dieses Wort dem entsprechenden Piktogramm zu (vgl. Unterabschnitt 6.3.5). Dies ist dann problematisch, wenn im weiteren Verlauf dieses Satzes ein für die zu übermittelnde Information relevanterer Begriff auftaucht. Bei dem Nutzer kann dadurch unter Umständen eine falsche Assoziation hervorgerufen werden. Dadurch kann die Explainability des Privacy Checks verringert werden.

Kapitel 8

Zusammenfassung und Ausblick

Dieses Kapitel fasst diese Arbeit zusammen. Zusätzlich gibt dieses Kapitel einen Ausblick, bei welchen auf die Verbesserungsmöglichkeit eingegangen und sich ergebende Forschungsfragen vorgestellt werden.

8.1 Zusammenfassung

Datenschutzrichtlinen und -erklärungen sind komplex und oft nicht verständlich für einen Nutzer. Oftmals tritt beim Nutzer beim Lesen ein Information Overkill auf und die Datenschutzerklärung wird nicht weiter zu Kenntnis genommen. Dies wird durch die mangelnde Fühlbarkeit der Folgen verstärkt.

In dieser Arbeit wird ein Tool entwickelt und implementiert, dass Datenschutzrichtlinien prägnant und anschaulich kommunizieren soll. Dieses Tool trägt den Namen Privacy Check.

In den Anforderungen an die Konzeptentwicklung des Privacy Checks wird festgelegt, dass dieser dem Nutzer eine prägnante und anschauliche Zusammenfassung präsentieren soll. Zur Erreichung der Prägnanz werden die zu übermittelnden Informationen auf die zu erhebenden Daten und die datenempfangenden Drittparteien reduziert. Die sich daraus ergebende Zusammenfassung der Datenschutzerklärung stützt sich auf einen Algortihmus, welcher die relevanten Informationen filtert. Die Entwicklung dieses Algorithmus stützt sich auf die Betrachtung von 50 Datenschutzerklärungen, anhand welcher Muster zur besseren Funktionsfähigkeit des Algorithmus herausgearbeitet werden. Dabei zeigt sich, dass 74% der Betreiber eine Liste mit erhobenen Daten führen und 64% die Erfassung von Daten in einem Fließtext kommunizieren.

Als Schlussfolgerung ergibt sich, dass die Filterung der relevanten Informationen mit Techniken aus dem NLP und dem Information Retrieval umgesetzt werden kann. Auf einen maschinellen Lernansatz wird aufgrund des Zeitrahmens dieser Arbeit verzichtet. Für die Anschaulichkeit der Zusammenfassung und des Privacy Checks werden Piktogramme eingesetzt, die dem Nutzer kommunizieren, welche Daten erhoben werden.

Des Weiteren wird für den Privacy Check in den Anforderungen ebenfalls festgelegt, dass dieser eine Hilfestellung zum Lesen einer Datenschutzerklärung geben soll. Diese Hilfestellung soll technische sowie datenschutzrelevante Begriffe erklären. Zusätzlich soll die Hilfestellung dem Nutzer zeigen, aus welchen Informationen die Zusammenfassung erstellt wurde. Diese Hilfestellung wird in Form einer FAQ in die Datenschutzerklärung einer Webseite eingebettet.

Der Privacy Check wird als Webextension für den Firefox implementiert. Die Filterung der relevanten Informationen wird jedoch auf einen Server ausgelagert.

In der Implementation der Filterung zeigen sich Probleme, welche durch den Algorithmus zur Filterung relevanter Daten entstehen. Diese Probleme können in dieser Arbeit nicht zufriedenstellend gelöst werden.

8.2 Ausblick

Im Folgenden werden mögliche Verbesserungen des Programms und eine weitere Forschungsfrage vorgestellt.

8.2.1 Verbesserung der Filter

Der Privacy Check benutzt keinen maschinellen Lernansatz. Dieser Ansatz könnte zufriedenstellendere Ergebnisse für die Filterung der relevanten Informationen geben. Nach Nüske et al. [17] bedarf dies jedoch eines ausgeprägten Know-hows auf diesem Gebiet und kann zudem sehr teure Implementierungskosten verursachen.

8.2.2 Verstärkung der Explainability durch eine Wissensdatenbank

Damit die Eplainability der Software verbessert werden kann, könnten neben den datenschutzrechtlichen Begriffen die Erklärung technischer Begriffe noch mehr ausgeweitet werden. Hierfür könnte eine Wissensdatenbank implementiert werden. Hiervon könnten Nutzer ohne oder mit geringem Know-how profitieren.

8.2.3 Untersuchung des Lernerfolgs des Privacy Checks

Diese Software bietet einem Nutzer eine kurze und prägnante Zusammenfassung. Da die Beschreibung von erfassten Daten ebenfalls vom Kontext abhängig ist, müsste untersucht werden, ob der Nutzer durch die hervorgerufenen Assoziationen von Bild und Text den Sinn richtig erfasst und Konsequenzen für sich selbst zieht. Eine mögliche Konsequenz vom Nutzer wäre, dass er weniger Daten von sich preisgibt. Es müsste ebenfalls untersucht werden, ob das Bewusstsein des Nutzers für Datenschutz dadurch verbessert wird. Zusätzlich könnte auch untersucht werden, inwieweit die Hilfestellung genutzt wird oder ob der Nutzer nur die prägnante Zusammenfassung benutzt.

Anhang A

Anhang - Liste betrachteter Datenschutzerklärungen

Im folgenden eine Auflistung der untersuchten Datenschutzerklärungen:

1. https://www.eresearch.uni-goettingen.de/privacy-policy/
2. https://mediadaten.heise.de/en/home/data-protection/
3. https://www.uni-siegen.de/start/kontakt/datenschutzerklaerung.html.en
4. https://www.uni-hamburg.de/en/datenschutz.html
5. https://commercial.cnn.com/privacy-policy
6. https://www.emo-hannover.de/en/privacy-policy/privacy-policy.xhtml
7. https://www.visit-hannover.com/en/Privacy
8. https://www.hannover-airport.de/en/the-company/sonderseiten/data-protection-statement/
9. https://www.bbc.co.uk/usingthebbc/privacy-policy/
10. https://www.hotel-berlin.de/en/privacy-policy/
11. https://airbrake.io/privacy
12. https://www.fu-berlin.de/en/redaktion/impressum/datenschutzhinweise/index.html
13. https://corporate.aboutyou.de/en/privacy-policy
14. https://gdpr.eu/privacy-policy/
15. https://www.centaurmedia.com/privacy/

16. https://secureprivacy.ai/privacy-policy-sp/
17. https://www.nytimes.com/privacy/privacy-policy#we-allow-for-personalized-advertising-on-times-services
18. https://www.politico.com/privacy-policy
19. https://www.insider-inc.com/privacy-policy#ccpa
20. https://www.digitaltrends.com/privacy-policy/
21. https://www2.tech.co/en__privacy_policy__techco/a9417?_ga=2.149398326.1044129999.1612596539-1585275001.1612596539
22. https://www.sky.com/help/articles/sky-privacy-and-cookies-notice
23. https://www.privacypolicies.com/our-privacy-policy/
24. https://www.termsfeed.com/legal/privacy-policy/
25. https://www.rca.ac.uk/contact-us/about-this-website/privacy-cookies/
26. https://corporate.britannica.com/privacy-policy/
27. https://www.royalcollege.ca/rcsite/about/privacy-e
28. https://www.rcseng.ac.uk/privacy-policy/
29. https://www.hec.edu/en/data-privacy-policy
30. https://www.uni-muenster.de/de/en/datenschutzerklaerung.html
31. https://www.uni-hannover.de/en/datenschutzerklaerung/#privacystatement
32. https://www.dowjones.com/privacy-notice/?mod=WSJ
33. https://www.visit-hannover.com/en/Privacy
34. https://securityconference.org/en/privacy-policy/
35. https://staatsbibliothek-berlin.de/en/extras/allgemeines/imprint/privacy-policy/
36. https://tripadvisor.mediaroom.com/us-privacy-policy
37. https://www.tu.berlin/en/footer/data-protection/
38. https://www.cambridge.org/about-us/legal-notices/privacy-notice
39. https://www.w3resource.com/privacy.php

40. https://www.oreilly.com/privacy.html
41. https://www.whatsapp.com/legal/privacy-policy
42. https://policies.google.com/privacy?hl=en-US#infocollect
43. https://www.facebook.com/about/privacy
44. https://www.amazon.com/gp/help/customer/display.html?nodeId=GX7NJQ4ZB8MHFR
45. https://newsprivacy.co.uk/single/
46. https://www.make-it-in-germany.com/en/footer-meta/privacy-policy
47. https://www.nbcuniversal.com/privacy?intake=NBC_News#accordionheader1
48. https://twitter.com/en/privacy
49. https://www.foxnews.com/privacy-policy
50. https://system1.com/terms/privacy-policy

Anhang B

Anhang - Installationsanleitungen

B.1 Start des Servers

1. Den Packet Manager anaconda starten. Flask als auch spaCy sind in „base“ installiert
2. in das Verzeichnis des Servers wechseln wo die server.py liegt
3. startscript ausführen

B.2 Einbinden des Plugins

1. about:debugging in Firefox öffnen
2. auf „dieser Firefox“ wechseln
3. Danach auf „Temporäres Add-on laden...“ drücken
4. Das Manifest des Privacy Checks auswählen

Anhang C

Anhang - Abkürzungsverzeichnis

- **DSGVO** Datenschutz-Grundverordnung (DSGVO)
- **EU** Europäischen Union
- **Art.** Artikel
- **NLP** Natural Language Processing
- **POS** Part of Speech Tagging
- **NER** Namend Entity Recognition
- **TF** Term Frequency and
- **IDF** Inverse Document Frequency Maßstab
- **P3P** Platform for Privacy Projekt
- **W3C** World Wide Web Consortium
- **XML** Extensible Markup Language
- **FAQ** Frequently Asked Questions
- **MDN** Mozilla Developer Network
- **HTML** Hypertext Markup Language
- **Web-CMS** Web Content Management System

Anhang D

Anhang - Inhalt der beiliegenden DVD

1. PDF dieser Bachelorarbeit
2. Git-Repository
    - Privacy Check Webextension
    - Der Server für den Privacy Check
    - Entwicklungsumgebung für das Material Design

Literaturverzeichnis

[1] J. Awwalu, S. Abdullahi, and A. Evwiekpaefe. Parts of speech tagging: A review of techniques. *FUDMA Journal of Sciences*, 4:712–721, 06 2020.

[2] D. Barker. *Web Content Management.* O'Reilly Media, Inc., 3 edition, 3 2016.

[3] P. Barr, J. Noble, and R. Biddle. Icons r icons. In *Proceedings of the Fourth Australasian User Interface Conference on User Interfaces 2003 - Volume 18*, AUIC '03, page 25–32, AUS, 2003. Australian Computer Society, Inc.

[4] S. Bhatia. Regular expressions. *Computer Apex*, 01 2005.

[5] F. Buschmann, R. Meunier, H. Rohnert, P. Sommerlad, and M. Stal. *Pattern-orientierte Software-Architektur: Ein Pattern-System.* Addison-Wesley, Bonn, 1998.

[6] L. Chazette and K. Schneider. Explainability as a non-functional requirement: challenges and recommendations. *Requirements Engineering*, 25, 12 2020.

[7] H. Christian, M. Agus, and D. Suhartono. Single document automatic text summarization using term frequency-inverse document frequency (tf-idf). *ComTech: Computer, Mathematics and Engineering Applications*, 7:285, 12 2016.

[8] N. Colic and F. Rinaldi. Improving spacy dependency annotation and pos tagging web service using independent ner services. *Genomics Informatics*, 17:e21, 06 2019.

[9] L. F. Cranor. P3p: making privacy policies more useful. *IEEE Security Privacy*, 1(6):50–55, 2003.

[10] L. F. Cranor, P. Guduru, and M. Arjula. User interfaces for privacy agents. 13(2):135–178, June 2006.

[11] D. M. V. Dr. Sara Elisa Kettner, Prof. Dr. Christian Thorun. Wege zur besseren informiertheit. Technical report, ConPolicy GmbH, Institut für Verbraucherpolitik, 2018.

[12] S. Joseph, K. Sedimo, F. Kaniwa, H. Hlomani, and K. Letsholo. Natural language processing: A review. *Natural Language Processing: A Review*, 6:207–210, 03 2016.

[13] J. Kanis and L. Skorkovská. Comparison of different lemmatization approaches through the means of information retrieval performance. pages 93–100, 09 2010.

[14] S. P. Makini, I. Oguntola, and D. Roy. Spelling their pictures: The role of visual scaffolds in an authoring app for young children's literacy and creativity. In *Proceedings of the Interaction Design and Children Conference*, IDC '20, page 372–384, New York, NY, USA, 2020. Association for Computing Machinery.

[15] A. Mansouri, L. Affendey, and A. Mamat. Named entity recognition approaches. *Int J Comp Sci Netw Sec*, 8, 01 2008.

[16] K. Mew. *Learning Material Design: Master Material Design and create beautiful, animated interfaces for mobile and web applications.* 2. Packt Publishing, 2015.

[17] N. Nüske, C. Olenberger, D. Rau, and F. Schmied. Privacy bots. *Datenschutz und Datensicherheit - DuD*, 25:28–32, 01 2019.

[18] P. Patel. A guide to material design, a modern software design language. 04 2016.

[19] S. Robertson. Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation - J DOC*, 60:503–520, 10 2004.

[20] X. Schmitt, S. Kubler, J. Robert, M. Papadakis, and Y. LeTraon. A replicable comparison study of ner software: Stanfordnlp, nltk, opennlp, spacy, gate. pages 338–343, 10 2019.

[21] G. Starke and P. Hruschka. *Software-Architektur kompakt angemessen und zielorientiert*, volume 2. Spektrum, Akad. Verl., Heidelberg, 2009.

[22] N. Tomuro, S. Lytinen, and K. Hornsburg. Automatic summarization of privacy policies using ensemble learning. pages 133–135, 03 2016.

[23] P. Voigt and A. Bussche. *EU-Datenschutz-Grundverordnung (DSGVO).* Springer Verlag, 01 2018.

[24] P. D. K. von Lewinsk und Dirk Pohl LL.B.r. Kommunikation von datenschutz – recht und (gute) praxis. Technical report, Stiftung Datenschutz, 2017.

[25] J. Webster and C. Kit. Tokenization as the initial phase in nlp. pages 1106–1110, 01 1992.

[26] W. Zhou, N. Smalheiser, and C. Yu. A tutorial on information retrieval: Basic terms and concepts. *Journal of biomedical discovery and collaboration*, 1:2, 02 2006.

Abbildungsverzeichnis