

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

Automatisierte Analyse und visuelle Aufbereitung von Datenschutzerklärungen

Bachelorarbeit

im Studiengang Informatik

von

Lukas Köhler

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

Hannover, 15.07.2021

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 15.07.2021

Lukas Köhler

Zusammenfassung

Bei der Benutzung des Internets fällt heutzutage eine große Menge von Daten an. Bereits bei dem Besuch einer Internetseite werden Informationen erhoben. Viele dieser Daten sind personenbezogen und geben Details preis, die eine Person möglicherweise nicht gern teilt. Möchte der Betreiber einer Internetseite Informationen über seine Nutzer speichern, so muss er dies in einer Datenschutzerklärung kenntlich machen. Der Inhalt jener zieht sich jedoch häufig über mehrere Seiten und ist ohne juristisches Fachwissen nur begrenzt zu verstehen.

Angesichts dieser Tatsache soll ein Konzept entwickelt werden, welches den Benutzer dabei unterstützt, einen Überblick über eine Datenschutzerklärung zu erhalten. Das Konzept umfasst Wege zur automatischen Erkennung von Datenschutzerklärungen, Informationen darüber, wie man diese mithilfe von Algorithmen aus dem Bereich des maschinellen Lernens analysieren kann und Möglichkeiten, um die Ergebnisse der Analyse dem Anwender darzustellen. Das entwickelte Konzept wird als Browser Extension umgesetzt und bietet eine intuitive Nutzung, ohne zu sehr in den üblichen Arbeitsablauf des Nutzers einzugreifen. In dieser Arbeit wird ein besonderes Augenmerk auf die Erarbeitung des Konzepts gelegt. Zusätzlich wird die Umsetzung beleuchtet. Abschließend wird auf zukünftige, potentielle Wege eingegangen, wie man die Ergebnisse dieser Arbeit erweitern könnte. Die Evaluierung des Konzepts mithilfe einer Nutzerstudie mit 65 Teilnehmern zeigte einen positiven Einfluss auf die Bearbeitung von typischen Aufgaben in Bezug zu Datenschutzerklärungen. Es wurde ein durchschnittlicher SUS-Score von 87.95 ermittelt.

Abstract

Nowadays, a large amount of data is generated through usage of the internet. Information is collected as soon as you visit a website. Much of this data is personal and reveals details that a person may not wish to share. If the operator of a website wants to store information about its users, he must make this clear in a privacy policy. However, the content of these often extend over several pages and can only be understood to a limited extent without in-depth legal knowledge.

In view of this fact a concept has to be developed that supports the user in obtaining an overview of a privacy policy. The concept includes ways to automatically recognize privacy statements, information on how to analyze them using machine learning algorithms and ways to clearly present the results to the user.

The developed concept is implemented as a browser extension and offers intuitive usage without interfering too much with the user's usual workflow. In this thesis, special attention is paid to the development of the concept. Additionally, the implementation will be highlighted. Finally, future potential ways to extend the results of this work are discussed. The evaluation of the concept with the help of a user study with 65 participants showed a positive influence on the processing of typical tasks related to privacy statements. An average SUS score of 87.95 was calculated.

Inhaltsverzeichnis

1	Einleitung	1
1.1	Problemstellung	2
1.2	Lösungsansatz	2
1.3	Struktur der Arbeit	3
2	Grundlagen	5
2.1	Privatsphäre	5
2.2	Datenschutz	5
2.3	Erklärbarkeit	6
2.4	Natural Language Processing	6
2.4.1	Tokenisierung	7
2.4.2	Stemming	7
2.4.3	N-Gramme	8
2.5	Klassifikationsverfahren	8
2.5.1	Naive-Bayes	9
2.5.2	Random Forest	10
2.5.3	Performanz-Metriken	12
3	Konzept und Planung	15
3.1	Suche der Datenschutzerklärung	16
3.1.1	Suche der korrekten Verlinkung	16
3.1.2	Überprüfung eines DSE-Kandidaten	18
3.2	Analyse einer Datenschutzerklärung	19
3.2.1	Grundlegende Funktionsweise	19
3.2.2	Datenexploration	20
3.2.3	Entwicklung eines Klassifikationsalgorithmus	21
3.2.4	Klassifizierung einer Datenschutzerklärung	22
3.3	Visualisierung der Ergebnisse	23
3.3.1	Piktogramme	23
3.3.2	Benutzeroberfläche	24
3.3.3	Detailansicht	25
3.3.4	Weitere Funktionen	25
3.4	Anforderungen an die Implementierung	27

4	Implementierung	29
4.1	Architektur	29
4.2	Verwendete Technologien	30
4.2.1	Browser Extension	30
4.2.2	Scikit-learn	31
4.2.3	Natural Language Toolkit	31
4.2.4	Typescript	32
4.2.5	React	32
4.3	Umsetzung	32
4.3.1	Iteratives Entwicklungsmodell	32
4.3.2	Webserver	33
4.3.3	Analyse-System	33
4.3.4	Benutzeroberfläche	34
4.4	Portabilität	36
5	Evaluation	37
5.1	Zielsetzung	37
5.2	Forschungsfragen	37
5.3	Methodik	39
5.3.1	Technische Umsetzung	40
5.3.2	Versuchsaufbau	40
5.4	Resultate	41
5.4.1	Resultate für RQ1	42
5.4.2	Resultate für RQ2	44
5.4.3	Resultate für RQ3	47
5.4.4	Resultate für RQ4	48
6	Diskussion	51
6.1	Interpretation der Ergebnisse	51
6.2	Limitierung der Ergebnisse	53
6.3	Fazit	53
7	Verwandte Arbeiten	55
7.1	Polisis und PriBot	55
7.2	P3P	56
7.3	Privacy Bird	56
7.4	Zusammenfassung mittels Ensemble Learning	56
8	Zusammenfassung und Ausblick	57
8.1	Zusammenfassung	57
8.2	Ausblick	58
A	Liste der analysierten Internetseiten	61

B Skript zur Studie	67
B.1 Skript der Experimentalgruppe	67
B.1.1 Einführung	67
B.1.2 Phase 1	67
B.1.3 Phase 2	68
B.1.4 Phase 3	69
B.1.5 Phase 4	70
B.1.6 Abschluss	72
B.2 Skript der Kontrollgruppe	72
B.2.1 Einführung	72
B.2.2 Phase 1	72
B.2.3 Phase 2	73
B.2.4 Phase 3	74
B.2.5 Phase 4	74
B.2.6 Abschluss	75
C Installationsanleitung	77
C.1 Installation der Browser Extension	77
C.1.1 Kompilierung	77
C.1.2 Installation	77
C.2 Installation und Start des Analyse-Servers	78
D Inhalt des beiliegenden USB-Sticks	79

Kapitel 1

Einleitung

Mit jedem Aufruf einer Internetseite geht der Nutzer einen Vertrag mit dem Anbieter dieser ein, welcher in einer sogenannten Datenschutzerklärung (DSE) erläutert wird. An dieser Stelle wird beschrieben, welche Daten über den Anwender gespeichert werden und auf welche Weise diese ausgewertet werden dürfen. Solche DSEs erstrecken sich häufig über mehrere Seiten und kosten den Leser viel Zeit, wenn er sie vollständig lesen und verstehen möchte, so McDonald et al. [1]. Neben dem zeitlichen Aspekt fällt es vielen Nutzern, besonders ohne juristisches Fachwissen, schwer, den Inhalt einer DSE vollständig zu verstehen, wie es Jensen et al. [2] in ihrer Arbeit beschreiben.

Die Mengen an Daten, die erhoben werden, steigen täglich [3]. Vor diesem Hintergrund ist es wichtig, den Datenschutz in Zukunft nicht aus den Augen zu verlieren [4]. In der Forschung werden verschiedene Herangehensweisen verfolgt, um DSEs für den Endnutzer verständlich zu machen und sein Datenschutzbewusstsein zu stärken (s. Kapitel 7). In dieser Arbeit wird der Ansatz verfolgt, DSEs für den Nutzer visuell aufzubereiten, um so für mehr Verständnis im Hinblick auf den Datenschutz zu sorgen. Hierfür wird ein System konzeptioniert und entwickelt, welches DSEs mithilfe von Algorithmen aus dem Bereich des maschinellen Lernens analysiert und dem Nutzer visuell aufbereitet darstellt.

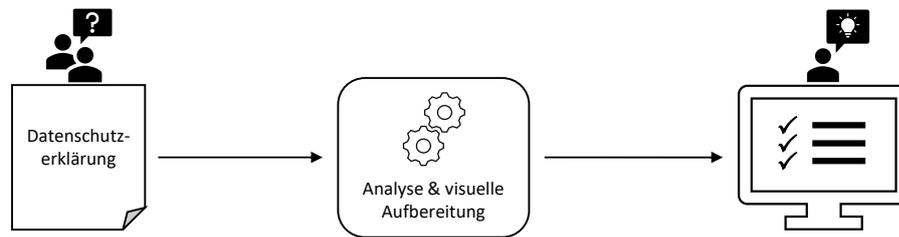


Abbildung 1.1: Überblick über den Nutzungsverlauf

In Abbildung 1.1 ist erkennbar, wie die Nutzung des Systems ablaufen soll. Der Anwender hat Probleme bei dem Verständnis einer DSE. Diese wird vom System verarbeitet und visuell aufbereitet dem Nutzer zur Verfügung gestellt.

1.1 Problemstellung

Bei der Nutzung des Internets fallen personenbezogene Daten an. Häufig wird jedoch nicht eindeutig mitgeteilt, wie bei der Sammlung, Verarbeitung und Weitergabe dieser vorgegangen wird. Somit kann nur schwer festgestellt werden, wie sich diese Vorgänge auf die Privatsphäre eines Nutzers auswirken. Unternehmen sind nach europäischem Recht dazu verpflichtet, eine DSE zu unterhalten, in welcher aufgeschlüsselt wird, wie der Konzern mit den Daten seiner Nutzer umgeht¹. Sie sollen als Entscheidungshilfe für diese dienen, wenn sie sich um ihren Datenschutz sorgen. Jedoch sind viele Menschen, besonders solche ohne juristischen Hintergrund, nicht selten überfordert sie in Gänze zu verstehen [2]. DSEs umfassen häufig mehrere Seiten und werden in einer Sprache zur Verfügung gestellt, die der Benutzer nicht versteht. Aus diesem Grund bieten sie keine sinnvolle Informationsquelle für Besucher einer Internetseite, die Interesse an ihrem Datenschutz haben.

1.2 Lösungsansatz

Um die gegebene Problemstellung zu bewältigen, soll im Rahmen dieser Arbeit ein System konzeptioniert werden, welches dem Nutzer die Möglichkeit bietet, einen Überblick über eine DSE zu erhalten. Für diesen Zweck soll eine gegebene DSE mithilfe von Algorithmen aus dem Bereich des maschinellen Lernens, genauer dem Natural Language Processing (NLP), analysiert und anschließend visuell aufbereitet dargestellt werden. Dieses System soll als Browser Extension (BE) implementiert werden, um den Nutzer in seiner täglichen Arbeit so wenig wie möglich zu stören und dennoch jederzeit abrufbar zu sein, wenn er sich Fragen im Bezug auf seine Privatsphäre stellt.

¹DSGVO Artikel 12 f.

Der Algorithmus für die Analyse der DSE wird mit einem Datensatz trainiert, welcher einzelne Abschnitte dieser in eine von zehn Kategorien einteilt. Eine nähere Einordnung der Informationen aus einer DSE kann aus diesem Grund nicht gegeben werden. Das Ziel dieser Arbeit ist es, DSEs englischsprachiger Internetseiten automatisiert zu analysieren und für den Nutzer visuell aufzubereiten, um ihm so einen Überblick über diese zu ermöglichen. Das soll dazu führen, dass der Nutzer ein besseres Gefühl im Hinblick auf seine Privatsphäre bekommt.

1.3 Struktur der Arbeit

Diese Arbeit ist folgendermaßen strukturiert: In Kapitel 2 werden die, für das Verständnis der Arbeit notwendigen, Grundlagen erläutert. Hier wird u. a. auf die Themen des *maschinellen Lernens* und *Datenschutzes* eingegangen. Im darauffolgenden Kapitel 3 wird die Planung und Konzeptionierung des Systems für die Zusammenfassung einer DSE beschrieben. Die Implementierung in Form einer BE wird in Kapitel 4 erläutert. Anschließend wird in Kapitel 5 beschrieben, wie eine Evaluation der implementierten Software mithilfe einer Nutzerstudie stattgefunden hat. Die Ergebnisse dieser werden in Kapitel 6 diskutiert. Weiter folgt in Kapitel 7 ein Ausblick auf Publikationen, die ein ähnliches Ziel wie diese Arbeit verfolgten. Abschließend wird in Kapitel 8 eine Zusammenfassung präsentiert und zusätzlich ein Ausblick gegeben, auf welche Weise das System in Zukunft erweitert werden könnte.

Kapitel 2

Grundlagen

In diesem Kapitel werden die wichtigsten Grundlagen dargestellt, welche für das Verständnis der Arbeit notwendig sind. Anfänglich wird erläutert, was Privatsphäre, Datenschutz und Erklärbarkeit im Rahmen dieser Arbeit bedeuten. Anschließend wird insbesondere auf die Themen der Verarbeitung von natürlicher Sprache und des maschinellen Lernens eingegangen.

2.1 Privatsphäre

Die Privatsphäre beschreibt den persönlichsten Bereich eines Menschen, in welchem es ihm erlaubt ist sich frei zu entfalten. Renaud et al. [5] definieren Privatsphäre auf folgende Weise: *„Privatsphäre ist das Recht einer Person, Grenzen zu definieren, diese zu erhalten und zu kontrollieren, welche festlegen, inwieweit die Gesellschaft mit diesen Grenzen interagieren und eingreifen darf. Gleichzeitig behält sie die volle Kontrolle über die von ihr erzeugten und auf sie bezogenen Informationen“*. Muss ein Mensch konstant eine Rolle in der Gesellschaft spielen und kann sich nicht in seiner Privatsphäre ausleben, so führt dies möglicherweise zu physischen und psychischen Problemen [6]. Durch die voranschreitende Entwicklung und Technisierung reduziert sich jedoch auch die Anzahl der Orte, die für eine Person als Rückzugspunkt angesehen werden können [6].

2.2 Datenschutz

Datenschutz bedeutet Bürgerinnen und Bürger vor unbefugter Erhebung, Speicherung und Weitergabe von Daten, die ihre Person betreffen, zu schützen [7]. Jeder Nutzer von moderner Technologie hinterlässt zunehmend mehr Daten im Internet, da es gibt kaum Interaktionen gibt, die keine digitalen Spuren hinterlassen [8]. Folglich wird es zunehmend schwerer die Kontrolle über die eigenen Daten zu behalten [9]. Viele dieser Daten werden zum Vorteil der Nutzer verwendet, beispielsweise um personalisierte

Vorschläge in Onlineshops anbieten zu können [10]. Jedoch nutzen viele Firmen diese Daten auch, um sie anschließend weiter zu verkaufen [11]. Aus diesem Grund wird Datenschutz oft auch als Verbraucherschutz aufgefasst [12]. Persönliche Daten werden in Zukunft nicht an Relevanz verlieren [4], dementsprechend ist es wichtig, Regeln für den Umgang digitaler Daten zu definieren, deren Einhaltung sicherzustellen und das Bewusstsein der Benutzer für den Datenschutz zu sensibilisieren [13].

2.3 Erklärbarkeit

Software wird in der heutigen Welt immer komplexer. Es werden Systeme genutzt, welche mithilfe von künstlicher Intelligenz Drohnen steuern [14], Tumore auf Röntgenbildern erkennen [15] oder Produktpreise dynamisch an die Konkurrenz anpassen [16]. Diese Systeme erfordern, besonders in den Bereichen wie dem Gesundheitswesen, wo Menschen mit ihnen in direktem Kontakt stehen, Erklärungen für ihre Arbeitsweise [17]. Aber auch Anwendungen, mit denen der normale Nutzer täglich in Berührung kommt, gewinnen an Komplexität. Ein Großteil der Benutzer wünscht sich Erklärungen für das Verhalten des Systems, wenn eine Reaktion ihren Erwartungen widerspricht [18]. Somit fällt der Erklärbarkeit in Software-Systemen eine immer wichtiger werdende Rolle zu [19].

2.4 Natural Language Processing

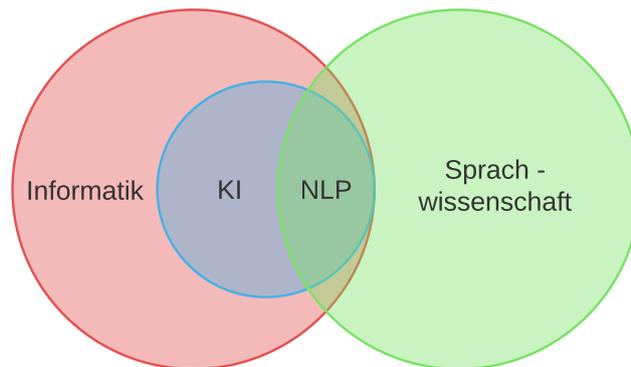


Abbildung 2.1: Venn-Diagramm Natural Language Processing

Das sogenannte *NLP* (dt. Computerlinguistik) ist ein Teilgebiet der Informatik, welches zu großen Teilen auf künstlicher Intelligenz beruht und die

Schnittstelle zu den Sprachwissenschaften darstellt. Es untersucht, wie natürliche Sprache mithilfe von Algorithmen analysiert und verarbeitet werden kann. In Abbildung 2.1 ist ein Venn-Diagramm dargestellt, welches aufzeigt, aus welchen Disziplinen das Natural Language Processing entstanden ist.

Natürliche Sprache, geschrieben oder gesprochen, stellt Computer seit jeher vor Probleme, denn sie beruht oft auf dem Kontext, in dem sie vorkommt. Beispielsweise könnte der Satz „Sarah folgte Max mit dem Fahrrad“ auf unterschiedliche Weisen interpretiert werden. Entweder fuhr Sarah mit dem Fahrrad, um Max zu folgen oder Max fuhr mit dem Fahrrad, während Sarah ihm folgte. Ohne die Intention des Sprechers oder einen anderen Kontext zu kennen, wie vorhergehende oder nachfolgende Sätze, ist es unmöglich diesen Satz korrekt zu deuten.

Aufgrund der komplexen und vielfältigen Struktur natürlicher Sprache ist es unmöglich, diese als System vollständig zu modellieren [20]. Daher ist es notwendig, gespeicherte Informationen so weit wie möglich herunter zu brechen, während die Semantik nicht verändert wird. Hierfür sollen folgende einige Wege beschrieben werden.

2.4.1 Tokenisierung

Der Prozess der Tokenisierung beschreibt die Segmentierung eines Quelltextes in Wörter (oder Sätze, Absätze, o.Ä.) an festgelegten Trennzeichen. So wird der Satz „Ach, was ein schöner Tag.“ in folgende Segmente aufgeteilt, wenn man an Satz- und Leerzeichen trennt:

„Ach“ | „was“ | „ein“ | „schöner“ | „Tag“

Um keine Informationen über den Kontext des Textes zu verlieren, ist es auch möglich, Satzzeichen mit einzubinden. In diesem Fall würde die anschließende Tokenisierung entstehen:

„Ach“ | „“ | „was“ | „ein“ | „schöner“ | „Tag“ | „.“

2.4.2 Stemming

Das Verfahren, um morphologische Variationen eines Wortes auf einen gemeinsamen Wortstamm zurückzuführen, wird in der Linguistik als Stemming (dt. Stammformreduktion) bezeichnet [21]. So können etwa die Begriffe „freiwillig“, „Freiwilliger“ und „Freiwilligkeit“ beispielsweise¹ auf den gemeinsamen Stamm „freiwill“ zurückgeführt werden. Solche Verfahren werden unter anderem eingesetzt, um die Ergebnisse von Suchmaschinen zu verbessern, da sie es ermöglichen, unterschiedliche Ausdrücke mit der gleichen Bedeutung auf ein einziges Wort zurückzuführen [22]. Dieses Vorgehen kann jedoch auch zu Problemen führen, wenn Wörter zwar

¹Bei Benutzung von <https://snowballstem.org/>

den selben Stamm, jedoch nicht die gleiche Bedeutung haben. Je nach genutztem Stemming-Algorithmus können beispielsweise die Wörter „*Maus*“ und „*Mauer*“ auf den gleichen Stamm „*Mau*“ zurückgeführt werden und verlieren somit ihre Bedeutung im Quelltext.

2.4.3 N-Gramme

Formale Definition

Sei Σ ein endliches Alphabet und sei $n \in \mathbb{N}$. Dann gilt:

Ist w ein Wort über dem Alphabet Σ , so ist w ein n -Gramm mit $w = (w_1, \dots, w_n) \in \Sigma^n$.

Erläuterung

Ein n -Gramm entsteht, wenn ein Quelltext in n aufeinanderfolgende Bestandteile aufgebrochen wird. Diese können, unter anderem, einzelne Wörter oder auch ganze Sätze sein. Die am häufigsten genutzten n -Gramme, sind Mono-, Bi-, Tri- und Tetragramme, welche respektive für 1-, 2-, 3- und 4-Gramme stehen. Wobei ein Monogramm nur ein einzelnes Element des Alphabets enthält. Um einen größeren Kontext abbilden zu können, bieten sich n -Gramme mit einem $n > 1$ an. Zerlegt man beispielsweise den Satz

„*Mara genießt die Sonne zur Mittagszeit.*“

in Trigramme anhand der Wörter, so entsteht die folgende Menge an Elementen:

1. „*Mara genießt die*“
2. „*genießt die Sonne*“
3. „*die Sonne zur*“
4. „*Sonne zur Mittagszeit*“

2.5 Klassifikationsverfahren

Bei Klassifikationsverfahren handelt es sich um Verfahren, welche es ermöglichen, neues Wissen aus bekannten Daten zu generieren. Um dieses Ziel zu erreichen, wird eine Menge an Eigenschaften eines Datenobjekts genau einer Klasse zugeordnet [23]. Bevor dies geschehen kann, muss der sogenannte Klassifikator trainiert werden. Hierfür wird ein bereits bekannter Datensatz, welcher die Eigenschaften eines Objekts und die dazugehörige Klasse enthält, genutzt, um dem Klassifikator abstrakte Muster für die Kategorisierung neuer Daten beizubringen. Anschließend kann dieser für genau diese Aufgabe angewendet werden. In diesem Kapitel werden einige genutzte Klassifikatoren erläutert.

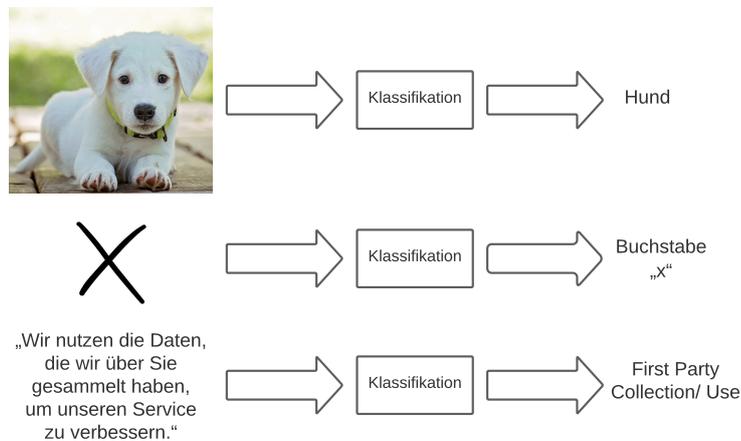


Abbildung 2.2: Einige Beispiele für Klassifikationen

In Abbildung 2.2 sind drei vereinfachte Beispiele für solche Klassifikationen dargestellt. Im ersten Teil der Abbildung sieht man das klassische Beispiel einer Bildklassifikation. Ein Algorithmus erhält als Eingabe Bilddaten eines Tieres. Anschließend ordnet er dieses Tier, anhand im Foto erkennbarer, signifikanter Eigenschaften, seiner Art zu. Im zweiten Beispiel ist ebenfalls eine Bildklassifikation zu erkennen. Hier wird jedoch versucht, handschriftlichen Text seiner Textcodierung in einem Computersystem zuzuordnen. Im letzten Beispiel ist eine Textklassifizierung zu sehen. Es wird ein Text, bereits in codierter Form gegeben, einer Kategorie zugeordnet.

2.5.1 Naive-Bayes

Bei dem *Naive-Bayes* Klassifizierer handelt es sich um einen Algorithmus, welcher auf dem Satz von Bayes beruht.

Satz von Bayes

Seien A, B zwei Zufallsereignisse und $P(B) > 0$.

Dann gilt:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \quad (2.1)$$

Dies bedeutet, dass sich die Eintrittswahrscheinlichkeit von A gegeben B unter anderem durch die Wahrscheinlichkeit das B unter der Bedingung A eingetreten ist, berechnen lässt.

Naive-Bayes-Klassifizierer

Basierend auf dem Wissen aus der Trainingsphase ordnet ein *Naive-Bayes-Klassifizierer* neue Daten der Klasse mit der höchsten Wahrscheinlichkeit zu [23]. Dieser Klassifikator geht von einer probabilistischen Unabhängigkeit der Eigenschaften einer Klasse aus. Obwohl dies in der Realität selten der Fall ist, liefert eine Klassifikation mittels Naive-Bayes häufig gute Ergebnisse [24].

2.5.2 Random Forest

Entscheidungsbäume

Ein Entscheidungsbaum (EB) ist ein geordneter und gerichteter Baum, welcher dazu dient, aufeinanderfolgende Entscheidungen zu repräsentieren und so die Beantwortung einer Fragestellung oder die Zuordnung von Objekten anhand von Eigenschaften in eine Klasse zu ermöglichen. Man beginnt bei der Wurzel des Baumes und betrachtet an jedem Knoten ein einzelnes Attribut. Durch dieses Attribut wird der Folgeknoten ausgewählt. Der Vorgang wird wiederholt, bis man an einem Blatt angelangt, welches die Antwort auf die eingangs gestellte Frage bzw. die Klassifikation des Objekts darstellt.

Beispiel eines Entscheidungsbaums

Attribute				Klassifizierung
Wetter	Luftfeuchtigkeit	Windig	Temperatur	Sport treiben
Sonnig	Normal	Ja	Heiß	Nein
Sonnig	Normal	Nein	Kalt	Nein
Sonnig	Hoch	Ja	Mild	Nein
Sonnig	Hoch	Nein	Mild	Nein
Regnerisch	Normal	Ja	Mild	Nein
Regnerisch	Normal	Nein	Mild	Ja
Regnerisch	Normal	Ja	Heiß	Nein
Regnerisch	Normal	Nein	Kalt	Ja
Bewölkt	Hoch	Ja	Kalt	Ja
Bewölkt	Hoch	Ja	Heiß	Ja
Bewölkt	Normal	Nein	Kalt	Ja
Bewölkt	Normal	Ja	Mild	Ja

Abbildung 2.3: Witterungsdaten

Abbildung 2.3 zeigt einen Datensatz in Form einer Tabelle. Sie enthält Witterungsdaten. Das momentane Wetter (Sonnig, Bewölkt, Regnerisch), die Luftfeuchtigkeit (Normal, Hoch), ob es windig ist (Ja, Nein) und die

momentane Temperatur (Kalt, Mild, Heiß) sind enthalten. Zusätzlich gibt es eine Spalte, welche die, zu der Witterung passenden, Klassifikation enthält und aussagt, ob es angebracht ist Sport zu treiben. Mithilfe dieses Datensatzes ist es möglich, einen Entscheidungsbaum zu erstellen. Dieser ist in Abbildung 2.4 zu sehen.

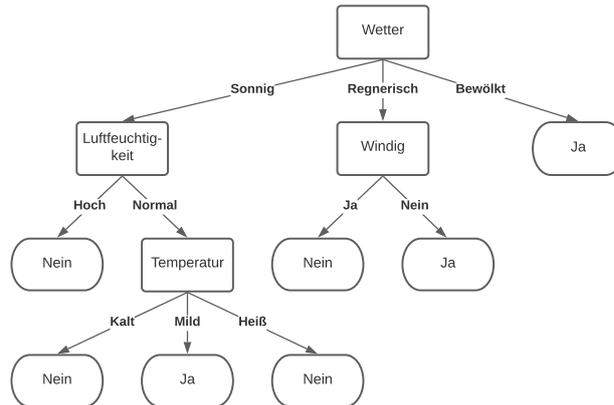


Abbildung 2.4: Der Entscheidungsbaum für das Sportbeispiel

Möchte man, dem Entscheidungsbaum unbekannte, Daten klassifizieren, so ist dies mit dem entstandenen Baum nun möglich. Übergibt man dem EB bspw. folgende Eingabe $E_1 = \{(Wetter = Sonnig), (Luftfeuchtigkeit = Normal), (Windig = Nein), (Temperatur = Mild)\}$ so kam diese nicht im Trainingsdatensatz vor. Der Baum kann diese, wie in Abbildung 2.5 dargestellt, trotzdem klassifizieren und kommt zu dem Ergebnis, dass die aktuelle Witterung zum Sport treiben geeignet ist.

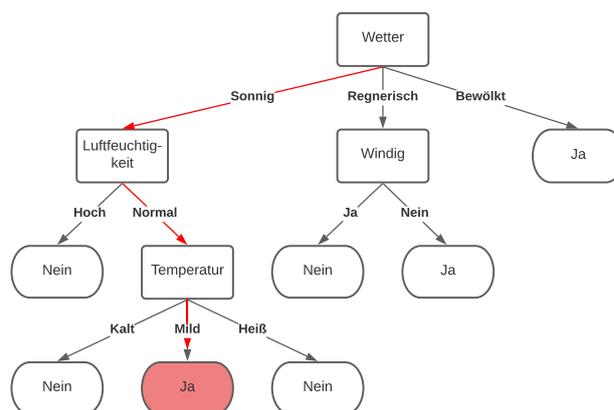


Abbildung 2.5: Klassifizierung von unbekanntem Daten

Random Forest

Ein *Random Forest* ist ein unkorrelierter Wald aus Entscheidungsbäumen. Dieses Verfahren ermöglicht es, einem Problem von Entscheidungsbäumen namens *Overfitting* zu entgegnen [25]. Sind die Trainingsdaten eines Entscheidungsbaums sehr komplex, so kann es passieren, dass dieser sich stark an die gegebenen Daten anpasst und die Ergebnisse sich, bei unbekanntem Daten, verschlechtern [26].

Ein Random Forest klassifiziert auf Basis einer Mehrheitsentscheidung. Jeder einzelne Baum trifft eine Entscheidung und jene Klasse, die am häufigsten ausgewählt wurde, stellt die endgültige Klassifizierung des Random Forests dar. Ein Random Forest Klassifizierer bietet bei bereits bekannten, sowie auch bei unbekanntem Daten eine bessere Einordnung als ein einzelner Entscheidungsbaum.

2.5.3 Performanz-Metriken

Ein Klassifikationsalgorithmus entscheidet sich möglicherweise nicht immer für die richtige Klasse des zu analysierenden Objekts. Soll die Qualität eines solchen Algorithmus bestimmt werden, so können für diesen Prozess unterschiedliche Metriken herangezogen werden.

In Tabelle 2.1 ist eine sogenannte *Konfusionsmatrix* dargestellt. Die in dieser Arbeit genutzten Klassifizierer teilen Daten in binäre Klassen ein. Das bedeutet, dass ein Datum entweder positiv oder negativ sein kann. Ordnet der Klassifizierer ein Objekt korrekt der positiven Klasse zu, so spricht man von einem *True Positive*. Berechnet der Algorithmus das selbe Ergebnis, die eigentliche Klasse ist aber negativ, so wird von einem *False Positive* gesprochen. Wird, mit der eigentlichen Klasse übereinstimmend, negativ klassifiziert, so ist dies ein *True Negative*, während eine negative Klassifizierung bei einem positiven Datum ein *False Negative* bedeutet.

		Actual	
		Positive	Negative
Prediction	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Tabelle 2.1: Konfusionsmatrix

Aufbauend auf diesen Metriken sind noch einige andere relevant für die Qualitätsmessung von Klassifikationsalgorithmen:

$$\text{Precision} = \frac{\sum \text{TP}}{\sum \text{TP} + \sum \text{FP}} \quad (2.2)$$

Die *Precision*-Metrik gibt an, wie viele, der als positiv kategorisierten, Objekte tatsächlich positiv sind.

Mithilfe der *Recall*-Metrik

$$\text{Recall} = \frac{\sum \text{TP}}{\sum \text{TP} + \sum \text{FN}} \quad (2.3)$$

kann berechnet werden, wie viele der positiven Objekte tatsächlich vom Algorithmus als positiv erkannt worden sind.

Letztlich ist der F_1 -Score zu nennen:

$$F_1\text{-Score} = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}} \quad (2.4)$$

Er ermöglicht die Kombination der Precision- und Recall-Metriken und berechnet sich durch das harmonische Mittel dieser.

Kapitel 3

Konzept und Planung

In diesem Kapitel wird ein Überblick über das gesamte Konzept des *Privacy Policy Check (PPC)* gegeben. Dafür wird zu Beginn dargestellt, wie der Nutzungsablauf im Ganzen ist. Dann wird ausführlich auf die einzelnen Schritte eingegangen und beschrieben, wie diese im Detail aussehen. Abschließend werden einige Anforderungen an die Implementierung des PPC erläutert.

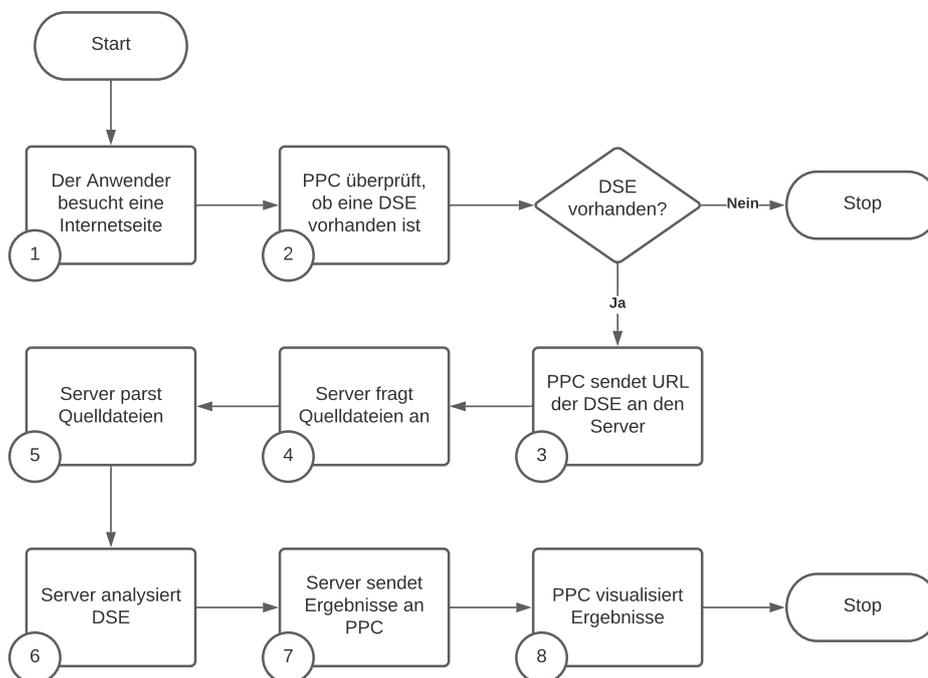


Abbildung 3.1: Der Nutzungsablauf des Privacy Policy Checks

In Abbildung 3.1 ist der genannte Ablauf zu sehen. Im ersten Schritt ruft der Anwender eine Internetseite auf (1). Nachfolgend wird durch den PPC überprüft, ob diese eine DSE enthält (2). Ist dies nicht der Fall, so stoppt der Vorgang. Ansonsten wird die gefundene URL an einen Server gesendet (3), welcher die Quelldateien der Internetseite anfordert (4). Als nächstes wird die Analyse der DSE vorbereitet, indem die geladenen Dateien geparkt werden (5). Nach dem Abschluss dieses Vorgangs werden die Texte analysiert (6). Im nächsten Schritt werden die Ergebnisse zurück an den PPC gesendet (7), welcher nun dafür zuständig ist, diese im letzten Schritt dem Nutzer darzustellen (8).

3.1 Suche der Datenschutzerklärung

In den folgenden Kapiteln wird beschrieben, wie bei der Suche nach einer DSE vorgegangen wird und wie daraufhin überprüft wird, ob es sich bei der gefundenen Internetseite tatsächlich um eine solche handelt. Für diesen Zweck wird darauf eingegangen, welche Maßnahmen getroffen worden sind, um Gemeinsamkeiten in der Verlinkung und Darstellung von DSEs zu finden. Zusätzlich werden zwei Algorithmen vorgestellt, welche hilfreich dabei sind, diese Aufgaben zu erfüllen.

3.1.1 Suche der korrekten Verlinkung

Besucht der Anwender eine Internetseite, so wird im zweiten Schritt (vgl. Abbildung 3.1) versucht, die DSE dieser zu finden. Da es kein einheitliches Format für die Verlinkung einer DSE auf einer Internetseite gibt, musste untersucht werden, auf welchem Weg diese möglichst zuverlässig gefunden werden kann.

Für diesen Zweck wurden händisch 100 englischsprachige Internetseiten analysiert, um Übereinstimmungen in der Art der DSE-Verlinkung zu finden. Die Auswahl der überprüften Seiten erfolgte über eine Auflistung der meistbesuchten Internetseiten¹, um sicherstellen zu können, dass diese auch Relevanz im täglichen Surfverhalten der Nutzer haben. Nachdem 100 englischsprachige Internetauftritte (vgl. Anhang A) ausgewählt worden sind, wurden sie manuell in einem Browser aufgerufen und auf Gemeinsamkeiten im Layout bzgl. der Verlinkung der DSE untersucht. Diese Untersuchung fand mit besonderem Blick auf Merkmale statt, welche ebenso durch einen Computer gefunden werden können. So wurde insbesondere darauf geachtet, auf welche Weise Links auf eine DSE beschriftet sind und wie die dahinter liegende URL aufgebaut ist. Diese Daten wurden festgehalten, um sie später miteinander vergleichen zu können und um Gemeinsamkeiten zu finden, welche einen verlässlichen Weg aufzeigen, die DSE einer Internetseite

¹<https://moz.com/top500>

automatisch zu finden.

Bei diesem Vorgang wurden folgende Daten gefunden: Auf 100% der untersuchten Seiten wurde das Wort „*privacy*“ (dt. Datenschutz) entweder innerhalb des Verzeichnispfads der URL oder der Subdomain gefunden. Oft wurde auch der Link zu der vorhandenen DSE mit diesem Wort versehen. Meist war es jedoch nicht das einzige Wort, welches genutzt wurde. Häufig stand es im Zusammenhang mit den Wörtern „*policy*“ (dt. Bestimmung), „*notice*“ (dt. Ankündigung) oder „*statement*“ (dt. Erklärung). Mit diesem Wissen wurde ein Algorithmus entwickelt, welcher zuverlässig in der Lage ist, die DSE einer Internetseite zu finden. Da es jedoch in seltenen Fällen vorkommen kann, dass eine Internetseite Verlinkungen zu mehreren DSEs enthält, bspw. bei Firmen, welche mehrere Produkte im Internet vertreiben, muss sich für eine von diesen entschieden werden. Aufgrund dessen, dass auf den meisten untersuchten Seiten, der kürzeste Link zu einer allgemeinen DSE geführt hat, wird in diesem Fall dieser Link verwendet.

Im folgenden ist der Algorithmus, welcher für die Suche der DSE verwendet wird, in Pseudocode beschrieben. Der beschriebene Algorithmus 1 arbeitet

Algorithmus 1 Algorithmus zur Suche einer Datenschutzerklärung

```

1: procedure GETPRIVACYLINK
2:   links ← getAllLinks()
3:   privacyLinks ← new List()
4:   for link ∈ links do
5:     if isPrivacyUrl(link) and link.text.contains("privacy") then
6:       privacyLinks.add(link)
7:     end if
8:   end for
9:   privacyPolicyLink ← getShortest(privacyLinks)
10:  return privacyPolicyLink
11: end procedure

```

auf folgende Weise. In Zeile zwei werden alle, auf der Seite vorhandenen, Links in einer Liste gespeichert. Anschließend wird eine neue Liste angelegt (3), welche die möglichen Kandidaten für den Link zu einer DSE enthält. Daraufhin wird über alle Links iteriert (4), wobei für jeden einzeln geschaut wird, ob dessen URL auf eine DSE hindeutet und ob der Text der Verlinkung das Wort „*privacy*“ enthält (5). Trifft dies nicht zu, so wird der gegebene Link ignoriert. Sollte jedoch ein Kandidat gefunden werden, so wird er in die Liste aufgenommen (6). Bevor der Link zur DSE in Zeile zehn zurückgegeben wird, wird in der vorhergehenden Zeile wie beschrieben der kürzeste Link aus den gefundenen Kandidaten ausgewählt.

3.1.2 Überprüfung eines DSE-Kandidaten

Wurde ein Link gefunden, so wird daraufhin überprüft, ob es sich bei der gegebenen Seite tatsächlich um eine DSE handelt. Dieser Schritt ist von essentieller Bedeutung, denn wird eine Seite als DSE erkannt, obwohl sie keine solche darstellt, werden die späteren Analyseergebnisse verfälscht. Das würde eventuell dazu führen, den Anwender weiter zu verwirren, anstatt ihn bei dem Verständnis einer vorliegenden DSE zu unterstützen. Um keine Internetseite irrtümlich als DSE einzustufen, wurden für diesen Schritt ebenfalls einige Vorüberlegungen angestellt. Demzufolge wurde auch hier die Liste der am häufigsten besuchten Internetseiten (vgl. Anhang A) begutachtet, um Gemeinsamkeiten in diesen zu finden. Es wurde festgestellt, dass häufig in *H1*- oder *Title*-Tags auf den Zweck der aktuell besuchten Internetseite hingewiesen wird. Im Falle einer DSE wird dies oft mit den Worten „*Privacy Notice*“, „*Privacy Statement*“, „*Privacy Policy*“, „*General Data Protection Regulation (GDPR)*“, „*Your Privacy*“ oder „*Data Protection*“ festgehalten. Diese Wörter werden im Folgenden als *Privatsphäreindikatoren* bezeichnet. Zusätzlich, wie bereits in Kapitel 3.1.1 erläutert, finden sich die DSEs häufig hinter einer URL, welche an einer oder mehreren Stellen das Wort „*privacy*“ enthält. Mit diesem Wissen ist es nun möglich, einen Algorithmus zu entwickeln, welcher verlässlich in der Lage ist, zu überprüfen, ob eine Internetseite eine DSE ist oder nicht. Das Ergebnis der Entwicklung wird mit Algorithmus 2 in Pseudocode beschrieben. Dieser

Algorithmus 2 Algorithmus zur Erkennung einer Datenschutzerklärung

```

1: procedure ISPRIVACYPOLICY(url)
2:   indicators  $\leftarrow$  0
3:   if isPrivacyUrl(url) then
4:     indicators  $\leftarrow$  indicators + 1
5:   end if
6:   for privacyIndicator  $\in$  privacyIndicators do
7:     for tag  $\in$  getAllH1Tags(url) do
8:       if tag.contains(privacyIndicator) then
9:         indicators  $\leftarrow$  indicators + 1
10:      end if
11:    end for
12:    for tag  $\in$  getAllTitleTags(url) do
13:      if tag.contains(privacyIndicator) then
14:        indicators  $\leftarrow$  indicators + 1
15:      end if
16:    end for
17:  end for
18:  return indicators  $\geq$  needed  $\triangleright$  needed  $\leftarrow$  2
19: end procedure

```

Algorithmus 2 arbeitet wie folgt beschrieben: In Zeile zwei wird eine Variable mit 0 initialisiert, welche die Indikatoren zählt, die für eine DSE sprechen. Anschließend wird überprüft, ob die gegebene URL für eine DSE spricht (3). Wenn dies der Fall ist, so wird der Zähler um eins erhöht (4). Ansonsten beginnt in Zeile sechs eine Schleife, welche über die Privatsphäreindikatoren iteriert und als erstes überprüft, ob ein H1-Tag solch einen Indikator enthält. Ist dies der Fall, so wird der Zähler inkrementiert (7-9). Anschließend wird das selbe Verfahren noch einmal für alle Title-Tags angewendet (12-14). Abschließend wird in Zeile 18 zurückgegeben, ob genug Indikatoren gefunden wurden, um davon ausgehen zu können, dass eine DSE vorliegt.

3.2 Analyse einer Datenschutzerklärung

In diesem Kapitel wird erläutert, wie die Analyse einer DSE im Detail abläuft. Aus diesem Grund wird zu Beginn auf die grundlegende Funktionsweise des Algorithmus eingegangen. Anschließend wird der genutzte Datensatz beleuchtet und die Konzeptionierung des eigentlichen Klassifikationsalgorithmus erläutert. Letztlich wird noch einmal der gesamte Vorgang einer Analyse dargestellt.

3.2.1 Grundlegende Funktionsweise

Bevor mit der Konzeptionierung eines Algorithmus zur Analyse von DSEs begonnen werden kann, muss als erstes dessen Arbeitsweise definiert werden. Möchte man einen Text analysieren, der in natürlicher Sprache vorliegt, gibt es für diese Aufgabe unterschiedliche Herangehensweisen. Hier sollen die zwei, welche für diese Arbeit in Betracht gezogen worden sind, gegenübergestellt und erläutert werden, warum sich für den Weg des maschinellen Lernens entschieden worden ist.

Nachteile statistischer Methoden

Zu Beginn der Konzeptionierungsphase wurde ein auf statistischen Methoden basierender Lösungsweg in Erwägung gezogen. Diese Ansätze liefern häufig gute Ergebnisse, ohne dass ein Datensatz benötigt wird, um einen Klassifizierer zu trainieren. So beschreiben Christian et al. [27] in ihrer Arbeit ein Verfahren, um Dokumente mithilfe des *Tf-idf*-Maßes zusammenzufassen, welches insgesamt gute Ergebnisse liefert. Allerdings zeigte eine frühere Arbeit inhärente Schwächen bei der Zusammenfassung von Texten in DSEs [28]. Aus diesem Grund wird ein Verfahren eingesetzt, welches auf Algorithmen aus dem Bereich des maschinellen Lernens setzt.

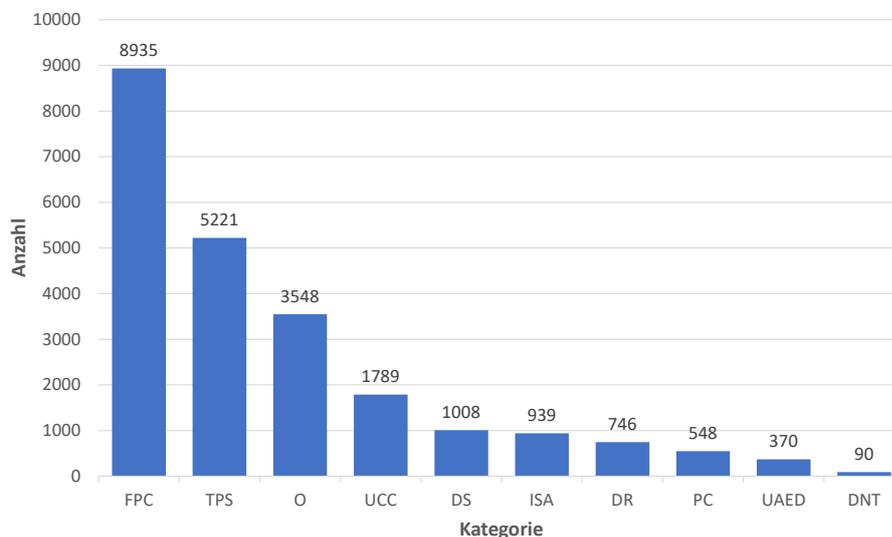


Abbildung 3.2: Anzahl der Einträge zu der jeweiligen Kategorie

Vorteile von maschinellem Lernen

Maschinelles Lernen bietet den Vorteil, dass dieses Verfahren unabhängig von der zugrundeliegenden Struktur einer DSE anwendbar ist, da der reine Text analysiert wird. Außerdem ist ein Klassifizierungsalgorithmus auf Basis von maschinellem Lernen frei von menschlichen Einflüssen, da er mithilfe von konkreten DSEs trainiert wird. Auf diese Weise können potentielle menschliche Fehler vermieden werden, die bei einer manuellen Eingabe der Parameter für eine Zusammenfassung einer DSE entstehen könnten. Zusätzlich bietet ein Naive-Bayes-Klassifizierer den Vorteil, dass er bereits mit wenig Trainingsdaten gute Ergebnisse erzielen kann [29].

3.2.2 Datenexploration

Ausgehend von der Funktionsweise des Systems wird ein Datensatz benötigt, um die Algorithmen für die Klassifikation zu trainieren. Für diese Arbeit wird der *OPP-115 Korpus* verwendet, welcher im Jahr 2016 von Wilson et al. für Lehr- und Forschungszwecke veröffentlicht wurde und 115 annotierte DSEs enthält [30]. Dabei handelt es sich um DSEs von englischsprachigen, US-amerikanischen Internetseiten. Jede einzelne dieser DSEs wurde von drei Doktoranden der Rechtswissenschaften bearbeitet, um den in der DSE beschriebenen Umgang mit Daten zu kategorisieren. Die Abschnitte der DSEs wurden in die folgenden zehn Kategorien eingeteilt:

First Party Collection (FPC) Informationen über Daten die vom Betreiber der Internetseite gesammelt werden und wie diese genutzt werden.	Third Party Sharing/ Collection (TPS) Informationen darüber, ob Daten mit Dritten getauscht werden und in welchem Umfang dies geschieht.
Data Security (DS) Informationen darüber, ob und wie vom Betreiber gespeicherte Daten geschützt werden.	User Access, Edit and Deletion (UAED) Informationen darüber, wie der Nutzer auf seine Daten zugreifen und diese bearbeiten oder löschen kann.
User Choice/ Control (UCC) Informationen darüber, ob und wie weit der Nutzer selber entscheiden kann, welche Daten über ihn gespeichert werden dürfen.	Data Retention (DR) Informationen darüber, wie und wie lange Daten seitens des Betreibers der Internetseite gespeichert werden.
Policy Change (PC) Informationen darüber, wie Veränderungen in der Datenschutzerklärung veröffentlicht werden.	Do Not Track (DNT) Informationen darüber, wie mit Do Not Track-Signalen umgegangen wird.
International and Specific Audiences (ISA) Informationen für Besucher aus dem Ausland und spezifische Personengruppen.	Other (O) Andere Informationen, wie beispielsweise Möglichkeiten zur Kontaktaufnahme mit dem Betreiber der Internetseite.

Tabelle 3.1: Die analysierten Kategorien

3.2.3 Entwicklung eines Klassifikationsalgorithmus

Um eine möglichst genaue Einordnung der Sätze in die, in Tabelle 3.1 aufgelisteten, Kategorien zu erreichen, wird für jede dieser ein eigener Klassifikationsalgorithmus trainiert. Auf diese Weise soll ein möglichst gutes Ergebnis bei der späteren Analyse von DSEs gewährleistet werden. Zu Beginn des Trainingsprozesses werden die Daten mithilfe eines Stemming-Algorithmus bearbeitet. Anschließend wird der Prozess der *Feature-Extraction* (dt. Gewinnung von Merkmalen) mithilfe eines sogenannten *CountVectorizers* [31] durchgeführt. Dies ermöglicht die spätere Verarbeitung durch Algorithmen. Daraufhin werden den ermittelten Features mittels des *Tf-idf-Maßes* [32] eine bestimmte Relevanz zugeordnet. Im abschließenden Schritt wird der Klassifikationsalgorithmus trainiert. In dieser Arbeit werden zwei verschiedene Algorithmen eingesetzt, je nachdem, wie gut die Ergebnisse waren, die während der Konzeptionierung geliefert worden sind. Einige Kategorien werden durch den Naive-Bayes-Algorithmus klassifiziert, während andere durch einen Random-Forest eingestuft werden.

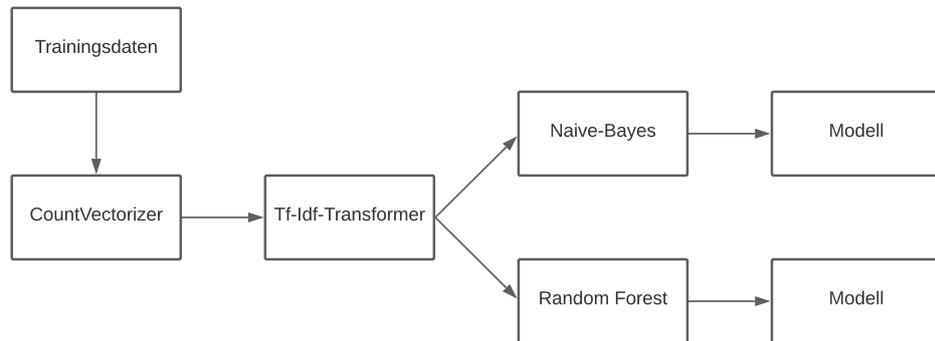


Abbildung 3.3: Trainings-Pipeline der Modelle

In Abbildung 3.3 ist der Trainingsprozess schematisch dargestellt. Die Trainingsdaten werden zu Beginn mithilfe eines *CountVectorizers* und *Tf-Idf-Transformers* bearbeitet. Daraufhin folgt, je nach Kategorie, das Training eines anderen Klassifikationsalgorithmus. Durch diesen Trainingsprozess entsteht ein Modell, das später für die Klassifikation eingesetzt werden kann.

3.2.4 Klassifizierung einer Datenschutzerklärung

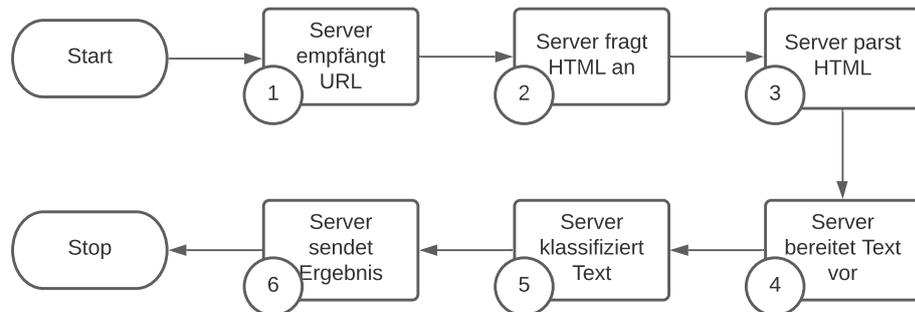


Abbildung 3.4: Die Analyse einer Datenschutzerklärung

Wird nach Abbildung 3.4 eine URL empfangen (1), so stößt das den Algorithmus zur Klassifizierung einer DSE an. Daraufhin werden als erstes die Quelldateien der DSE angefordert (2), um diese im nächsten Schritt für die Analyse vorzubereiten. Zu diesem Zweck wird als erstes überprüft, ob der jeweilige HTML-Tag für den Nutzer sichtbar ist oder nicht (3). Diese

Überprüfung findet statt, da in einem HTML-Dokument häufig Metadaten, wie beispielsweise die Codierung der Datei oder der Titel der Seite, enthalten sind. Solche Informationen sind für die Analyse irrelevant und können ignoriert werden. Nachdem dieser Prozess abgeschlossen ist, werden die Texte in Segmente mit einer Mindestlänge von 75 Zeichen aufgeteilt. Sollte ein Segment kürzer sein, so wird es an das vorherige Segment angehängt. Auf diese Weise soll ein gewisser Kontext beibehalten werden. Dieser ist bei der späteren Klassifikation von Vorteil, um ein Segment besser in die zugehörige Kategorie einteilen zu können. Anschließend wird der so resultierende Text weiter bearbeitet. Hierfür werden alle Satzzeichen entfernt, die enthaltenen Wörter in Minuskeln² umgewandelt und schließlich wird der Text mithilfe eines Stemming-Algorithmus in seine Stammform reduziert (4). Darauf folgt die eigentliche Klassifikation der Segmente (5). Zu diesem Zweck wird für jedes Segment und jede Kategorie berechnet, wie hoch die Wahrscheinlichkeit ist, dass das jeweilige Segment in die Kategorie passt. Übersteigt dieser Wert dabei einen Schwellenwert, welcher für jede Kategorie einzigartig ist, so wird angenommen, dass das Segment der Kategorie zugehörig ist. Sollten während der Verarbeitung fehlerhafte oder zu kurze Texte entstanden sein, so werden diese im letzten Schritt entfernt und anschließend wird das Ergebnis der Analyse zurück an den PPC gesendet (6).

3.3 Visualisierung der Ergebnisse

Der letzte aber dennoch sehr wichtige Schritt im Programmablauf, welcher in Abbildung 3.1 zu sehen ist, ist der achte. In diesem werden die Ergebnisse der Analyse visualisiert und dem Benutzer dargestellt. Hierbei kommt es auf eine klare, vereinfachte und durch eine geeignete Repräsentation dargestellte Visualisierung an, welche im Gegensatz zur ursprünglichen DSE die kognitive Last des Nutzers verringern soll [33].

3.3.1 Piktogramme

Um dem Anwender einen schnellen Überblick über eine DSE bieten zu können, sollen sogenannte Piktogramme eingesetzt werden. Bei Piktogrammen handelt es sich um Bilder, welche bei dem Verständnis einer Systemfunktionalität unterstützen [34] und dazu beitragen, DSEs vergleichbarer zu machen [35]. Daher werden sie in dieser Arbeit eingesetzt, um die Verständlichkeit der Analyseergebnisse zu verbessern und dem Nutzer zu helfen, sich in einer ihm unbekanntem DSE schneller zurechtzufinden. Sie werden von Anwendern besser verstanden als reine Texte und ermöglichen eine schnellere Erfassung gezeigter Inhalte, da es für Menschen einfacher ist, Bilder zu identifizieren als Texte zu lesen und aufzunehmen [36].

²Wörter in Kleinbuchstaben



Abbildung 3.5: Die genutzten Piktogramme

In Abbildung 3.5 ist eine Übersicht über alle verwendeten Piktogramme mit der zugehörigen Kategorie gegeben. Zu Beginn dieser Arbeit war es geplant, die Piktogramme einer anderen Initiative zu nutzen. Mittlerweile findet man diese aber nur noch im „Archiv des Internets“³. Aus diesem Grund wurde sich dafür entschieden, die Piktogramme eines schweizerischen Vereins⁴ einzusetzen. Dieser hat es sich zur Aufgabe gemacht, das Datenschutzbewusstsein von Internetnutzern zu verbessern. Sie verfolgen den Ansatz, den Betreibern von Internetseiten kostenfrei Icons zur Verfügung zu stellen, die diese nutzen können, um ihre DSE transparenter zu gestalten. Auf diese Weise sollen die Anwender einen schnelleren Überblick erlangen und die wichtigsten Informationen über die Verarbeitung ihrer Daten aufnehmen können.

3.3.2 Benutzeroberfläche

Um die in Abschnitt 3.3 genannten Ziele zu realisieren, soll eine graphische Benutzeroberfläche konstruiert werden, die sowohl hohe Affordance als auch Visibility aufweist. Nutzer sollen keinesfalls aufgrund von zu hohem Informationsgehalt und der damit einhergehenden kognitiven Last überfordert werden. Aus diesem Grund folgt die Benutzeroberfläche dem Prinzip „So wenig wie möglich, so viel wie nötig“.

Interagiert ein Anwender direkt mit dem System, so dient die Benutzeroberfläche als erster Einstiegspunkt. Wurde eine DSE analysiert, werden alle Kategorien aufgelistet die in dieser Erwähnung finden. Neben den Kategorien werden, wie in Unterabschnitt 3.3.1 erläutert, passende Piktogramme dargestellt. Diese sollen den Nutzer dabei unterstützen, das Gesehene schnell und einfach aufzunehmen und zu verstehen. Zusätzlich soll ein Button angezeigt werden, über welchen auf Wunsch weitere Detailinformationen

³<https://archive.org/>

⁴<https://www.privacy-icons.ch/>



Abbildung 3.6: Initiales Mockup der Benutzeroberfläche

abgerufen werden können. In Abbildung 3.6 ist ein initiales Mockup der Benutzeroberfläche zu sehen. Das System zeigt dem Nutzer die Kategorien an, die in der aktuellen DSE erwähnt worden sind. Neben diesen Kategorien befinden sich die, zu der jeweiligen Kategorie passenden, Piktogramme. Zusätzlich wird der Name des Systems angezeigt. Dies soll dazu dienen, dass der Nutzer jederzeit weiß, an welcher Stelle er sich im System befindet und welchen Zustand es momentan hat.

3.3.3 Detailansicht

Zusätzlich zu der eigentlichen Benutzeroberfläche wird eine Detailansicht angeboten. In dieser werden sämtliche Kategorien noch einmal übersichtlich neben ihrem zugehörigen Piktogramm angezeigt und eine kurze Erläuterung für jede einzelne gegeben. Darüber hinaus befindet sich auch eine Auflistung der, zu der jeweiligen Kategorie passenden, Texte in dieser Detailansicht. Sollte eine DSE Informationen über eine Kategorie enthalten, so wird das zugehörige Piktogramm eingefärbt, um das Vorhandensein von Informationen zu signalisieren. Diese Detailansicht soll dem Nutzer helfen, wenn er sich über die Kategorien, in welche die Abschnitte eingeteilt werden, informieren möchte. Zusätzlich ist sie dienlich, wenn er tiefere Informationen über die Sätze, die den Kategorien zugeordnet worden sind, erfahren möchte.

3.3.4 Weitere Funktionen

Das System bietet neben den bisher genannten Funktionen noch weitere. Unter anderem soll, wenn sich der Nutzer direkt auf einer DSE befindet, ein klickbares Icon neben jeder Kategorie angezeigt werden. Wählt der Nutzer

eines dieser Icons aus, so wird der, zu der angeklickten Kategorie passende Text innerhalb der DSE farblich hervorgehoben.

Abschließend ist eine Benachrichtigungsfunktionalität zu nennen. Ruft der Anwender eine Internetseite auf und es wird eine DSE gefunden, so wird, wenn die automatische Analyse abgeschlossen ist, eine Benachrichtigung ausgegeben. Diese soll dazu dienen, dem Nutzer mitzuteilen, dass der PPC neue Informationen für ihn bereithält. Es wäre auch möglich, diese direkt anzuzeigen. Allerdings soll der Nutzer nicht in seinem Arbeitsablauf gestört werden, daher wurde sich für diese Benachrichtigungsfunktion entschieden.

3.4 Anforderungen an die Implementierung

Um die Konzeptionierung abzuschließen, ist es von Vorteil, eine Anforderungsanalyse durchzuführen. Diese soll dabei unterstützen das Konzept auf eine möglichst effektive Weise umzusetzen. Es wurden für das Konzept einige explizite Anforderungen, mit besonderem Hinblick auf die visuelle Darstellung der Analyseergebnisse, gestellt. Daher liegt ein besonderer Fokus auf der Benutzbarkeit der Applikation. Zusätzlich spielt die Funktionalität eine wichtige Rolle, denn um den Benutzer bei dem Verständnis einer DSE zu unterstützen, sollte die Anwendung zuverlässig arbeiten.

Andere Qualitätsaspekte, wie beispielsweise die Sicherheit oder Kompatibilität, können nicht vernachlässigt werden. Jedoch muss hier kein Schwerpunkt gesetzt werden, denn die Implementierungsweise über eine BE schreibt bereits Grundregeln für diese vor, welche über die genutzten Schnittstellen implementiert sind.

Folgende grundlegende Anforderungen wurden hierbei in Gesprächen mit dem Betreuer dieser Arbeit ermittelt:

R1 Die Applikation soll einfach erlernbar sein, um auch neuen Nutzern die Möglichkeit zu geben, schnell einen Überblick über eine DSE zu erhalten.	R2 Das Programm muss einfache Möglichkeiten bieten, einen Überblick über eine DSE zu erhalten.
R3 Die Anwendung muss, soweit möglich, autonom arbeiten, um die Anzahl von Benutzerfehlern zu minimieren.	R4 Die wichtigen Informationen der DSE sollten klar erkennbar und schnell erreichbar sein.
R5 Da das Programm als BE entwickelt werden soll, ist ein modularer Aufbau der einzelnen Funktionen von besonderer Wichtigkeit.	R6 Die analysierten Sätze sollen korrekt in die dazugehörige Kategorie eingeordnet werden.
R7 Um die Applikation als BE in mehreren Browsern nutzbar zu machen, ist es wichtig, auf die genutzten Schnittstellen zu achten, um so die Übertragbarkeit auf andere Browser zu gewährleisten.	R8 Die Dokumentation des entwickelten Programmcodes ist von essentieller Wichtigkeit, um die Weiterentwicklung des Projekts zu ermöglichen.

Tabelle 3.2: Anforderungen an die Implementierung

Kapitel 4

Implementierung

In diesem Kapitel wird beschrieben, wie das im vorhergegangenen Kapitel dargestellte Konzept implementiert wurde. Hierfür wird zu Beginn auf die Architektur des Systems eingegangen. Daraufhin werden die verwendeten Technologien erläutert und erklärt, wie die explizite Umsetzung stattfand. Abschließend wird darauf eingegangen, wie die Portabilität der Applikation umgesetzt wurde.

4.1 Architektur

Die Architektur des PPC ist aus mehreren Schichten und Komponenten aufgebaut. Mit dem Ziel, den Anwender so wenig wie möglich in seinem täglichen Arbeitsablauf zu stören, wurde sich dafür entschieden, eine BE zu entwickeln. Somit kann der Anwender sich die Ergebnisse einer DSE-Analyse direkt im Browser anzeigen lassen, ohne die Applikation wechseln zu müssen. Diese Art der Implementierung bringt jedoch einige Restriktionen mit sich. So kann eine BE nicht den vollen Funktionsumfang des zugrundeliegenden Browsers oder Betriebssystems nutzen. Außerdem bringt eine solche Implementierung Beeinträchtigungen in Bezug auf die Performanz und die verfügbaren Bibliotheken mit sich. Aus diesen Gründen wurde sich dafür entschieden, die Analyse einer DSE (vgl. Abschnitt 3.2) auf einen Server auszulagern. Dieser wurde in Form einer *Flask-Applikation* entwickelt. Die BE ist in dieser Architektur für die Suche einer DSE (vgl. Abschnitt 3.1) und die Darstellung der Analyseergebnisse (vgl. Abschnitt 3.3) zuständig.

In Abbildung 4.1 ist ein Überblick über die Architektur des Systems gegeben. Die für diese Arbeit relevanten Teile sind in blau dargestellt, während die eigenständig entwickelten noch einmal dunkler eingefärbt wurden. Eine BE wird im Rahmen des Browsers ausgeführt. Sie kann auf einige Schnittstellen des jeweiligen Browsers zugreifen. Ein Anwender kann mehrere Extensions innerhalb eines Browsers installieren, ohne dass diese miteinander interagieren können. Besucht der Nutzer des PPC eine

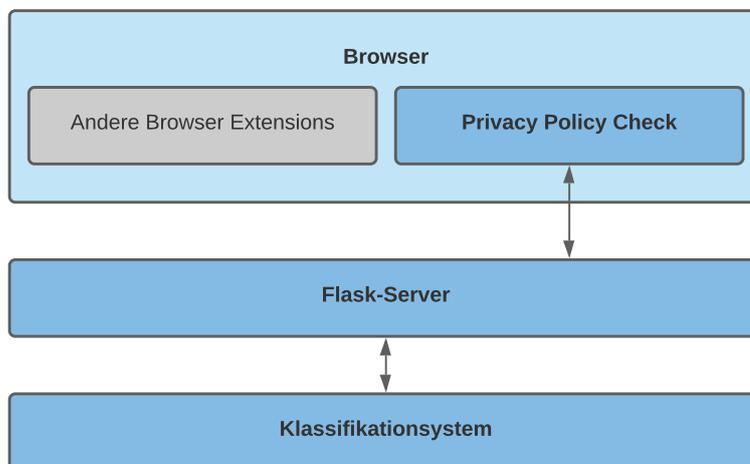


Abbildung 4.1: Die Architektur des Systems

Internetseite, so wird überprüft, ob diese eine DSE enthält. Ist diese Suche erfolgreich, so spricht der PPC die Flask-Applikation an, welche wiederum die Klassifikation anstößt. Nachdem auch dieser Prozess abgeschlossen ist, werden die Ergebnisse an die Flask-Anwendung zurückgegeben. Diese sendet die Ergebnisse zurück an den PPC, welcher daraufhin dafür zuständig ist, diese darzustellen.

4.2 Verwendete Technologien

Für die Umsetzung des vorgestellten Konzepts wurde eine Vielzahl von Technologien genutzt. Die relevantesten werden im Folgenden vorgestellt.

4.2.1 Browser Extension

Eine BE ist ein Programm, welches die Funktionalität eines Browsers erweitert. Sie sind in mehrere Dateien aufgeteilt, welche innerhalb des Programmablaufs unterschiedliche Rollen spielen. Folgend soll eine kurze Einführung in diese gegeben werden.

Im sogenannten **Manifest** einer BE werden Metadaten über diese festgehalten. So findet man in dieser Datei unter anderem den Namen der BE und Informationen über dessen Entwickler. Zusätzlich wird hier die gesamte Struktur der Extension festgelegt und beschrieben, welche Berechtigungen vom Browser benötigt werden, um die Funktionalität zu erfüllen.

Viele Aufgaben, für die eine BE zur Problemlösung eingesetzt werden kann, benötigen einen Service, welcher konstant im Hintergrund läuft. Dieser kann

bei Bedarf angesprochen werden und zur Lösung beitragen. Für diesen Zweck sind sogenannte **Background-Skripte** vorhanden, welche bei der Installation einer Extension im Browser gestartet und anschließend konstant im Hintergrund ausgeführt werden.

Content-Skripte werden im Kontext der Internetseite ausgeführt, die der Nutzer der BE aufruft. Da sie somit Zugriff auf das *Document Object Model* (DOM) der jeweiligen Seite haben, können sie Details aus diesem auslesen, Informationen an die eigentliche Extension weiterreichen und das DOM zusätzlich verändern.

Das **Popup** einer BE ist die graphische Benutzeroberfläche, in welcher Informationen dargestellt werden. Es dient als Einstiegspunkt in die Interaktion zwischen Extension und Anwender.

Eine **direkte Kommunikation** zwischen einzelnen Skripten ist nicht möglich. Um Informationen übertragen zu können, wird das sogenannte *Message-Passing* (dt. Nachrichtenaustausch) Paradigma verwendet. Das bedeutet, dass einzelne Nachrichten explizit gesendet und empfangen werden müssen, um Daten übergeben zu können.

Eine andere Möglichkeit, Daten zu speichern, ist über den sogenannten lokalen Speicher gegeben. Dieser hat Ähnlichkeiten mit einem Datenbanksystem. Skripte können hier Informationen persistent abspeichern, später auf diese zugreifen und außerdem benachrichtigt werden, wenn sich Informationen im Speicher ändern.

4.2.2 Scikit-learn

Scikit-learn ist eine freie, plattformunabhängige Bibliothek für Algorithmen aus dem Bereich des maschinellen Lernens [37]. Sie ist für die Programmiersprache *Python* entwickelt worden und unterstützt verschiedene Klassifikationsverfahren, unter anderem *Random Forest* und *Naive Bayes*, welche für diese Arbeit relevant sind. Es werden sogenannte Pipelines zur Verfügung gestellt, die es ermöglichen, mit gegebenen Trainingsdaten die Algorithmen für die Klassifikation auf eine schnelle und einfache Weise zu trainieren.

4.2.3 Natural Language Toolkit

Das *Natural Language Toolkit* (NLTK) ist eine Kollektion mehrerer Bibliotheken für *Python*, welche sich mit dem Thema der *Computerlinguistik* befassen [38]. Es existieren viele verschiedene Bibliotheken, die eine ähnliche Funktionalität und Performanz bieten [39]. Es wurde sich für NLTK entschieden, da diese Bibliothek eine einfach anzusprechende Schnittstelle für Stemming-Algorithmen bietet.

4.2.4 Typescript

Bei *Typescript* handelt es sich um eine, auf dem *ECMAScript-6*-Standard basierende, Programmiersprache. Sie stellt eine Obermenge von Javascript dar und erweitert es, unter anderem, mit statischer Typisierung. Dies ermöglicht eine schnellere Erkennung und Entfernung von Fehlern im Programmcode [40]. Bevor Typescript ausgeführt werden kann, wird es mithilfe eines Compilers zu gewöhnlichem Javascript kompiliert, um die Kompatibilität mit Browsern zu gewährleisten [40].

4.2.5 React

React ist ein Javascript-Framework für die Entwicklung von Benutzeroberflächen. Es erleichtert die Erstellung dieser auf mehrere Weisen. Das Framework ermöglicht die Definition von Komponenten, welche im gesamten Projekt wiederverwendet werden können. Zusätzlich stehen bereits viele solcher Komponenten als vordefinierte Bibliotheken zur Verfügung, welche einfach eingebunden werden können. Als Beispiel einer solchen Bibliothek ist *Material-UI*¹ zu nennen. Diese implementiert das von Google entworfene *Material Design*² in React. Einige Material-UI-Komponenten wurde auch für die Entwicklung der Benutzeroberfläche des PPC genutzt.

4.3 Umsetzung

In diesem Kapitel wird beschrieben, wie das in Kapitel 3 entwickelte Konzept konkret umgesetzt worden ist.

4.3.1 Iteratives Entwicklungsmodell

Da sich die Architektur des Systems aus mehreren, aufeinander aufbauenden, Bestandteilen zusammensetzt, wurde bei der Entwicklung iterativ vorgegangen. Auf diese Weise wurde sichergestellt, dass jede Komponente korrekt arbeitet, bevor sie von einem späteren Prozessschritt benötigt wird.

1. **Webserver:** Zu Beginn wurde ein Webserver entwickelt, welcher für die Kommunikation zwischen der BE und dem Analyse-System zuständig ist. Dieser sendet und empfängt Nachrichten im HTTP-Format. Wird eine Nachricht ausgehend von der BE empfangen, so stößt der Server die Analyse der DSE an.
2. **Analyse-System:** Nachdem der Webserver lauffähig war, wurde damit begonnen, das System zur Analyse einer DSE aufzusetzen. Die für die Analyse zuständigen Elemente befinden sich in einer logisch

¹<https://material-ui.com/>

²<https://material.io/design>

getrennten Einheit, welche bei Bedarf durch den Server angesprochen wird.

3. **Browser Extension:** Anschließend wurde die BE entwickelt. Damit diese vollständig funktionsfähig ist, muss der im ersten Schritt entwickelte Webserver ansprechbar sein. Zusätzlich muss das Analyse-System zur Verfügung stehen.

Die in dieser Auflistung beschriebenen Komponenten der Architektur werden in den folgenden Unterabschnitten noch intensiver beleuchtet.

4.3.2 Webserver

Da der eigentliche Webserver in der genutzten Architektur eine untergeordnete Rolle spielt und nur eine Schnittstelle für die Kommunikation zwischen der BE und dem Analyse-System zur Verfügung stellen muss, wurde nach einer Applikation mit wenig *Overhead* gesucht. Im Ökosystem der Programmiersprache Python, in welcher das eigentliche Analyse-System geschrieben wurde, existiert eine Vielzahl solcher Anwendungen. Für diese Arbeit wurden die beiden Webframeworks *Flask* und *Django* in Betracht gezogen. Während beide Systeme Vor- und Nachteile mit sich bringen, hat man sich für Flask entschieden. Diese Entscheidung fiel insbesondere aus zwei Gründen. Im Vordergrund stand, wie bereits erläutert, der zu vermeidende *Overhead*. Eine Django-Applikation bringt von sich aus bereits deutlich mehr Funktionalitäten mit, die für den hier angestrebten Einsatzzweck nicht notwendig sind und somit *Overhead* darstellen. Zusätzlich ist eine Flask-Applikation bereits mit sehr wenig geschriebenem Quelltext einsatzfähig und bietet die Möglichkeit, durch einen modularen Aufbau und viele verfügbare Bibliotheken, ohne größere Hindernisse in Zukunft erweitert werden zu können.

4.3.3 Analyse-System

Das Analyse-System besteht aus drei verschiedenen Komponenten, die allesamt eine wesentliche Rolle in der Analyse einer DSE spielen. Zu Beginn wird das Modul *Privacy Check* durch den Webserver angesprochen. Dieses ist dafür zuständig, den eigentlichen Inhalt der DSE zu laden. Anschließend wird der Quelltext der Internetseite mithilfe des *Text Parser*-Moduls geparkt. In diesem Schritt wird das HTML entfernt und der Stemming-Algorithmus angewendet. Ist dieser Vorgang abgeschlossen, beginnt die eigentliche Klassifikation der DSE im Modul *Classifier* wie in Unterabschnitt 3.2.4 beschrieben. Die Klassifikationsalgorithmen arbeiten jedoch nicht ohne Fehler. So wurde exemplarisch für das der *First Party Collection*-Kategorie zugehörige Modell die in Tabelle 4.1 ersichtlichen Metriken berechnet. Aus diesem Grund wurden für jede klassifizierbare Kate-

Zugehörigkeit	Precision	Recall	F ₁ -Score
Nicht zugehörig	0.94	0.94	0.94
Zugehörig	0.88	0.86	0.87

Tabelle 4.1: Berechnete Metriken

gorie zwei Schwellenwerte festgelegt. Einer dieser Schwellenwerte beschreibt die durch den Klassifikationsalgorithmus berechnete Wahrscheinlichkeit für die Zugehörigkeit eines Abschnitts zu einer Kategorie. Der andere legt die Mindestanzahl an Abschnitten fest, die zu einer jeweiligen Kategorie zugeordnet werden müssen, um als relevant gewertet zu werden. Diese beiden Werte sind für jede Kategorie einzeln festgelegt. Sie wurden gewählt, um die Anzahl der falschen Klassifikationen zu minimieren und die Anforderung **R6** zu erfüllen.

4.3.4 Benutzeroberfläche

Folgend wird, aufbauend auf dem in Abschnitt 3.3 entwickelten Konzept, die Benutzeroberfläche des PPC beschrieben.

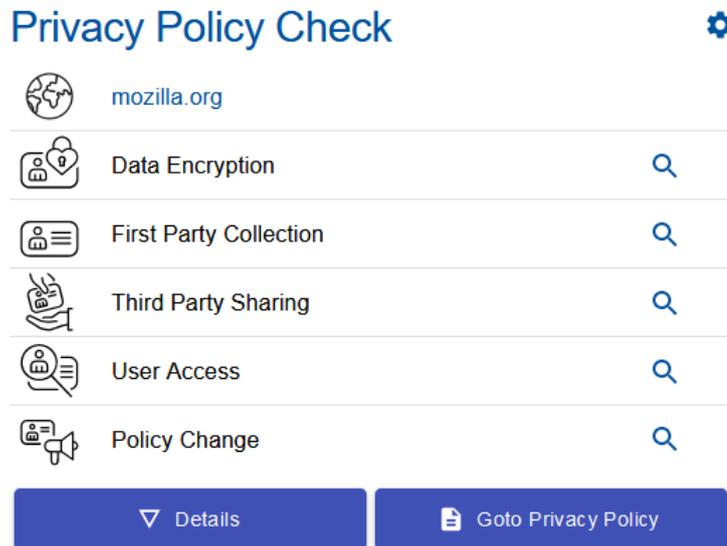


Abbildung 4.2: Die Benutzeroberfläche des Privacy Policy Check

In Abbildung 4.2 ist diese zu sehen. Für die Realisierung wurde auf das Framework *React* zusammen mit *Typescript* gesetzt. Die Grundstruktur wurde, im Vergleich zum entwickelten Konzept (vgl. Abbildung 3.6), beibehalten,

um die dort beschriebene Funktionalität zu bewahren und die Anforderungen **R1**, **R2** und **R4** zu erfüllen. Jedoch wurden einige Funktionen hinzugefügt und andere Details geändert. Es werden die in Unterabschnitt 3.2.2 vorgestellten und innerhalb einer DSE gefundenen Kategorien präsentiert. Außerdem werden die in Unterabschnitt 3.3.1 beschriebenen Piktogramme genutzt, um die Nutzer bei dem Verständnis der angezeigten Texte zu unterstützen. Neben diesen Kategorien lässt sich zusätzlich ein Lupensymbol finden. Mit einem Klick auf dieses wird der, der Kategorie zugehörige, Text innerhalb der DSE markiert. Zum Konzept hinzugekommen ist der „Goto Privacy Policy“ Button, welcher den Benutzer direkt auf die DSE der momentan besuchten Internetseite führt. Zusätzlich gibt es in der rechten oberen Ecke eine Schaltfläche, welche die Optionen des PPC öffnet.

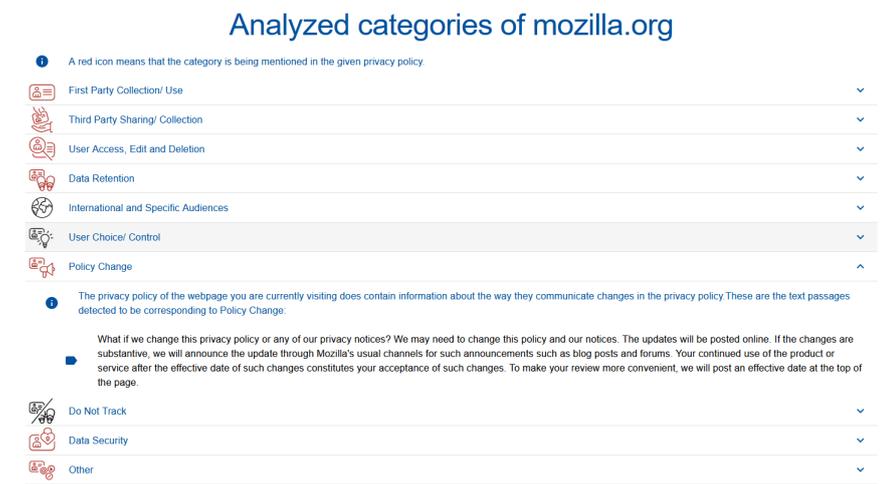


Abbildung 4.3: Der Detailbereich des Privacy Policy Check

In Abbildung 4.3 ist die *Details*-Seite des PPC dargestellt. Hier wird eine Auflistung aller Kategorien für den Nutzer präsentiert. Solche, die in der DSE der momentan besuchten Internetseite gefunden wurden, werden durch ein rotes Piktogramm markiert. Alle Weiteren werden durch ein schwarzes repräsentiert. Über dieser Auflistung befindet sich ein kurzer Informationstext, der den Nutzer über den Inhalt der gerade besuchten DSE aufklären soll. Mit einem Klick auf den jeweiligen Namen der Kategorie kann diese aufgeklappt werden. Wenn dies geschieht, zeigt sich auch hier noch einmal ein weiterer Informationstext. Dieser dient dazu, die Bedeutung der jeweiligen Kategorie zu erklären. Unter dieser Erklärung befindet sich eine weitere Liste. Hier werden alle Abschnitte der momentanen DSE aufgeführt, die als zu der jeweiligen Kategorie zugehörig erkannt worden sind.

4.4 Portabilität

Eine Anforderung an die Implementierung war es, dass die entwickelte BE im Firefox-Browser lauffähig ist. Mit Hinblick auf die Vielfalt der von Anwendern genutzten Browser wurde jedoch von Beginn an darauf geachtet, die Applikation portabel zu gestalten (vgl. Abschnitt 3.4) und somit **R7** zu erfüllen. Aus diesem Grund wurde ein sogenannter *Polyfill* verwendet. Hierbei handelt es sich um einen Programmcode, der Funktionen, die manche Browser nicht nativ anbieten, nachrüsten soll. Auf diese Weise ist es möglich, die Applikation in den Browsern Firefox, Chrome und Opera bereitzustellen.

Weiterhin wurde die, für die Analyse einer DSE zuständige, Flask-Applikation auf einen Server ausgelagert. Möchte ein Anwender Informationen über eine DSE erhalten, so spricht die entwickelte BE diesen an. Dementsprechend müssen die Nutzer keinen weiteren Service starten, sondern können ohne Störung mit ihrem täglichen Arbeitsablauf beginnen.

Wie in Tabelle 4.2 ersichtlich, wurde die Funktionalität in allen drei genannten Browsern unter einem Windows 10-System der Version 10.0.19043 überprüft. Zusätzlich wurde die Portabilität der BE auf einem Linux Pop!_OS und einem MacOS System im Firefox-Browser getestet.

Betriebssystem (Version)	Browser (Version)
Windows 10 (10.0.19043)	Firefox (89.0.2)
Windows 10 (10.0.19043)	Chrome (91.0.4472.124)
Windows 10 (10.0.19043)	Opera (77.0.4054.203)
Linux Pop!_OS (21.04)	Firefox (89.0.2)
MacOS (11.4)	Firefox (89.0.2)

Tabelle 4.2: Die getesteten Systemkonfigurationen

Kapitel 5

Evaluation

In diesem Kapitel wird die Evaluation des entwickelten Konzeptes vorgestellt. Für diesen Zweck wurde eine Nutzerstudie durchgeführt. Zu Beginn wird die Zielsetzung erläutert und zusätzlich die Forschungsfragen (eng. Research Questions (RQs)) beschrieben, die mithilfe dieser beantwortet werden sollen. Als nächstes wird auf die Methodik eingegangen. Letztlich werden die Ergebnisse vorgestellt.

5.1 Zielsetzung

Die Zielsetzung dieser Studie wurde mithilfe der Vorlage aus der Arbeit von Wohlin et al. [41] entwickelt.

Es wird das entwickelte Konzept **analysiert um** den Einfluss und die Wahrnehmung einer software-seitigen Unterstützung für Datenschutzrichtlinien zu bewerten **mit Bezug** auf Bedienbarkeit, Effektivität und Verständlichkeit **aus der Sicht von** Endbenutzern **im Kontext einer** Studie, die über das Internet durchgeführt und die Think-Aloud-Methode [42] genutzt wird, in Anwesenheit eines Experimentators.

5.2 Forschungsfragen

Mithilfe des *Goal-Question-Metric* Paradigmas wurden RQs von der Zielsetzung abgeleitet. Sie sollen dazu beitragen, die Zielsetzung erfüllen zu können. In Abbildung 5.1 sind die, zu der jeweiligen RQ evaluierten, Metriken abgebildet. Zusätzlich ist in der linken oberen Ecke eine Legende vorhanden. Ein durchgezogener Strich von einer RQ zu einer Metrik bedeutet, dass diese Metrik für die RQ relevant ist.

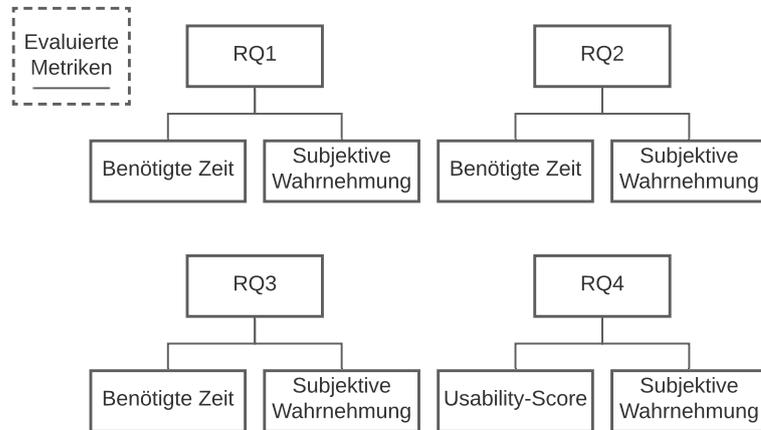


Abbildung 5.1: Die Forschungsfragen und die dazugehörigen Metriken

RQ1 Unterstützt die Browser Extension den Nutzer, die Datenschutzerklärung einer Internetseite schneller zu finden?

Die meisten Internetseiten pflegen eine DSE. Es gibt jedoch keine Vorgabe, auf welche Weise diese verlinkt werden müssen. Aus diesem Grund benötigen Nutzer häufig bereits viel Zeit, um eine DSE zu suchen, bevor sie überhaupt damit beginnen können, diese auszuwerten. Somit hören Nutzer, die an ihrem Datenschutz interessiert sind, möglicherweise bereits bei diesem Schritt auf und nutzen die Internetseite, ohne sich vorher in der DSE informiert zu haben. Um evaluieren zu können, ob die entwickelte BE einen Einfluss auf die Zeit hat, die ein Nutzer benötigt, um eine DSE zu finden, wurde die Zeit gemessen, die der Proband für diese Aufgabe benötigt. Zusätzlich wurden den Teilnehmern innerhalb der Studie einige Fragen diesbezüglich gestellt, um das Empfinden einfangen zu können.

RQ2 Unterstützt die Browser Extension dabei, gesuchte Informationen in einer Datenschutzerklärung schneller zu finden?

Interessiert sich ein Nutzer nicht für die gesamte DSE, sondern möchte nur bestimmte Aspekte der Datenverarbeitung in Erfahrung bringen, so stellt ihn diese Aufgabe möglicherweise vor eine Herausforderung. Viele DSEs bieten kein Inhaltsverzeichnis an, welches die Anwender nutzen könnten, um direkt zu dem Punkt zu springen, der sie interessiert. Daher wurde den Teilnehmern der Studie in beiden Gruppen (vgl. Unterabschnitt 5.3.2) die Aufgabe gestellt, nacheinander zwei Informationen innerhalb der DSE von Mozilla und Netflix zu suchen. Durch diese Aufgabe soll untersucht werden,

ob das entwickelte Konzept die Zeit beeinflusst, die ein Nutzer benötigt, um Informationen innerhalb einer DSE zu suchen. Auch hier wurden die Teilnehmer zusätzlich nach ihrem subjektiven Empfinden befragt, nachdem sie die Aufgabe erfüllt hatten.

RQ3 Unterstützt die Browser Extension den Nutzer bei dem Verständnis einer Datenschutzerklärung?

DSEs sind für den durchschnittlichen Internetnutzer, ohne juristischen Hintergrund, häufig schwer zu verstehen [2]. Mit dem Ziel, den Anwender bei dem Verständnis einer DSE zu unterstützen, wurde die, in dieser Arbeit vorgestellte, BE entwickelt. Den Teilnehmern der Experimentalgruppe wurden in den Phasen 2 bis 4 (vgl. Unterabschnitt 5.3.2) unterschiedliche Fragen zum Nutzen der BE gestellt. Der Kontrollgruppe wurden Fragen gestellt, die das Verständnis einer DSE ohne technische Unterstützung evaluieren sollen. Die Antworten auf diese wurden analysiert, um den Einfluss der BE beurteilen zu können. Zusätzlich spielt die, für RQ1 und RQ2 gemessene Zeit, für die Beantwortung dieser RQ eine entscheidende Rolle.

RQ4 Empfinden die Nutzer die Browser Extension als einfach zu bedienen?

Um diese RQ beantworten zu können, wurden die Probanden nach der Bearbeitung der Aufgaben gebeten, den *SUS*-Fragebogen [43] auszufüllen. Dieser wurde als eine „quick and dirty“ Evaluierungsmethode für die Bedienbarkeit von Systemen entworfen, denn Bedienbarkeit kann nicht im absoluten Sinne gemessen werden. Sie muss in Anbetracht des Kontexts und Anwendungsbereichs der jeweiligen Applikation ermittelt werden. Der Fragebogen besteht aus zehn Fragen, die anhand einer 5-Punkte-Likert-Skala beantwortet werden. Mithilfe der Ergebnisse dieses Fragebogens lässt sich der sogenannte *SUS-Score* berechnen, welcher sich im Intervall [0, 100] bewegen kann. Es wird die Bewertung der einzelnen Fragen aufsummiert, wobei die schlechteste Antwort mit einer 0 und die bestmögliche mit einer 4 codiert wird. Daraufhin wird diese Summe mit 2.5 multipliziert, um so auf das finale Ergebnis zu kommen. Dieses stellt explizit **keinen** Prozentwert, sondern eine dimensionslose Kennzahl dar [44]. Im Anhang Tabelle B.1 ist eine Übersicht des *SUS*-Fragebogens zu finden.

5.3 Methodik

Für diese Studie wurde ein „synchronous remote usability test“ (sRUT, synchroner Distanz-Usability-Test) durchgeführt. Das bedeutet, dass die Teilnehmer sich nicht, wie üblich in einem Labor gegenüber sitzen, sondern mithilfe eines Konferenzsystems über das Internet miteinander kommunizieren.

5.3.1 Technische Umsetzung

Die folgenden Programme wurden im Rahmen der Studie verwendet und stellen somit die technischen Voraussetzungen dar, die erfüllt sein mussten, um an der Studie teilnehmen zu können:

TeamViewer bzw. AnyDesk

Bei TeamViewer¹ und AnyDesk² handelt es sich um Programme für den Fernzugriff auf Computer. Diese Programme wurden genutzt, um den Teilnehmern Zugriff auf den Rechner zu ermöglichen, auf welchem die Studie durchgeführt worden ist.

BigBlueButton

Bei BigBlueButton³ handelt es sich um ein Konferenzsystem. In dieser Studie wurde es genutzt, um die Kommunikation zwischen Studienteilnehmer und -leiter zu ermöglichen.

OBS

OBS⁴ ist eine Software, die Videoaufnahmen des Bildschirms ermöglicht. Diese wurde genutzt, um die Teilnehmer bei der Durchführung der Studie aufnehmen zu können.

Internetverbindung

Um an der Studie teilnehmen zu können, war eine schnelle und stabile Internetverbindung die Voraussetzung.

5.3.2 Versuchsaufbau

Zu Beginn wurde die Studie mit einer Versuchsgruppe geplant. Es sollten Aufgaben mithilfe des PPC und andere ohne diesen durchgeführt werden. Mit dem Ziel eine Eingewöhnung sowie Lerneffekte, welche die Studienergebnisse beeinflussen könnten, zu verhindern, wurde ein *Between-Groups*-Studiendesign gewählt. Hierfür wurden die Teilnehmer in zwei Gruppen aufteilt. Die Experimentalgruppe, welche den PPC nutzte, um die Aufgaben zu bewältigen und die Kontrollgruppe, welche die Aufgaben ohne Unterstützung lösen musste. Auf diese Weise konnten dieselben Aufgaben in beiden Gruppen gestellt werden, ohne dass vorherige Interaktionen mit einer DSE das Ergebnis verfälschen könnten. Die Probanden wurden den Gruppen

¹<https://www.teamviewer.com/de/>

²<https://anydesk.com/de>

³<https://bigbluebutton.org/>

⁴<https://obsproject.com/de>

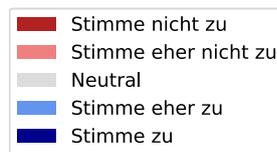


Abbildung 5.2: Legende der geschlossenen Fragen

zufällig zugeordnet. Die an die Probanden gestellten Fragen unterschieden sich leicht in beiden Gruppen.

Es wurde die Zeit für jede Aufgabe gemessen, die den Probanden gestellt wurde.

Während der Studiendurchführung wurde die Think-Aloud-Methode eingesetzt [42]. Diese verlangt von den Teilnehmern, dass sie jeden Gedanken, den sie während der Studie haben, laut aussprechen sollen.

Beide Studien wurden in vier Phasen aufgeteilt. Zu Beginn der ersten Phase wurden den Probanden einige allgemeine Fragen zu ihrer Person gestellt. Anschließend wurden die Teilnehmer über ihr Surfverhalten und ihre Gedanken zu Datenschutz befragt.

Die zweite Phase begann mit der Aufgabe, die DSE von Mozilla⁵ zu suchen. Nachdem die Probanden diese Aufgabe erfüllt hatten, wurden sie gebeten, nacheinander zwei unterschiedliche Datenpunkte in der DSE zu suchen. Im Anschluss hieran sollte ein Fragebogen ausgefüllt werden. Die dritte Phase der Studie verlief analog, nur dass die DSE von Netflix⁶ genutzt wurde. Zum Abschluss der Studie wurden die Teilnehmer gebeten, einen letzten Fragebogen auszufüllen. Ein detailliertes Manuskript der Studie ist im Anhang B zu finden. Die Dauer pro Teilnehmer belief sich auf ungefähr 20 bis 40 Minuten, je nachdem in welcher Gruppe sich der Proband befand.

5.4 Resultate

Die Studie wurde mit insgesamt **65 Teilnehmern** durchgeführt. Von diesen befanden sich 33 in der Experimental- und 32 in der Kontrollgruppe. Es waren, bis auf grundlegende Englischkenntnisse, keine weiteren Kenntnisse für eine Studienteilnahme nötig. In beiden Studiengruppen wurden offene und geschlossene Fragen gestellt, wobei es für die geschlossenen Fragen eine einheitliche Skala von „keine Zustimmung“/ „Nie“ bis „volle Zustimmung“/ „Immer“ gab. In Abbildung 5.2 ist die Legende der geschlossenen Fragen abgebildet. Die offenen Fragen wurden in zwei Phasen kategorisiert. In der ersten Phase wurden die Antworten der Probanden durch den Betreuer und den Autor dieser Arbeit getrennt voneinander kategorisiert. Anschließend wurden die Ergebnisse zusammengetragen und bei Antworten, für die kein

⁵<https://www.mozilla.org/en-US/privacy/>

⁶<https://help.netflix.com/legal/privacy/>

Konsens über die zugeordnete Kategorie bestand, wurde diskutiert, bis sich dieser fand. Dieser Prozess wurde in der zweiten Phase wiederholt, um eine genaue Zuordnung der Antworten in Kategorien zu ermöglichen. In der ersten Phase wurde eine Übereinstimmung von 123/134 Antworten erreicht, in der zweiten Phase wurden 131 Antworten in die gleiche Kategorie eingeordnet.

Die Probanden in der Experimentalgruppe waren durchschnittlich 24 Jahre alt, während der Altersdurchschnitt in der Kontrollgruppe 22 Jahre betrug. 52 (80%) der Teilnehmer gaben an, Studierende zu sein und 64 (98.46%) surfen mindestens einmal täglich im Internet. 49 (75.38%) stimmten der Aussage „Mir ist Datenschutz wichtig“ mindestens eher zu. Dies steht im Kontrast zu den Angaben der Teilnehmer, wie oft sie darauf achten, ob eine Internetseite eine DSE besitzt. Hier gaben 47 (72.31%) der Probanden an, dies selten bis nie zu tun. Weiterhin führten nur 6 (9.23%) Teilnehmer an, die DSE einer Internetseite, die sie gerade besuchen oft zu lesen. Drei (4.62%) Probanden gaben an, dies immer zu tun und 9 (13.84%) taten dies gelegentlich. Zusätzlich gaben 50 (76.92%) Personen an, dass sie bei unterschiedlichen Diensten auch unterschiedlichen Wert auf ihren Datenschutz legen. Als Grund wurde häufig die Art und Menge der Daten angeführt, die der jeweilige Dienst über sie sammelt. Bei Internetseiten, die viele oder persönliche Daten über die Probanden sammeln würden, gaben sie an, auch mehr Wert auf ihren Datenschutz zu legen. Die Probanden der Kontrollgruppe wurden abschließend gefragt, ob sie das Gefühl haben, umfassend über ihren Datenschutz aufgeklärt worden zu sein. Dieser Aussage stimmten jeweils 16 (50%) Teilnehmer zu bzw. nicht zu.

5.4.1 Resultate für RQ1

In diesem Kapitel werden die, für die RQ1 relevanten, Resultate dargestellt.

Statistische Tests

Statistische Tests stellen eine Möglichkeit dar, die Validität einer Hypothese zu überprüfen. Diese werden im Folgenden dafür genutzt, um zu prüfen, ob die DSE unter der Verwendung der entwickelten BE schneller gefunden werden kann. Dementsprechend ist im Kontext des Experiments folgende Nullhypothese zu überprüfen:

H_0 : Es existiert kein signifikanter Unterschied zwischen dem Auffinden einer DSE mit oder ohne der entwickelten BE.

Für diesen Zweck müssen die folgenden Datensätze betrachtet werden. Die Zahlen 1 bzw. 2 im Index stehen für die Nummer der Aufgabe. Die Wörter „mit“ und „ohne“ sagen aus, dass die Zeit jeweils mit bzw. ohne Extension gemessen worden ist.

Datensatz	Teststatistik W	p-Wert
$t_{1,RQ1,mit}$	0.969	0.458
$t_{1,RQ1,ohne}$	0.959	0.264
$t_{2,RQ1,mit}$	0.929	0.032
$t_{2,RQ1,ohne}$	0.942	0.827

Tabelle 5.1: Die Teststatistiken und p-Werte für die RQ1-Messungen

- $t_{1,RQ1,mit}$: Benötigte Zeiten zum Finden der Mozilla-DSE mit Extension.
- $t_{1,RQ1,ohne}$: Benötigte Zeiten zum Finden der Mozilla-DSE ohne Extension.
- $t_{2,RQ1,mit}$: Benötigte Zeiten zum Finden der Netflix-DSE mit Extension.
- $t_{2,RQ1,ohne}$: Benötigte Zeiten zum Finden der Netflix-DSE ohne Extension.

Die Normalverteilung der jeweiligen Datensätze wurde mithilfe des Shapiro-Wilk-Tests [45] geprüft. Da Zeitdaten positiv verzerrt sind, wurden diese zunächst in eine logarithmische Darstellung umgewandelt [46]. Zusätzlich bietet diese Transformation einen Schutz gegenüber statistischen Ausreißern. Bis auf Gruppe $t_{2,RQ1,mit}$ konnte mittels des Shapiro-Wilk-Tests die Normalverteilung bestätigt werden. Die Teststatistiken und die dazugehörigen p-Werte können der Tabelle 5.1 entnommen werden. Gemäß Mordkoff [47] wird dennoch von einer Normalverteilung ausgegangen, da die Stichprobengröße 33 Teilnehmer beträgt und somit der zentrale Grenzwertsatz „greift“.

Sofern nicht anders angegeben, stehen die Abkürzungen M für das arithmetische Mittel und SD für die Standardabweichung der Stichproben. Mittels eines Zweistichproben-Gauß-Tests (Z-Test) [48] wurde ein signifikanter Unterschied der mittleren DSE-Auffindungszeit auf mozilla.org zwischen Nutzern ($M \approx 25.88$, $SD \approx 29.77$) und Nicht-Nutzern ($M \approx 25.88$, $SD \approx 29.77$) der entwickelten BE bei einem 5%-Signifikanzniveau gefunden; $z \approx -7.59$, $p < 0.01$. Falls nicht anders angegeben, wird im Folgenden weiter von diesem Signifikanzniveau ausgegangen. Ebenso konnte dies auf netflix.com für die Nutzer- ($M \approx 27.93$, $SD \approx 33.76$) und die Nicht-Nutzer-Gruppe ($M \approx 31.81$, $SD \approx 16.64$) gezeigt werden; $z \approx -2.13$, $p \approx 0.03$.

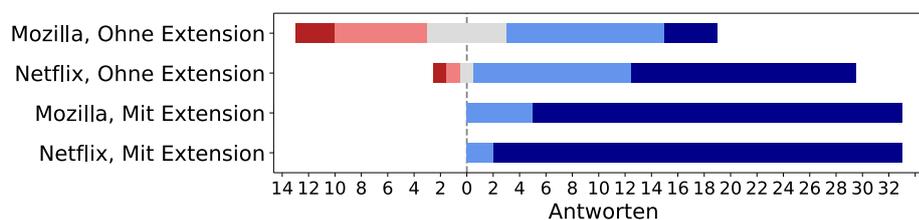


Abbildung 5.3: Reaktionen auf die Aussage „Die Datenschutzerklärung war einfach zu finden.“

Qualitative Analyse

Abbildung 5.3 zeigt die Verteilung der Reaktionen auf die Aussage „Die Datenschutzerklärung war (mithilfe der Browser Extension) einfach zu finden“ in Form eines Balkendiagramms. Beginnend von oben zeigt der erste Balken die Reaktion der Nutzer auf der DSE von Mozilla, die die BE nicht nutzten. Der zweite beschreibt die Reaktion der Nutzer ohne BE auf der DSE von Netflix. Die unteren zwei Balken zeigen die Reaktionen der Nutzer, welche die BE auf den DSEs von Mozilla respektive Netflix nutzten.

Teilnehmer mit Unterstützung der entwickelten Extension stimmten ohne Ausnahme in beiden Fällen der Aussage mindestens eher zu (100%).

10 (31.25%) Probanden, die die BE nicht nutzten, gaben im Falle der DSE von Mozilla an, die Aussage mindestens eher abzulehnen. Hier gaben 6 (18.75%) Teilnehmer an, der Aussage neutral gegenüberzustehen und 16 (50%) stimmten ihr mindestens eher zu.

Die Personen, die die DSE von Netflix ohne Extension suchen sollten, reagierten folgendermaßen: Jeweils eine Person (6.25%) stimmte der Aussage nicht bzw. eher nicht zu, eine weitere Person (3.13%) gab eine neutrale Antwort. Die restlichen 29 (90.62%) Probanden stimmte mindestens eher zu.

Die Antworten der Teilnehmer fielen bei Mozilla ohne Extension möglicherweise negativer als in den anderen Fällen aus, da Mozilla für jedes ihrer Produkte eine DSE anbot, jedoch nach einer spezifischen gefragt wurde. Netflix hingegen verlinkte nur eine DSE auf ihrer Internetpräsenz.

5.4.2 Resultate für RQ2

In diesem Kapitel werden die, für die RQ2 relevanten, Resultate dargestellt.

Statistische Tests

Die Hypothese „Die BE ermöglicht eine schnellere Suche von Informationen innerhalb einer DSE im Vergleich zu herkömmlichen Methoden“ soll analog zu Unterabschnitt 5.4.1 überprüft werden. Zu diesem Zweck ist im Kontext des Experiments folgende Nullhypothese H_0 zu überprüfen:

H_0 : Es existiert kein signifikanter Unterschied zwischen dem Auffinden von Informationen innerhalb einer DSE mit oder ohne der entwickelten BE.

Demnach müssen folgende Datensätze betrachtet werden. Hier stehen, analog zu den Datensätzen aus Unterabschnitt 5.4.1, die Zahlen im Index für die Nummer der Aufgabe und die Wörter „mit“ und „ohne“ für die (Nicht-)Nutzung der BE.

- $t_{3,RQ2,mit}$: Benötige Zeiten zum Auffinden der zur ersten Aufgabe spezifischen Information auf mozilla.org mit Extension
- $t_{3,RQ2,ohne}$: Benötige Zeiten zum Auffinden der zur ersten Aufgabe spezifischen Information auf mozilla.org ohne Extension
- $t_{4,RQ2,mit}$: Benötige Zeiten zum Auffinden der zur zweiten Aufgabe spezifischen Information auf mozilla.org mit Extension
- $t_{4,RQ2,ohne}$: Benötige Zeiten zum Auffinden der zur zweiten Aufgabe spezifischen Information auf mozilla.org ohne Extension
- $t_{5,RQ2,mit}$: Benötige Zeiten zum Auffinden der zur ersten Aufgabe spezifischen Information auf netflix.com mit Extension
- $t_{5,RQ2,ohne}$: Benötige Zeiten zum Auffinden der zur ersten Aufgabe spezifischen Information auf netflix.com ohne Extension
- $t_{6,RQ2,mit}$: Benötige Zeiten zum Auffinden der zur zweiten Aufgabe spezifischen Information auf netflix.com mit Extension
- $t_{6,RQ2,ohne}$: Benötige Zeiten zum Auffinden der zur zweiten Aufgabe spezifischen Information auf netflix.com ohne Extension

Für die genannten Datensätze wurde analog zu Unterabschnitt 5.4.1 die Normalverteilung mithilfe des Shapiro-Wilk-Tests bestätigt. Die Teststatistiken mit den zugehörigen p-Werten sind in der Tabelle 5.2 zu finden. Erneut wurden Zeitdaten in eine logarithmische Darstellung transformiert.

Die durchgeführten Zweistichproben-Gauß-Tests liefern folgende Ergebnisse: Es wurde ein signifikanter Unterschied der mittleren Ausführungszeit zwischen $t_{3,RQ2,mit}$ ($M \approx 42.61$, $SD \approx 34.04$) und $t_{3,RQ2,ohne}$ ($M \approx 59.53$, $SD \approx 42.02$) gefunden; $z \approx -2.10$, $p \approx 0.04$. Ebenfalls wurde ein signifikanter Unterschied der mittleren Ausführungszeit zwischen $t_{4,RQ2,mit}$ ($M \approx 26.45$, $SD \approx 32.02$) und $t_{4,RQ2,ohne}$ ($M \approx 52.26$, $SD \approx 42.13$) gefunden; $z \approx -3.86$, $p < 0.01$. Es konnte kein signifikanter Unterschied der mittleren Ausführungszeit zwischen $t_{5,RQ2,mit}$ ($M \approx 138.27$, $SD \approx 81.69$) und $t_{5,RQ2,ohne}$ ($M \approx 145.38$, $SD \approx 85.55$) festgestellt werden; $z \approx -0.26$, $p \approx 0.80$. Außerdem konnte kein signifikanter Unterschied der mittleren Ausführungszeit zwischen $t_{6,RQ2,mit}$ ($M \approx 54.09$, $SD \approx 60.36$) und $t_{6,RQ2,ohne}$ ($M \approx 67.69$, $SD \approx 51.99$) festgestellt werden; $z \approx -1.19$, $p \approx 0.24$.

Datensatz	Teststatistik W	p-Wert
$t_{3,RQ2,mit}$	0.966	0.378
$t_{3,RQ2,ohne}$	0.960	0.271
$t_{4,RQ2,mit}$	0.975	0.628
$t_{4,RQ2,ohne}$	0.982	0.866
$t_{5,RQ2,mit}$	0.959	0.247
$t_{5,RQ2,ohne}$	0.982	0.862
$t_{6,RQ2,mit}$	0.948	0.113
$t_{6,RQ2,ohne}$	0.970	0.492

Tabelle 5.2: Die Teststatistiken und p-Werte für die RQ2-Messungen

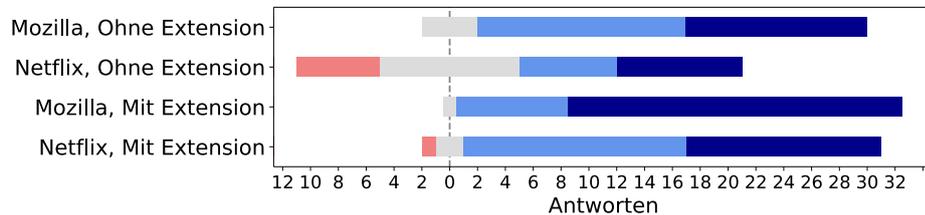


Abbildung 5.4: Reaktionen auf die Aussage „Die Informationen aus der Datenschutzerklärung waren einfach zu finden.“

Qualitative Analyse

In Abbildung 5.4 sind die Reaktionen der Teilnehmer auf die Aussage „Die Informationen aus der Datenschutzerklärung waren (mithilfe der Browser Extension) einfach zu finden“ zu sehen. Der Aufbau des Balkendiagramms ist analog zu dem aus Abbildung 5.3. Die ersten beiden Balken stellen die Reaktion der Probanden ohne Extension für beide Internetseiten auf die Aussage dar, die zweiten für die Nutzer der Extension. Es ist ersichtlich, dass auf der DSE von Mozilla weder die Teilnehmer mit noch ohne Extension eine negative Reaktion zeigten. Lediglich die Anzahl der neutralen (Mit: 1 (3.03%), Ohne: 4 (12.5%)), eher zustimmenden (Mit: 8 (24.24%), Ohne: 15 (46.87%)) und zustimmenden (Mit: 24 (72.73%), Ohne: 13 (40.63%)) Aussagen unterscheidet sich.

Auf der DSE von Netflix ergaben sich folgende Reaktionen auf die Aussage: 6 (18.75%) Probanden ohne Extension stimmten eher nicht zu, 10 (31.25%) gaben an der Aussage neutral gegenüberzustehen und 16 (50%) stimmten mindestens eher zu. Unter den Nutzern der Extension stimmte eine Person (3.03%) eher nicht zu, zwei (6.06%) blieben neutral und 30 (90.91%) gaben an, der Aussage mindestens eher zuzustimmen.

Die DSE von Netflix enthielt mehr Informationen und war textuell deutlich länger als die DSE von Mozilla.

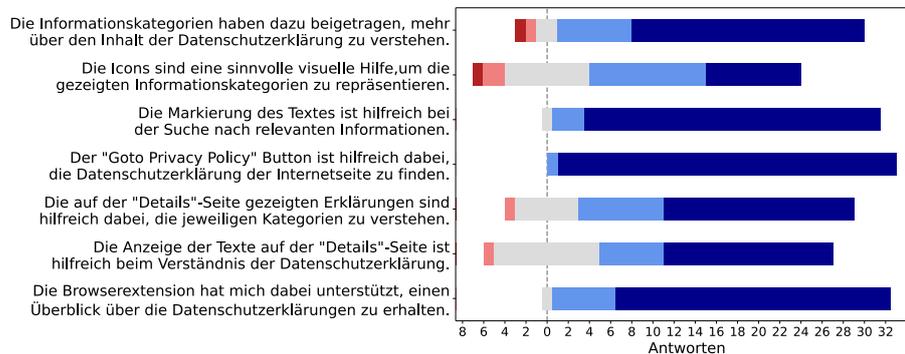


Abbildung 5.5: Angaben zum Verständnis einer Datenschutzerklärung

5.4.3 Resultate für RQ3

Mit dem Ziel, die Frage: „Unterstützt die Browser Extension den Nutzer bei dem Verständnis einer Datenschutzerklärung?“ beantworten zu können, sind einige Punkte zu berücksichtigen. Zum einen spielen die, in Unterabschnitt 5.4.1 und Unterabschnitt 5.4.2 analysierten Zeiten zum Auffinden einer DSE und zum Suchen von Informationen in dieser, eine Rolle. Zusätzlich sind noch weitere Faktoren von Relevanz, welche in diesem Abschnitt beleuchtet werden sollen. Einige der hierfür relevanten Fragen sind in Abbildung 5.5 zu sehen. Die für diesen Zweck genutzten Balkendiagramme sind analog zu den in Unterabschnitt 5.4.2 genutzten zu verstehen.

Die Nutzer der Extension wurden, jeweils nach ihren Aufgaben innerhalb einer DSE, über einige Funktionen des Systems befragt. So wurden sie gefragt, ob die dargestellten Informationskategorien innerhalb der BE-Benutzeroberfläche bei dem Verständnis einer DSE helfen. Jeweils eine Person (6.06%) stimmte dieser Frage nicht bzw. eher nicht zu. Zwei blieben neutral (6.06%) und 29 (87.88%) stimmten mindestens eher zu.

Anschließend wurden die Probanden der Aussage gegenübergestellt, ob die in Unterabschnitt 3.3.1 dargestellten Piktogramme bei dem Verständnis der Informationskategorien helfen würden. Drei Personen (9.09%) stimmten der Aussage nicht bzw. eher nicht zu, 8 (24.24%) blieben neutral und 22 (66.67%) stimmten mindestens eher zu.

Zusätzlich wurde die Frage gestellt, ob die Markierung des Textes innerhalb einer DSE hilfreich bei der Suche nach relevanten Informationen sei. Hier blieb eine Person (3.03%) neutral, die restlichen 32 Teilnehmer (96.97%) stimmten mindestens eher zu.

Dann wurden die Probanden über ihre Meinung befragt, ob der „Goto Privacy Policy“-Button, welcher die Nutzer direkt auf eine DSE einer Internetseite führt, hilfreich sei, um diese zu finden. Hier stimmten 32 Personen (96.97%) zu, eine (3.03%) stimmte eher zu.

Ferner wurden den Teilnehmern zwei Fragen über die Details-Seite des PPC

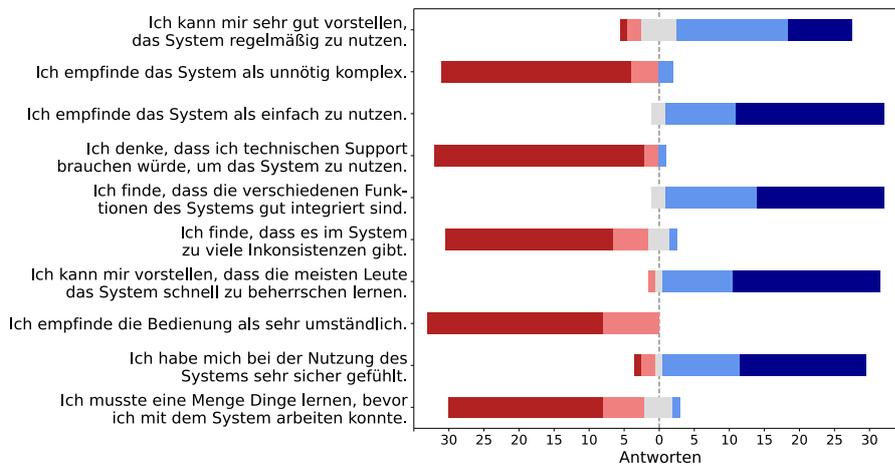


Abbildung 5.6: Verteilung der SUS-Bewertungen

gestellt. Die erste sollte in Erfahrung bringen, wie gut die dortigen Erklärungen die angezeigten Informationskategorien erklären. Eine Person (3.03%) stimmte der Aussage eher nicht zu, 6 (18.18%) blieben neutral und 26 (78.79%) stimmten mindestens eher zu. Die zweite Frage wurde gestellt, um in Erfahrung zu bringen, ob die Anzeige der, zu einer Kategorie passenden, Abschnitte aus einer DSE hilfreich bei dem Verständnis dieser sind. Auch hier stimmte eine Person (3.03%) eher nicht zu, aber 10 (30.30%) blieben neutral und 22 (66.67%) stimmten mindestens eher zu.

Abschließend wurden die Probanden mit der Aussage konfrontiert, ob die BE sie dabei unterstützt hat, einen Überblick über die DSEs zu erhalten. Hier blieb eine Person neutral (3.03%), während 32 (96.97%) ihr mindestens eher zustimmten. Als Grund führten die Probanden hier oft den, durch die Extension ermöglichten, schnelleren Zugriff auf die Informationen innerhalb einer DSE an.

5.4.4 Resultate für RQ4

In Abbildung 5.6 ist die Verteilung der einzelnen SUS-Bewertungen zu sehen. Jede zweite Frage des Fragebogens ist semantisch negativ formuliert. Das bedeutet, dass „Stimme nicht zu“ die beste Auswahloption darstellt. Aus diesem Grund kam es während der Durchführung der Studie gelegentlich dazu, dass Teilnehmer versehentlich die Antwort wählten, die dem Gegensatz ihrer eigentlichen Intention entsprach. Diese Fehler konnten jedoch den Probanden mitgeteilt und anschließend von ihnen korrigiert werden.

Der ermittelte SUS-Score beträgt im Durchschnitt 87,95 ($SD \approx 9.76$) und liegt somit um 19.95 Punkte eindeutig über dem der durchschnittlichen Web-Applikation von 68 [49].

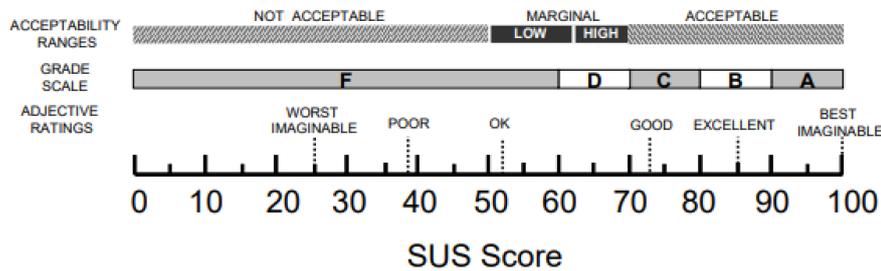


Abbildung 5.7: Abbildung der SUS-Bewertung auf weitere Skalen [49]

Die Normalverteilung der Stichprobe wurde mit einem Shapiro-Wilk-Test ($W \approx 0.88$, $p < 0.0016$) überprüft. Dieser Test konnte die Normalverteilung der Stichprobe nicht bestätigen, da die Alternativhypothese (Sie entspricht **nicht** der Normalverteilung) mit $p < 0.05$ bei einem 5%-Signifikanzniveau angenommen werden muss. Analog zu Unterabschnitt 5.4.1 wird erneut von einer Normalverteilung der Stichprobe aufgrund des zentralen Grenzwertsatzes und der Stichprobengröße von 33 Teilnehmern ausgegangen. Ein Einstichproben-Gauß-Test zwischen der Stichprobe und der durchschnittlichen SUS-Bewertung einer Web-Applikation (68) [49] zeigt einen signifikanten Unterschied zugunsten der entwickelten BE; $z \approx 11.57$, $p < 0.01$. Nachfolgend konnte zugunsten der BE zwischen der Stichprobe und 84 ebenfalls ein signifikanter Unterschied mittels eines Einstichproben-Gauß-Tests nachgewiesen werden; $z \approx 2.29$, $p \approx 0.02$.

Weiter wurde das 95%-Konfidenzintervall als $[84.49, 91.41]$ ermittelt. Aus diesem Grund kann davon ausgegangen werden, dass sich der echte SUS-Score zu 95% innerhalb dieses Intervalls bewegt. Somit handelt es sich gemäß Abbildung 5.7 um ein System mit mindestens überdurchschnittlicher bzw. beinahe exzellenter Bedienbarkeit [49]. Allerdings zeigte die Resonanz der Teilnehmer während der Studie doch einige Schwächen der Implementierung. So merkten verschiedene Probanden beispielsweise an, dass sie sich Feedback wünschen würden, wenn eine Textmarkierung auf einer Internetseite angefordert wurde, da diese Aufgabe einige Sekunden in Anspruch nehmen kann.

Kapitel 6

Diskussion

An dieser Stelle sollen die Ergebnisse der Evaluation analysiert und interpretiert werden. Zusätzlich wird über die Limitierungen der Studie gesprochen und abschließend ein Fazit gezogen.

6.1 Interpretation der Ergebnisse

RQ1 Viele Anwender legen in ihrem täglichen Leben aus verschiedenen Gründen wenig Wert auf ihren Datenschutz. Die Gründe hierfür sind vielfältig. 47 der 65 Teilnehmer gaben an, selten bis nie darauf zu achten, ob eine Internetseite eine DSE besitzt und nur sechs der Probanden äußerten, dass sie diese oft lesen würden, bevor sie eine Internetseite nutzen. Vier Personen waren der Meinung, dass es ausweglos sei, sich gegen „Datenkraken“ zu wehren, da diese sowieso bereits alle wichtigen Informationen über die Person besäßen. 12 Teilnehmer sagten aus, dass sie ihr Datenschutz nicht interessieren würde. Anderen war die Bedeutung des Datenschutzes zwar bewusst, er ist ihnen aber den Aufwand, den er macht, nicht wert.

Mit dem Ziel mehr Nutzer für ihren Datenschutz zu interessieren, ist es wichtig, die Belastung, die dieser verursacht, zu minimieren. Viele Anwender sind möglicherweise schon nicht bereit, die Zeit zu investieren, um die DSE auf einer Internetseite zu suchen. In der durchgeführten Studie wurde gezeigt, dass Nutzer mithilfe der entwickelten BE signifikant weniger Zeit benötigten, um die genannte Aufgabe zu erledigen. Auf diese Weise wurde es möglich gemacht, die Informationen der DSE für den Nutzer schneller erreichbar zu machen, was auch durch die Aussagen der Teilnehmer gestützt wird. Somit lässt sich, mit den im Rahmen dieses Experiments festgehaltenen Daten, sagen, dass die **RQ1** mit „Ja“ beantwortet werden kann. Die entwickelte BE unterstützt den Nutzer dabei, die DSE einer Internetseite zu finden.

RQ2 Für die Beantwortung dieser Frage wurden in der Studie neben der subjektiven Meinung der Teilnehmer auch die Zeit gemessen, die sie für die Suche nach Informationen innerhalb einer DSE brauchten. Die Nutzer gaben an, dass die BE sie dabei unterstützt hat, die gesuchten Informationen schneller zu finden. Auf der Datenschutzerklärung von Mozilla wurde zudem ein signifikanter Unterschied zwischen den Nutzern und Nicht-Nutzern der BE gemessen. Dies war bei der DSE von Netflix nicht der Fall. Dort waren die Teilnehmer im Durchschnitt auch schneller, der Zeitgewinn war jedoch nicht signifikant. Dies liegt möglicherweise an der Länge und dem Inhalt der DSE. Je mehr Daten ein Konzern über seine Nutzer sammelt, desto mehr Ergebnisse zeigt auch der PPC an. Dies führt dazu, dass auch hier Nutzer mehr Zeit benötigen, um den Inhalt der DSE zu verstehen. Insgesamt lässt sich auch hier, durch die im Experiment gesammelten Daten, festhalten, dass die BE den Nutzer dabei unterstützt, Informationen innerhalb einer DSE zu finden.

RQ3 Soll die Frage beantwortet werden, ob die entwickelte BE Nutzer bei dem Verständnis einer DSE unterstützt, so müssen hier verschiedene Metriken in Betracht gezogen werden. Zum einen spielt die Zeit eine Rolle, die die Nutzer benötigen, um eine DSE zu verstehen. Zum anderen ist die subjektive Wahrnehmung der Probanden von essentieller Wichtigkeit. Denn wenn Nutzer Probleme haben, mit der Anwendung umzugehen, Funktionen nicht anwenden können oder Ergebnisse nicht interpretieren können, so ist ein entwickeltes Programm nicht dienlich, um den vorgesehenen Zweck zu erfüllen. Durch RQ1 und RQ2 wurde gezeigt, dass die benötigte Zeit, um diese Aufgaben zu erfüllen, minimiert worden ist. Zusätzlich gaben viele Nutzer an, dass die genutzten Piktogramme sie bei dem Verständnis des Textes unterstützen würden und auch die Markierung von wichtigen Passagen innerhalb einer DSE wurde als äußerst hilfreich angesehen. Außerdem wurde gezeigt, dass die implementierte „Details“-Seite dem Verständnis des Inhalts einer DSE dienlich sei. Durch die hier gesammelten Daten lässt sich somit auch diese RQ positiv beantworten.

RQ4 Die Bedienbarkeit einer Applikation spielt im täglichen Arbeitsablauf eines Anwenders eine wichtige Rolle. Bietet ein Programm eine weniger gute Bedienbarkeit, so nutzen Anwender dieses möglicherweise nicht oder wechseln auf ein anderes Produkt. Mit dem Ziel die Bedienbarkeit der entwickelten BE zu evaluieren, wurde im Rahmen der durchgeführten Nutzerstudie der sogenannte *SUS-Score* berechnet. Dieser ordnet den Aussagen von Nutzern eine Kennzahl zu, die die Bedienbarkeit von Anwendungen vergleichbar machen soll. Der PPC erreichte einen *durchschnittlichen SUS-Score von 87.95* und liegt damit weit über dem einer durchschnittlichen Web-Applikation von 68. Damit wurde gezeigt, dass die entwickelte BE eine überdurchschnittliche Bedienbarkeit aufweist.

6.2 Limitierung der Ergebnisse

In der Experimentalgruppe befanden sich 33 Probanden, während in der Kontrollgruppe 32 zu finden waren. Zusätzlich gaben 80% der Teilnehmer an, Studierende zu sein. Diese Eigenschaften repräsentieren nur einen kleinen Teil der Bevölkerung. Daher sind die Ergebnisse der Studie möglicherweise nicht repräsentativ.

Die an die Probanden gestellten Aufgaben und Fragen stammen vom Autor. Obwohl eine Pilotierung mit dem Betreuer dieser Arbeit vor Beginn der Studie durchgeführt wurde, kann nicht garantiert werden, dass alle Fragen objektiv gestellt worden sind.

Hinzu kommt, dass die Teilnehmer nicht alle dieselben Vorkenntnisse hatten. Einige hatten Probleme mit der englischen Sprache, während andere sich noch nie mit Datenschutz auseinandergesetzt haben.

Eine weitere Limitierung der durchgeführten Studie ist der begrenzte Rahmen, in dem diese durchgeführt wurde. Nutzer besuchen täglich viele verschiedene Internetseiten, im Rahmen des Experiments konnten jedoch nur zwei betrachtet werden.

Aufgrund der COVID-19 Pandemie wurden die Experimente vollständig online durchgeführt, dies stellt jedoch keine Limitierung der Ergebnisse dar, wie Andreasen et al. [50] in ihrer Arbeit zeigten. Ein Problem, welches durch die getätigten Sprachaufnahmen der Probanden ersichtlich wurde, ist, dass diese die Fragebögen möglicherweise nicht nach ihrem Empfinden beantworten haben. So gaben einige Teilnehmer im Gespräch an, dass sie Probleme hatten, die DSE zu verstehen. Sie äußerten dies jedoch nicht in der anschließenden Beantwortung des Fragebogens, sondern gaben dort durchweg positive Antworten.

6.3 Fazit

Aufgrund des beschränkten Rahmens, in welchem die BE getestet werden konnte, bietet es sich an, diese durch zukünftige Forschung weiter zu evaluieren. Es wäre von Vorteil, die BE mit einer heterogeneren Nutzergruppe, welche nicht zu 80% aus Studenten besteht, zu testen, sie einer ebenso vielschichtigeren Kontrollgruppe gegenüberzustellen und weitere Internetseiten für die Untersuchung durch Probanden auszuwählen.

Jedoch zeigen die Ergebnisse, der im Rahmen dieser Studie durchgeführten Experimente, bereits ein vielversprechendes Potential der entwickelten BE. Die Teilnehmer fanden DSEs und Informationen in diesen schneller und gaben an, dass viele Funktionen sie bei dem Verständnis einer DSE unterstützt haben. Hier wurden beispielsweise die Markierung von relevantem Text innerhalb einer DSE oder die Anzeige der, in DSEs enthaltenen, Kategorien in der Benutzeroberfläche genannt. Auf diese Weise ist es ermöglicht

worden, die Probanden auf ihren Datenschutz aufmerksam zu machen. Dies stellt einen wichtigen Schritt in der Stärkung und weiteren Entwicklung des Datenschutzbewusstseins der Nutzer dar. Das entwickelte Programm integriert sich nahtlos in den täglichen Arbeitsablauf der Anwender und wurde von diesen sehr positiv aufgenommen. Viele der Teilnehmer wünschten sich eine Veröffentlichung der Extension, damit sie diese in ihren Browser einbinden können.

Kapitel 7

Verwandte Arbeiten

In diesem Kapitel werden wissenschaftliche Publikationen vorgestellt, die ein sehr ähnliches Ziel wie diese Arbeit verfolgten. Zusätzlich wird zu jeder Publikation darauf eingegangen, inwiefern sie sich von dieser Arbeit unterscheidet.

7.1 Polisis und PriBot

Harkous et al. [51] entwickelten in ihrer Arbeit ein System, um die DSE einer Internetseite automatisiert zu analysieren und die Ergebnisse dem Benutzer darzustellen. Während sie für die Analyse einer DSE *Deep-Learning*-Verfahren nutzten, stellten sie die Ergebnisse dieser Analyse auf zwei Wegen dar. Zunächst wurde ein Chatbot namens *PriBot* entwickelt [52], an welchen die Anwender konkrete Fragen über eine DSE stellen können. Diese werden daraufhin von dem Bot beantwortet. Weiter können aufbereitete, strukturierte DSE über eine Internetseite angefragt werden.

Die Publikation von Harkous et al. unterscheidet sich folgenden Punkten von dieser Arbeit. Zum einen wurden *neuronale Netzwerke* für die Klassifikation der Texte verwendet und zum anderen wurde das System nicht als BE implementiert. Möchte ein Nutzer Informationen über eine DSE erhalten, so muss er einen externen Service ansteuern. Diesen Schritt sind womöglich viele Anwender nicht bereit zu gehen, da dies zusätzlichen Arbeitsaufwand bedeuten würde. Im Umkehrschluss führt das möglicherweise zu einer Vernachlässigung ihres Datenschutzes.

Stattdessen wird in dieser Arbeit bewusst ein anderer Ansatz verfolgt. Es wird eine BE entwickelt, die sich direkt in die Arbeitsumgebung des Nutzers (den Browser wie bspw. Firefox) integriert, sodass kein weiterer Schritt notwendig ist. Somit wird dem Anwender eine bequem nutzbare softwareseitige Unterstützung für das Verständnis einer DSE geboten.

7.2 P3P

Cranor stellte in ihrer Arbeit [53] die „Platform for Privacy Preferences (P3P)“ vor, ein Projekt des World Wide Web Consortium (W3C) mit dem Ziel, das Format von DSEs auf Anbieterseite zu vereinheitlichen. Es wurde ein System vorgeschlagen, welches auf dem Extensible Markup Language (XML) Dateiformat basiert. Dieses Format bietet den Vorteil, dass es durch Computer verarbeitbar ist. Auf diese Weise soll ermöglicht werden, die wichtigsten Punkte einer DSE dem Benutzer ohne weitere Systeme direkt im Browser darstellen zu können.

Die P3P unterscheidet sich drastisch von dieser Arbeit. Das System sieht vor, dass Anbieter einer Internetseite eine XML-Datei auf einem bestimmten Pfad hinterlegen. Diese Datei dient dazu, dem Nutzer Informationen bereitzustellen. Aus diesem Grund setzt der PPC auf Algorithmen aus dem Bereich des maschinellen Lernens und dem NLP, um auch DSE interpretieren zu können, die in natürlicher Sprache verfasst worden sind.

7.3 Privacy Bird

Cranor et al. [54] beschreiben in ihrer Arbeit den *Privacy Bird*. Dieser wurde ebenfalls durch eine Initiative des W3C entwickelt. Hierbei handelt es sich um ein System, welches die in Abschnitt 7.2 beschriebenen XML-Dateien auslesen und verarbeiten kann. Besitzt eine Internetseite eine DSE in P3P-Form, so benachrichtigt der Privacy Bird den Anwender und stellt die enthaltenen Informationen in einer übersichtlichen Form dar.

Das Ziel der Arbeit von Cranor et al. und das dieser sind sich sehr ähnlich. Der Privacy Bird besitzt infolge der zugrunde liegenden Funktionsweise jedoch die gleichen Limitierungen wie in Abschnitt 7.2 beschrieben. Er kann die enthaltenen Informationen nur darstellen, wenn eine DSE in XML-Format vorliegt, während der PPC diese Aufgabe auch ohne bewältigen kann.

7.4 Zusammenfassung mittels Ensemble Learning

Tomuro et al. [55] entwickelten ein System, welches mithilfe von *Ensemble Learning* DSEs zusammenfassen kann. Bei dieser Technik handelt es sich um eine Kombination von Verfahren aus dem Bereich des maschinellen Lernens. Zusätzlich wurden manuell wichtige Schlagworte gezählt, um Sätze aus DSEs in fünf Kategorien einzuordnen. Die Ergebnisse dieses Verfahrens konnten über eine Internetseite abgerufen werden.

Die Arbeit von Tomuro et al. unterscheidet sich wieder in zwei wesentlichen Punkten von dieser. Möchte der Anwender eine DSE zusammenfassen, so muss er erst den entwickelten Service aufrufen. Außerdem nutzt der PPC nur maschinelles Lernen, ohne Schlagwortzählung, um DSEs zusammenzufassen.

Kapitel 8

Zusammenfassung und Ausblick

Dieses Kapitel dient dazu, einen abschließenden Überblick über die Inhalte und die wichtigsten Ergebnisse der Arbeit zu geben.

8.1 Zusammenfassung

Die Sammlung von personenbezogenen Daten wird in Zukunft nicht an Relevanz verlieren. Daher bleibt es ein wichtiges Ziel, das Bewusstsein von Nutzern für ihren Datenschutz zu verbessern. In dieser Arbeit wurde eine BE entwickelt, welche dazu dienen soll, diesem Ziel einen Schritt näher zu kommen. Dafür wurde, mithilfe von Angaben aus originalen DSEs, ein System konzeptioniert und implementiert, welches dem Anwender eine DSE visuell aufbereitet und zusammengefasst darstellt. Das System beinhaltet einen Mechanismus für die automatisierte Erkennung einer DSE auf einer Internetseite, Algorithmen für die Analyse dieser und Möglichkeiten, um die Ergebnisse dem Anwender übersichtlich darzustellen.

Die Erkennung einer DSE und die Darstellung der Analyseergebnisse wurden erfolgreich als BE implementiert, welche sich in fast allen gängigen Browsern nutzen lässt. Die eigentliche Analyse wurde aus Performanzgründen auf einen Server ausgelagert. Um eine korrekte Funktionsweise des Systems zu garantieren, sind alle Bestandteile von essentieller Wichtigkeit.

Auch wenn statistische Methoden vielversprechende Ansätze für die Analyse von DSEs boten, wurde sich für einen Ansatz aus dem Bereich des maschinellen Lernens entschieden, um eine Unabhängigkeit von menschlichen Adjustierungsfehlern zu gewährleisten und konkrete DSEs besser abbilden zu können. Daher wurde der in Unterabschnitt 3.2.2 erläuterte *OPP-115-Korpus* verwendet, um einen Algorithmus für die Klassifizierung von Texten aus einer DSE zu entwickeln. Dieser ordnet Abschnitte einer DSE passenden Kategorien zu. Ein hoch priorisierter Punkt während der Konzeptionierung

war die Benutzeroberfläche. Sie wurde mithilfe des React-Frameworks entwickelt und bietet einen übersichtlichen Weg, die in einer DSE enthaltenen Kategorien zu visualisieren und damit mehr über den Inhalt zu erfahren. Nutzer haben die Möglichkeit, sich zu einer Kategorie passende Texte innerhalb einer DSE farblich markieren zu lassen. Ferner können sie auf einer extra für diesen Zweck entworfenen Seite die Details der Analyseergebnisse betrachten und an dieser Stelle mehr über die analysierten Kategorien erfahren.

Die nach der Konzeptionierungs- und Implementierungsphase durchgeführte Nutzerstudie zeigte einen signifikanten Zeitgewinn bei der Suche nach einer DSE und ebenso einen Zeitgewinn bei der Suche nach Informationen innerhalb dieser. Außerdem wurde ein durchschnittlicher *SUS-Score* von 87.95 berechnet. Dieser sagt aus, dass die entwickelte Applikation eine mindestens überdurchschnittliche bis hin zu einer exzellenten Bedienbarkeit vorweist [49].

8.2 Ausblick

Zu den Problemen, die im Rahmen dieser Arbeit nicht gelöst werden konnten, gehört eine tiefere Klassifikation der Inhalte einer DSE und der entsprechenden Visualisierung innerhalb der Benutzeroberfläche. So bietet das entwickelte System dem Anwender die Möglichkeit, eine übersichtliche Kategorisierung einer DSE zu erhalten. Es arbeitet jedoch nur korrekt, wenn der gesamte Inhalt einer DSE in einem einzigen HTML-Dokument zu finden ist und das Dokument keine Informationen enthält, die erst durch weitere Klicks sichtbar werden. Möchte man diesem Problem begegnen, so können Mechanismen entwickelt werden, welche es ermöglichen, den gesamten Inhalt einer DSE, die auf mehrere Seiten verteilt ist, zu erfassen. Zusätzlich arbeitet der, für diese Arbeit trainierte, Klassifikationsalgorithmus nicht in jedem Fall korrekt. Ein weiteres Problem des verwendeten Datensatzes ist es, dass dieser nur englischsprachige DSEs kategorisiert. Infolgedessen bietet es sich an, andere Klassifikationsverfahren zu evaluieren oder einen Datensatz für das Training dieser manuell zusammenzustellen.

Außerdem ist es naheliegend, die entwickelte BE mit weiterer Forschung zu untersuchen, denn wie bereits in Abschnitt 6.2 erläutert, unterliegt die durchgeführte Studie einigen Limitierungen. In der Nutzerstudie wurden einige Punkte an der entwickelten BE kritisiert. Die zehn Kategorien (vgl. Unterabschnitt 3.2.2) helfen bereits dabei, einen guten Überblick über eine DSE zu erhalten. Sollen jedoch spezifische oder tiefere Fragen beantwortet werden, muss häufig in der DSE selber nach diesen Informationen gesucht werden. Beispielsweise spielen Cookies im Datenschutzbewusstsein der Nutzer eine wichtige Rolle, diese werden vom PPC jedoch hauptsächlich der Kategorie „First Party Collection/ Use“ zugeordnet, welche bereits viele

verschiedene Informationen enthält. Ein Ansatz, um dieser Problematik zu entgegnen, wäre es auch hier, einen vielfältigeren Datensatz zu verwenden, um auch detailreiche Fragen beantworten zu können oder es den Nutzern zu ermöglichen, mithilfe einer Suchfunktion die Ergebnisse des PPC schneller zu erkunden. Darüber hinaus wurden einige der genutzten Piktogramme kritisiert, die bei der Aufnahme der gezeigten Informationen unterstützen sollen. Hier wurde angemerkt, dass einige dieser nicht eindeutig der jeweiligen Kategorie zuzuordnen waren. Infolgedessen würde es sich anbieten, die verwendeten Piktogramme von Grund auf neu zu entwickeln und mithilfe einer separaten Nutzerstudie zu evaluieren, um die Erkennbarkeit der Benutzeroberfläche zu verbessern. Letztlich wurde angemerkt, dass die Textmarkierung auf einer Internetseite unter Umständen einige Sekunden in Anspruch nimmt und gelegentlich nicht den erwünschten Text markiert. Als Maßnahme bietet sich in diesem Fall ein verbesserter Algorithmus für die Suche und Markierung des geforderten Textes an.

Neben der hier aufgeführten Kritik wurden innerhalb der Studie auch einige Erweiterungen für das System vorgeschlagen. So könnte eine Datenbank angelegt werden, welche bereits analysierte DSEs enthält. Auf diese Weise wäre es möglich, dem Nutzer Änderungen in einer DSE zu signalisieren, indem dieser eine Benachrichtigung erhält, wenn Änderungen stattgefunden haben. Zusätzlich würde es diese Datenbank ermöglichen, Kommentare von Nutzern zu speichern, die weitere Anmerkungen bzgl. einer DSE haben. Durch diese nutzergenerierten Anmerkungen könnten weitere wichtige Informationen dargestellt werden.

Anhang A

Liste der analysierten Internetseiten

Folgende Internetseiten wurden mit dem Ziel analysiert, Gemeinsamkeiten im Bezug auf die Verlinkung der DSE zu finden. Zusätzlich wurde jede DSE dieser Seiten auf Merkmale untersucht, die deutlich machen, dass es sich bei dieser auch tatsächlich um eine solche handelt. Wurde eine Internetseite in mehreren Sprachen angeboten, so wurde jeweils die englischsprachige (USA) Version der Internetseite betrachtet. Die Seiten wurden bis auf wenige Ausnahme von folgender Internetseite ausgewählt: <https://moz.com/top500/>. Dies wurde getan, um das normale Surf-Verhalten von Internetnutzern bestmöglich abzubilden. Hat eine Firma auf die DSE des Mutterkonzerns verlinkt, so wird nur diese hier aufgelistet. Diese Analyse wurde am 17.04.2021 durchgeführt. Dort fand auch der letzte Aufruf statt.

1. <https://www.mozilla.org/en-US/privacy/>
2. <https://www.help.netflix.com/legal/privacy/>
3. <https://www.twitter.com/en/privacy/>
4. <https://www.npmjs.com/policies/privacy/>
5. <https://www.docs.github.com/en/github/site-policy/github-privacy-statement/>
6. https://www.meta.wikimedia.org/wiki/Privacy_policy/
7. <https://www.redditinc.com/policies/privacy-policy/>
8. <https://www.ebay.com/help/policies/member-behaviour-policies/user-privacy-notice-privacy-policy?id=4260/>
9. <https://www.corporate.walmart.com/privacy-security/>

10. <https://www.imdb.com/privacy/>
11. <https://www.privacy.thewaltdisneycompany.com/en/current-privacy-policy/>
12. <https://www.foxnews.com/privacy-policy/>
13. <https://www.privacy.microsoft.com/en-us/privacystatement/>
14. <https://www.about.usps.com/who/legal/privacy-policy/>
15. https://www.paypal.com/de/webapps/mpp/ua/privacy-full?locale.x=en_DE/
16. <https://www.etsy.com/legal/privacy/>
17. <https://www.nytimes.com/privacy/privacy-policy/>
18. <https://www.privacy.zillowgroup.com/>
19. <https://www.quora.com/about/privacy/>
20. <https://www.twitch.tv/p/en/legal/privacy-notice/>
21. <https://www.apple.com/legal/privacy/en-ww/>
22. <https://www.healthline.com/about/privacy-policy/>
23. <https://www.accuweather.com/de/privacy/>
24. <https://www.cambridge.org/about-us/legal-notices/privacy-notice/>
25. <https://www.about.urbandictionary.com/privacy/>
26. https://www.genius.com/static/privacy_policy/
27. <https://www.fandango.com/policies/privacy-policy/>
28. <https://www.policies.immediate.co.uk/privacy/>
29. <https://www.forbes.com/privacy/english/>
30. <https://www.diy.com/customer-support/privacy/>
31. <https://www.songfacts.com/blog/pages/songfacts-privacy-policy-and-legal-notice/>
32. <https://www.pmc.com/privacy-policy/>
33. <https://www.hearst.com/-/us-magazines-privacy-notice/>
34. <https://www.xero.com/about/legal/privacy/>

35. <https://www.wordpress.org/about/privacy/>
36. <https://www.istockphoto.com/en/legal/privacy-policy/>
37. <https://www.brandbucket.com/privacy-policy/>
38. <https://www.squareup.com/us/en/legal/general/privacy/>
39. <https://www.latimes.com/privacy-policy/>
40. <https://www.cloudflare.com/privacypolicy/>
41. <https://www.policies.google.com/privacy?hl=en-US>
42. <https://www.linkedin.com/legal/privacy-policy/>
43. https://www.europa.eu/european-union/abouteuropa/privacy-policy_en/
44. <https://www.adobe.com/privacy/policy.html>
45. <https://www.vimeo.com/privacy/>
46. <https://www.warnermediaprivacy.com/policycenter/b2c/WMNS/en-us/>
47. <https://www.facebook.com/privacy/explanation/>
48. https://www.amazon.com/gp/help/customer/display.html?nodeId=468496&ref_=footer_privacy/
49. <https://telegram.org/privacy/>
50. <https://www.dropbox.com/privacy/>
51. <https://www.w3.org/Consortium/Legal/privacy-statement-20140324/>
52. <https://www.creativecommons.org/privacy/>
53. <https://www.slideshare.net/privacy/>
54. <https://www.verizonmedia.com/policies/ie/en/verizonmedia/privacy/index.html/>
55. <https://www.issuu.com/legal/privacy/>
56. <https://www.myspace.com/pages/privacy/>
57. <https://www.thomsonreuters.com/en/privacy-statement.html/>
58. <https://www.theguardian.com/help/privacy-policy/>
59. <https://www.opera.com/privacy/>

60. <https://www.whatsapp.com/legal/privacy-policy-eea?lang=en/>
61. <https://www.dailymotion.com/legal/privacy/>
62. <https://www.policy.medium.com/medium-privacy-policy-f03bf92035c9/>
63. <https://www.jimdo.com/info/privacy/>
64. https://www.hugedomains.com/privacy_policy.cfm/
65. <https://www.un.org/en/about-us/privacy-notice/>
66. <https://www.fandom.com/privacy-policy/>
67. <https://www.huffpost.com/static/privacy-policy/>
68. <https://www.dowjones.com/privacy-notice/>
69. <https://www.dailymail.co.uk/home/article-7759273/Privacy-Cookies-Policy-Policy.html/>
70. <https://www.apache.org/foundation/policies/privacy.html>
71. <https://www.cpanel.net/privacy-policy/>
72. <https://www.independent.co.uk/service/privacy-notice-a6184181.html/>
73. <https://www.youradchoices.com/privacy-policy/>
74. <https://www.washingtonpost.com/discussions/2021/01/01/privacy-policy/>
75. <https://cm.usatoday.com/privacy/>
76. <https://support.scribd.com/hc/en-us/articles/210129366-Privacy-policy/>
77. <https://www.samsung.com/us/account/privacy-policy/>
78. <https://www.endurance.com/privacy/privacy/>
79. <https://www.bloomberg.com/notices/privacy/>
80. <https://automattic.com/privacy/>
81. <https://www.condenast.com/privacy-policy#privacypolicy>
82. <https://www.harvard.edu/privacy-statement/>
83. https://www.mediafire.com/policies/privacy_policy.php

84. https://store.steampowered.com/privacy_agreement/
85. <https://www.booking.com/content/privacy.html>
86. <https://www.newsprivacy.co.uk/single/>
87. <https://www.who.int/about/policies/privacy/>
88. <https://www.insider-inc.com/privacy-policy/>
89. <https://www.dan.com/legal/privacy-policy/>
90. <https://www.redventures.com/legal/privacy-policy/>
91. <https://www.4shared.com/privacy.jsp>
92. <https://www.plesk.com/legal/#privacy-policy/>
93. <https://policy.pinterest.com/en/privacy-policy/>
94. <https://www.dotdash.com/dash-privacy-policy/>
95. https://terms.yelp.com/privacy/en_us/20200101_en_us/
96. <https://discord.com/privacy/>
97. <https://www.greenpeace.org/usa/privacy-policy/>
98. <https://legal.trustpilot.com/for-reviewers/end-user-privacy-terms/>
99. <https://www.spotify.com/us/privacy/>
100. https://www.asus.com/us/Terms_of_Use_Notice_Privacy_Policy/Privacy_Policy/

Anhang B

Skript zur Studie

B.1 Skript der Experimentalgruppe

B.1.1 Einführung

Willkommen zur Studie für meine Bachelorarbeit „Automatisierte Analyse und visuelle Aufbereitung von Datenschutzerklärungen“. Bereits vorab vielen Dank für Ihre Teilnahme. Bitte denken Sie daran, dass Ihre Stimme und der Bildschirm aufgenommen werden. Die Studie besteht aus vier Phasen. Sollten während der Studie Fragen aufkommen, bitte stellen Sie sie. Denken Sie bitte auch daran, dass dies eine „Think-Aloud“-Studie ist. Das bedeutet, dass Sie jeden Gedanken, den Sie während der Studie haben, laut aussprechen sollen. Sehen Sie beispielsweise einen Knopf, so könnten Sie „Ich weiß nicht genau, was dieser Button macht. Ich werde ihn jetzt einmal betätigen und schauen, was dann passiert“ sagen.

Haben Sie noch Fragen?

Wenn nicht, würde ich nun mit Ihrem Einverständnis mit der Aufnahme des Bildschirms und Ihrer Stimme beginnen.

B.1.2 Phase 1

Als Erstes werden Ihnen einige allgemeine Frage gestellt - Bitte beantworten Sie diese so ehrlich wie möglich.

Frage 1 Wie alt sind Sie? (Zahleingabe)

Frage 2 Was machen Sie beruflich? (Mehrfachauswahl)

Frage 3 Wie oft surfen Sie im Internet? (1: „Nie“ bis 5: „Mehrere Male täglich“)

Frage 4 Aus welchem Grund surfen Sie im Internet? (Mehrfachauswahl aus: „Beruflich“, „Privat“, „Sonstiges“)

Frage 5 Mir ist Datenschutz wichtig. (1: „Stimme nicht zu“ bis 5: „Stimme zu“)

Frage 6 Warum nicht? (Freitext, wird nur gezeigt, wenn die Antwort auf

Frage 5 im Bereich zwischen 1 und 3 lag)

Frage 7 Achten Sie im Allgemeinen darauf, ob die Webseite, die Sie gerade besuchen, eine Datenschutzerklärung besitzt? (1: „Nie“ bis 5: „Immer“)

Frage 8 Wie oft lesen Sie die Datenschutzerklärung einer Internetseite? (1: „Nie“ bis 5: „Immer“)

Frage 9 Sind manche Datenschutzerklärungen für Sie wichtiger als andere, obwohl Sie beide Dienste nutzen? Legen Sie beispielsweise mehr Wert auf Ihren Datenschutz bei WhatsApp, als bei Netflix? („Ja“ oder „Nein“)

Frage 10 Warum legen Sie bei manchen Diensten mehr Wert auf Ihren Datenschutz? (Freitext, wird nur angezeigt, wenn die Antwort auf Frage 9 „Ja“ war)

B.1.3 Phase 2

Erläuterung Sie sehen nun den Firefox-Browser. Oben rechts neben der URL-Leiste, sehen Sie ein ausgegrautes Logo der Leibniz Universität Hannover. Wenn Sie eine Internetseite besuchen und eine Datenschutzerklärung gefunden wurde, so wird das Logo blau und eine kleine Notification erscheint. Klicken Sie dann gern auf das Logo.

Besuchen Sie nun gern mozilla.org, entweder über den Favoritenlink in der Mitte des Bildschirms oder das Lesezeichen in der oberen Leiste. Ist das Icon blau geworden, so klicken Sie gern darauf.

Nehmen Sie sich jetzt gern Zeit, um die Oberfläche des Programms zu erkunden. Sollte etwas unklar sein, fragen Sie gern.

Nutzen Sie bitte die Browser Extension als Unterstützung der folgenden Aufgaben.

Aufgabe 1 Bitte suchen Sie nun die Datenschutzerklärung von mozilla.org. (Zeitmessung)

Bitte beantworten Sie mir nun folgende Fragen:

Aufgabe 2 Sammelt Mozilla Daten über Sie und wenn ja, auf welche Weise geschieht dies? (Zeitmessung)

Aufgabe 3 Beschützt Mozilla Ihre Daten und wenn ja, auf welche Weise? (Zeitmessung)

Bitte sagen Sie mir Bescheid, wenn Sie Fragen haben oder Sie fertig sind.

Füllen Sie nun den Fragebogen bezüglich dieser Phase aus.

Frage 11 Durch die Browser Extension konnte ich die Datenschutzerklärung einfach finden. (1: „Stimme nicht zu“ bis 5: „Stimme zu“)

Frage 12 Durch die Browser Extension konnte ich die gewünschten

Informationen in der Datenschutzerklärung einfach finden. (1: „Stimme nicht zu“ bis 5: „Stimme zu“)

Frage 13 Zu welchen Themen haben Sie, zusätzlich zu Ihrer eigentlichen Aufgabe, Informationen in der Datenschutzerklärung gefunden? (Mehrfachauswahl aus: „Weitergabe von Daten an Dritte“, „Datenspeicherung“, „Veränderungen in der Datenschutzerklärung“, „Datensicherheit“, „Do-Not-Track“, „Gebrauch der Nutzerdaten“, „Zugriff, Bearbeiten und Löschen der eigenen Daten“, „Steuerung der gespeicherten Daten“, „Verkauf von Daten“, „Informationen über Cookies“, „Kontaktaufnahmemöglichkeiten“, „Sonstiges“)

Frage 14 Ich bin mir sicher zu wissen, welche Daten der Betreiber der Internetseite über mich sammelt und was mit diesen passiert. („Ja“ oder „Nein“)

Frage 15 Welche Daten werden gesammelt und was passiert mit ihnen? (Freitext)

Frage 16 Haben Sie die kurzen Informationskategorien, wie zum Beispiel „First Party Collection“, in der Benutzeroberfläche der Browser Extension gesehen? („Ja“ oder „Nein“)

Frage 17 Was könnten diese bedeuten? (Freitext, wird nur angezeigt, wenn die Antwort auf Frage 16 „Ja“ war)

Frage 18 Die gezeigten Informationskategorien haben dazu beigetragen, mehr über den Inhalt der Datenschutzerklärung zu verstehen. (1: „Stimme nicht zu“ bis 5: „Stimme zu“, wird nur angezeigt, wenn die Antwort auf Frage 16 „Ja“ war)

Frage 19 Haben Sie die Icons neben den Informationskategorien gesehen? („Ja“ oder „Nein“)

Frage 20 Die Icons sind eine sinnvolle visuelle Hilfe, um die gezeigten Informationskategorien zu repräsentieren. (1: „Stimme nicht zu“ bis 5: „Stimme zu“, wird nur angezeigt, wenn die Antwort auf Frage 19 „Ja“ war)

Frage 21 Haben Sie auf die Lupe neben den Texten geklickt? („Ja“ oder „Nein“)

Frage 22 Die Markierung des Textes ist hilfreich bei der Suche nach relevanten Informationen. (1: „Stimme nicht zu“ bis 5: „Stimme zu“, wird nur angezeigt, wenn die Antwort auf Frage 21 „Ja“ war)

B.1.4 Phase 3

Denken Sie bitte auch hier an die Browser Extension.

Aufgabe 4 Bitte besuchen Sie nun die Datenschutzerklärung von netflix.com. (Zeitmessung)

Bitte beantworten Sie mir nun folgende Fragen:

Aufgabe 5 Nutzt Netflix Cookies? Wenn ja, wieso? (Zeitmessung)

Aufgabe 6 Sammelt Netflix Daten über Sie? Wenn ja, auf welche Weise? (Zeitmessung)

Bitte sagen Sie mir Bescheid, wenn Sie Fragen haben oder Sie fertig sind.

Füllen Sie nun den Fragebogen bezüglich dieser Phase aus.

Frage 23 Durch die Browser Extension konnte ich die Datenschutzerklärung einfach finden. (1: „Stimme nicht zu“ bis 5: „Stimme zu“)

Frage 24 Durch die Browser Extension konnte ich die gewünschten Informationen in der Datenschutzerklärung einfach finden. (1: „Stimme nicht zu“ bis 5: „Stimme zu“)

Frage 25 Zu welchen Themen haben Sie, zusätzlich zu Ihrer eigentlichen Aufgabe, Informationen in der Datenschutzerklärung gefunden? (Mehrfachauswahl aus: „Weitergabe von Daten an Dritte“, „Datenspeicherung“, „Veränderungen in der Datenschutzerklärung“, „Datensicherheit“, „Do-Not-Track“, „Gebrauch der Nutzerdaten“, „Zugriff, Bearbeiten und Löschen der eigenen Daten“, „Steuerung der gespeicherten Daten“, „Verkauf von Daten“, „Informationen über Cookies“, „Kontaktaufnahmemöglichkeiten“, „Sonstiges“)

Frage 26 Ich bin mir sicher zu wissen, welche Daten der Betreiber der Internetseite über mich sammelt und was mit diesen passiert. („Ja“ oder „Nein“)

Frage 27 Welche Daten werden gesammelt und was passiert mit ihnen? (Freitext)

Frage 28 Haben Sie den „Goto Privacy Policy“ Button gesehen? („Ja“ oder „Nein“)

Frage 29 Der „Goto Privacy Policy“ Button ist hilfreich dabei, die Datenschutzerklärung der Internetseite zu finden. (1: „Stimme nicht zu“ bis 5: „Stimme zu“, wird nur angezeigt, wenn die Antwort auf Frage 28 „Ja“ war)

Frage 30 Haben Sie den „Details“ Button gesehen? („Ja“ oder „Nein“)

Frage 31 Die auf der „Details“-Seite gezeigten Erklärungen sind hilfreich dabei, die jeweiligen Kategorien zu verstehen. (1: „Stimme nicht zu“ bis 5: „Stimme zu“, wird nur angezeigt, wenn die Antwort auf Frage 30 „Ja“ war)

Frage 32 Die Anzeige der, zu der jeweiligen Kategorie passenden, Texte auf der „Details“-Seite ist hilfreich beim Verständnis der Datenschutzerklärung. (1: „Stimme nicht zu“ bis 5: „Stimme zu“, wird nur angezeigt, wenn die Antwort auf Frage 30 „Ja“ war)

B.1.5 Phase 4

Bitte füllen Sie zum Abschluss die letzte Seite des Fragebogens aus.

ID	Aussage
A1	Ich kann mir sehr gut vorstellen, das System regelmäßig zu nutzen.
A2	Ich empfinde das System als unnötig komplex.
A3	Ich empfinde das System als einfach zu nutzen.
A4	Ich denke, dass ich technischen Support brauchen würde, um das System zu nutzen.
A5	Ich finde, dass die verschiedenen Funktionen des Systems gut integriert sind.
A6	Ich finde, dass es im System zu viele Inkonsistenzen gibt.
A7	Ich kann mir vorstellen, dass die meisten Leute das System schnell zu beherrschen lernen.
A8	Ich empfinde die Bedienung als sehr umständlich.
A9	Ich habe mich bei der Nutzung des Systems sehr sicher gefühlt.
A10	Ich musste eine Menge Dinge lernen, bevor ich mit dem System arbeiten konnte.

Tabelle B.1: Der SUS-Fragebogen

Frage 33 Bitte lesen Sie jede der folgenden Aussage sorgfältig durch und wählen Sie dann, wie stark Sie ihr zustimmen. (Hierbei handelt es sich um den SUS-Fragebogen, für die eigentlichen Fragen s. Tabelle B.1, 1: „Stimme nicht zu“ bis 5: „Stimme zu“)

Frage 34 Die Browser Extension hat mich dabei unterstützt, einen Überblick über die Datenschutzerklärungen zu erhalten. (1: „Stimme nicht zu“ bis 5: „Stimme zu“)

Frage 35 Warum haben Sie sich so entschieden? (Freitext)

Frage 36 Das aktuelle Icon der Browser Extension ist blau, wenn eine Datenschutzerklärung analysiert wurde und grau, wenn dies nicht der Fall ist. Denken Sie, es wäre hilfreich mit mehr als diesen beiden Farben zu arbeiten? Beispielsweise rot, wenn viele oder wichtige persönliche Daten gesammelt werden. (Mehrfachauswahl aus: „Nein, zwei Farben genügen“, „Ja, eine zusätzliche Farbe wäre hilfreich um mehr Informationen bezüglich der gesammelten Daten zu erlangen“, „Ja, mehr Farben wären besser um anzuzeigen zu können, wie viele Daten gesammelt werden“, „Sonstiges“)

Frage 37 Haben Sie Verbesserungsvorschläge für die Darstellung der Ergebnisse? (Freitext)

B.1.6 Abschluss

Vielen Dank für Ihre Teilnahme. Sie können TeamViewer nun beenden, deinstallieren und den BigBlueButton-Tab schließen.

B.2 Skript der Kontrollgruppe

B.2.1 Einführung

Willkommen zur Studie für meine Bachelorarbeit „Automatisierte Analyse und visuelle Aufbereitung von Datenschutzerklärungen“. Bereits vorab vielen Dank für Ihre Teilnahme. Bitte denken Sie daran, dass Ihre Stimme und der Bildschirm aufgenommen werden. Die Studie besteht aus vier Phasen. Sollten während der Studie Fragen aufkommen, bitte stellen Sie sie. Denken Sie bitte auch daran, dass dies eine „Think-Aloud“-Studie ist. Das bedeutet, dass Sie jeden Gedanken, den Sie während der Studie haben, laut aussprechen sollen. Sehen Sie beispielsweise einen Knopf, so könnten Sie „Ich weiß nicht genau, was dieser Button macht. Ich werde ihn jetzt einmal betätigen und schauen, was dann passiert“ sagen.

Haben Sie noch Fragen?

Wenn nicht, würde ich nun mit Ihrem Einverständnis mit der Aufnahme des Bildschirms und Ihrer Stimme beginnen.

B.2.2 Phase 1

Als Erstes werden Ihnen einige allgemeine Frage gestellt - Bitte beantworten Sie diese so ehrlich wie möglich.

Frage 1 Wie alt sind Sie? (Zahleingabe)

Frage 2 Was machen Sie beruflich? (Mehrfachauswahl)

Frage 3 Wie oft surfen Sie im Internet? (1: „Nie“ bis 5: „Mehrere Male täglich“)

Frage 4 Aus welchem Grund surfen Sie im Internet? (Mehrfachauswahl aus: „Beruflich“, „Privat“, „Sonstiges“)

Frage 5 Mir ist Datenschutz wichtig. (1: „Stimme nicht zu“ bis 5: „Stimme zu“)

Frage 6 Warum nicht? (Freitext, wird nur gezeigt, wenn die Antwort auf Frage 5 im Bereich zwischen 1 und 3 lag)

Frage 7 Achten Sie im Allgemeinen darauf, ob die Webseite, die Sie gerade besuchen, eine Datenschutzerklärung besitzt? (1: „Nie“ bis 5: „Immer“)

Frage 8 Wie oft lesen Sie die Datenschutzerklärung einer Internetseite? (1: „Nie“ bis 5: „Immer“)

Frage 9 Sind manche Datenschutzerklärungen für Sie wichtiger als andere, obwohl Sie beide Dienste nutzen? Legen Sie beispielsweise mehr Wert auf Ihren Datenschutz bei WhatsApp, als bei Netflix? („Ja“ oder „Nein“)

Frage 10 Warum legen Sie bei manchen Diensten mehr Wert auf Ihren

Datenschutz? (Freitext, wird nur angezeigt, wenn die Antwort auf Frage 9 „Ja“ war)

B.2.3 Phase 2

Erläuterung Sie sehen nun den Firefox-Browser. Besuchen Sie jetzt mozilla.org, entweder über den Favoritenlink in der Mitte des Bildschirms oder das Lesezeichen in der oberen Leiste.

Aufgabe 1 Bitte suchen Sie nun die Datenschutzerklärung von mozilla.org. (Zeitmessung)

Bitte beantworten Sie mir nun folgende Fragen:

Aufgabe 2 Sammelt Mozilla Daten über Sie und wenn ja, auf welche Weise geschieht dies? (Zeitmessung)

Aufgabe 3 Beschützt Mozilla Ihre Daten und wenn ja, auf welche Weise? (Zeitmessung)

Bitte sagen Sie mir Bescheid, wenn Sie Fragen haben oder Sie fertig sind.

Füllen Sie nun den Fragebogen bezüglich dieser Phase aus.

Frage 11 Die Datenschutzerklärung der Internetseite war für mich leicht zu finden. (1: „Stimme nicht zu“ bis 5: „Stimme zu“)

Frage 12 Die gewünschten Informationen aus der Datenschutzerklärung waren für mich leicht zu finden. (1: „Stimme nicht zu“ bis 5: „Stimme zu“)

Frage 13 Der Inhalt der Datenschutzerklärung war für mich leicht zu verstehen. (1: „Stimme nicht zu“ bis 5: „Stimme zu“)

Frage 14 Zu welchen Themen haben Sie, zusätzlich zu Ihrer eigentlichen Aufgabe, Informationen in der Datenschutzerklärung gefunden? (Mehrfachauswahl aus: „Weitergabe von Daten an Dritte“, „Datenspeicherung“, „Veränderungen in der Datenschutzerklärung“, „Datensicherheit“, „Do-Not-Track“, „Gebrauch der Nutzerdaten“, „Zugriff, Bearbeiten und Löschen der eigenen Daten“, „Steuerung der gespeicherten Daten“, „Verkauf von Daten“, „Informationen über Cookies“, „Kontaktaufnahmemöglichkeiten“, „Sonstiges“)

Frage 15 Ich bin mir sicher zu wissen, welche Daten der Betreiber der Internetseite über mich sammelt und was mit diesen passiert. („Ja“ oder „Nein“)

Frage 16 Welche Daten werden gesammelt und was passiert mit ihnen? (Freitext, wird nur angezeigt, wenn die Antwort auf Frage 15 „Ja“ war)

B.2.4 Phase 3

Aufgabe 4 Bitte besuchen Sie nun die Datenschutzerklärung von netflix.com. (Zeitmessung)

Bitte beantworten Sie mir nun folgende Fragen:

Aufgabe 5 Nutzt Netflix Cookies? Wenn ja, wieso? (Zeitmessung)

Aufgabe 6 Sammelt Netflix Daten über Sie? Wenn ja, auf welche Weise? (Zeitmessung)

Bitte sagen Sie mir Bescheid, wenn Sie Fragen haben oder Sie fertig sind.

Füllen Sie nun den Fragebogen bezüglich dieser Phase aus.

Frage 17 Die Datenschutzerklärung der Internetseite war für mich leicht zu finden. (1: „Stimme nicht zu“ bis 5: „Stimme zu“)

Frage 18 Die gewünschten Informationen aus der Datenschutzerklärung waren für mich leicht zu finden. (1: „Stimme nicht zu“ bis 5: „Stimme zu“)

Frage 19 Der Inhalt der Datenschutzerklärung war für mich leicht zu verstehen. (1: „Stimme nicht zu“ bis 5: „Stimme zu“)

Frage 20 Zu welchen Themen haben Sie, zusätzlich zu Ihrer eigentlichen Aufgabe, Informationen in der Datenschutzerklärung gefunden? (Mehrfachauswahl aus: „Weitergabe von Daten an Dritte“, „Datenspeicherung“, „Veränderungen in der Datenschutzerklärung“, „Datensicherheit“, „Do-Not-Track“, „Gebrauch der Nutzerdaten“, „Zugriff, Bearbeiten und Löschen der eigenen Daten“, „Steuerung der gespeicherten Daten“, „Verkauf von Daten“, „Informationen über Cookies“, „Kontaktaufnahmemöglichkeiten“, „Sonstiges“)

Frage 21 Ich bin mir sicher zu wissen, welche Daten der Betreiber der Internetseite über mich sammelt und was mit diesen passiert. („Ja“ oder „Nein“)

Frage 22 Welche Daten werden gesammelt und was passiert mit ihnen? (Freitext, wird nur angezeigt, wenn die Antwort auf Frage 21 „Ja“ war)

B.2.5 Phase 4

Bitte füllen Sie zum Abschluss die letzte Seite des Fragebogens aus.

Frage 23 Ich habe das Gefühl, umfassend über meinen Datenschutz aufgeklärt worden zu sein. (1: „Stimme nicht zu“ bis 5: „Stimme zu“)

Frage 24 Warum haben Sie sich so entschieden? (Freitext)

B.2.6 Abschluss

Vielen Dank für Ihre Teilnahme. Sie können TeamViewer nun beenden, deinstallieren und den BigBlueButton-Tab schließen.

Anhang C

Installationsanleitung

Hier wird eine kurze Erklärung für die Installation und den Start des entwickelten Systems gegeben.

C.1 Installation der Browser Extension

C.1.1 Kompilierung

Möchten Sie die BE installieren, so liegen im Ordner „/privacy-check/extension-builds/“ bereits Dateien für diesen Zweck, die Sie verwenden können. Überspringen Sie den Rest dieses Abschnitts (Unterabschnitt C.1.1), wenn sie verwendet werden sollen.

Möchten Sie die Browser Extension selber kompilieren, so gehen Sie wie folgt vor (Sie benötigen „NodeJS“ und den Package-Manager „Yarn“):

1. Gehen Sie in den Ordner „/privacy-check/“.
2. Führen Sie „yarn install“ aus, um möglicherweise nicht vorhandene Abhängigkeiten zu installieren.
3. Führen Sie „yarn run build:<browser>“ aus.
Ersetzen Sie „<browser>“ durch Ihren Browser. Die kompatiblen Browser sind Firefox, Chrome und Opera. Sie müssen den Namen des jeweiligen Browser im Kommando klein schreiben, wobei sie für Opera auch die Chrome-Version kompilieren müssen.

C.1.2 Installation

Installation in Firefox:

1. Öffnen Sie Firefox.
2. Geben Sie „about:debugging“ in die URL-Leiste des Browsers ein.

3. Klicken Sie auf „Dieser Firefox“.
4. Klicken Sie anschließend auf „Temporäres Add-on laden...“.
5. Wählen Sie, in dem sich öffnenden Dateibrowser, die (wie in Unterabschnitt C.1.1 beschrieben) kompilierte „firefox.xpi“-Datei aus.

Installation in Chrome:

1. Öffnen Sie Chrome.
2. Geben sie „chrome://extensions/“ in die URL-Leiste des Browsers ein.
3. Aktivieren sie den Entwicklermodus.
4. Schieben Sie, per Drag-And-Drop, die (wie in Unterabschnitt C.1.1 beschrieben) kompilierte „chrome.zip“-Datei in das Erweiterungen-Fenster des Browsers.

Installation in Opera:

1. Öffnen Sie Opera.
2. Geben sie „opera://extensions/“ in die URL-Leiste des Browsers ein.
3. Aktivieren sie den Entwicklermodus.
4. Schieben Sie, per Drag-And-Drop, die (wie in Unterabschnitt C.1.1 beschrieben) kompilierte „chrome.zip“-Datei in das Erweiterungen-Fenster des Browsers.

C.2 Installation und Start des Analyse-Servers

Für die Installation des Analyse-Servers benötigen Sie die Programme „Docker“ und „Docker-Compose“. Sind diese installiert, führen Sie das Kommando „docker-compose up“ aus. Dieser Befehl lädt alle benötigten Abhängigkeiten herunter und startet anschließend den Analyse-Server.

Anhang D

Inhalt des beiliegenden USB-Sticks

Auf dem beiliegenden USB-Stick findet sich folgender Inhalt:

- Dieses Dokument („BA-Koehler2021.pdf“).
- Die Latex-Quelldateien dieses Dokuments („BA-Koehler2021.zip“).
- Ein Order namens „Videos“. Hier sind die Videoaufnahmen zu finden, die während der Studie entstanden sind.
- Das Git-Repository („ba-koehler“), welches genutzt wurde, um das in dieser Arbeit vorgestellte System zu entwickeln.
Darin lässt sich folgender Inhalt finden:
 - Im Ordner „/OPP-115/“: Der für das Training der Algorithmen genutzte Datensatz.
 - Im Ordner „/privacy-check/“: Die Privacy Policy Check Browser Extension.
 - Im Ordner „/server/“: Der Flask-Server zusammen mit dem Klassifikationsalgorithmus.
 - Im Ordner „/Studie/“: Informationen über die durchgeführte Studie. Zusätzlich Skripte und andere Dateien, die für die Auswertung der Studie genutzt worden sind.

Akronyme

Bezeichnung	Beschreibung
BE	Browser Extension 2, 3, 27, 29–33, 36, 38, 39, 42–49, 51–53, 55, 57, 58, 77
DSE	Datenschutzerklärung 1–3, 16–27, 29, 30, 32, 33, 35, 36, 38–48, 51–53, 55–59, 61
EB	Entscheidungsbaum 10, 11
GDPR	General Data Protection Regulation 18
HTML	Hypertext Markup Language 22, 23, 33, 58
NLP	Natural Language Processing 2, 6, 56
P3P	Platform for Privacy Preferences 56
PPC	Privacy Policy Check 15, 16, 23, 26, 29, 30, 32, 34, 35, 40, 47, 52, 56, 58, 59
RQ	Research Question 37, 39, 42–44, 46, 51, 52, 85
URL	Uniform Resource Locator 16–19, 22
W3C	World Wide Web Consortium 56
XML	Extensible Markup Language 56

Abbildungsverzeichnis

1.1	Überblick über den Nutzungsverlauf	2
2.1	Venn-Diagramm Natural Language Processing	6
2.2	Einige Beispiele für Klassifikationen	9
2.3	Witterungsdaten	10
2.4	Der Entscheidungsbaum für das Sportbeispiel	11
2.5	Klassifizierung von unbekanntem Daten	11
3.1	Der Nutzungsablauf des Privacy Policy Checks	15
3.2	Anzahl der Einträge zu der jeweiligen Kategorie	20
3.3	Trainings-Pipeline der Modelle	22
3.4	Die Analyse einer Datenschutzerklärung	22
3.5	Die genutzten Piktogramme	24
3.6	Initiales Mockup der Benutzeroberfläche	25
4.1	Die Architektur des Systems	30
4.2	Die Benutzeroberfläche des Privacy Policy Check	34
4.3	Der Detailbereich des Privacy Policy Check	35
5.1	Die Forschungsfragen und die dazugehörigen Metriken	38
5.2	Legende der geschlossenen Fragen	41
5.3	Reaktionen auf die Aussage „Die Datenschutzerklärung war einfach zu finden.“	44
5.4	Reaktionen auf die Aussage „Die Informationen aus der Datenschutzerklärung waren einfach zu finden.“	46
5.5	Angaben zum Verständnis einer Datenschutzerklärung	47
5.6	Verteilung der SUS-Bewertungen	48
5.7	Abbildung der SUS-Bewertung auf weitere Skalen [49]	49

Tabellenverzeichnis

2.1	Konfusionsmatrix	12
3.1	Die analysierten Kategorien	21
3.2	Anforderungen an die Implementierung	27
4.1	Berechnete Metriken	34
4.2	Die getesteten Systemkonfigurationen	36
5.1	Die Teststatistiken und p-Werte für die RQ1-Messungen . . .	43
5.2	Die Teststatistiken und p-Werte für die RQ2-Messungen . . .	46
B.1	Der SUS-Fragebogen	71

Literaturverzeichnis

- [1] A. M. McDonald and L. Cranor, “The cost of reading privacy policies,” *A Journal of Law and Policy for the Information Society*, vol. 4, pp. 543–568, 2008.
- [2] C. Jensen and C. Potts, *Privacy Policies as Decision-Making Tools: An Evaluation of Online Privacy Notices*. New York, NY, USA: Association for Computing Machinery, 2004, p. 471–478.
- [3] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua, and S. Guo, “Protection of big data privacy,” *IEEE Access*, vol. 4, pp. 1821–1834, 2016.
- [4] J. Van Den Hoven, *Information Technology, Privacy, and the Protection of Personal Data*, ser. Cambridge Studies in Philosophy and Public Policy. Cambridge University Press, 2008, p. 301–321.
- [5] K. Renaud and D. Gálvez-Cruz, “Privacy: Aspects, definitions and a multi-faceted privacy preservation approach,” in *2010 Information Security for South Africa*, 2010, pp. 1–8.
- [6] S. M. Jourard, “Some psychological aspects of privacy,” *Law and Contemporary Problems*, vol. 31, no. 2, pp. 307–318, 1966.
- [7] Dudenredaktion. (o. J.) Datenschutz. [Online]. Available: <https://www.duden.de/node/30540/revision/30569>
- [8] T. Reigeluth, “Why data is not enough: Digital traces as control of self and self-control,” *Surveillance and Society*, vol. 12, no. 2, pp. 243–254, 05 2014.
- [9] W. T. DeVries, “Protecting privacy in the digital age,” *Berkeley Technology Law Journal*, vol. 18, no. 1, pp. 283–312, 2003.
- [10] B. Knijnenburg, M. Willemsen, Z. Gantner, H. Soncu, and C. Newell, “Explaining the user experience of recommender systems,” *User Modeling and User-Adapted Interaction*, vol. 22, pp. 441–504, 10 2012.

- [11] J. Sovern, “Opting in, opting out, or no options at all: The fight for control of personal information,” *Washington Law Review*, vol. 74, no. 4, pp. 1033–1118, 1999.
- [12] P. Schmechel, *Verbraucherdatenschutzrecht in der EU-Datenschutz-Grundverordnung*, 1st ed., ser. Verbraucherrecht 2.0 - Verbraucher in der digitalen Welt. Baden-Baden: Nomos Verlagsgesellschaft mbH & Co. KG, Dec 2017, pp. 265–302.
- [13] J. E. Cohen, “What privacy is for,” *Harv. L. Rev.*, vol. 126, pp. 1904–1933, 2012.
- [14] Y. S. Mandloi and Y. Inada, “Machine learning approach for drone perception and control,” in *Engineering Applications of Neural Networks*, J. Macintyre, L. Iliadis, I. Maglogiannis, and C. Jayne, Eds. Cham: Springer International Publishing, 2019, pp. 424–431.
- [15] J. Amin, M. Sharif, M. Raza, T. Saba, and M. A. Anjum, “Brain tumor detection using statistical and machine learning method,” *Computer Methods and Programs in Biomedicine*, vol. 177, pp. 69–79, 2019.
- [16] R. Gupta and C. Pathak, “A machine learning framework for predicting purchase by online customers based on dynamic pricing,” *Procedia Computer Science*, vol. 36, pp. 599–605, 2014, complex Adaptive Systems Philadelphia, PA November 3-5, 2014.
- [17] N. Burkart and M. F. Huber, “A survey on the explainability of supervised machine learning,” *Journal of Artificial Intelligence Research*, vol. 70, p. 245–317, Jan 2021.
- [18] L. Chazette, O. Karras, and K. Schneider, “Do end-users want explanations? analyzing the role of explainability as an emerging aspect of non-functional requirements,” in *2019 IEEE 27th International Requirements Engineering Conference (RE)*, 2019, pp. 223–233.
- [19] L. Chazette and K. Schneider, “Explainability as a non-functional requirement: challenges and recommendations,” *Requirements Engineering*, vol. 25, no. 4, pp. 493–514, Dec. 2020.
- [20] D. Guthrie, B. Allison, W. Liu, L. Guthrie, and Y. Wilks, “A closer look at skip-gram modelling,” in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*. Genoa, Italy: European Language Resources Association (ELRA), May 2006.
- [21] J. B. Lovins, “Development of a stemming algorithm,” *Mechanical Translation and Computational Linguistics*, vol. 11, pp. 22–31, 1968.

- [22] A. Uyar, “Google stemming mechanisms,” *Journal of Information Science*, vol. 35, no. 5, pp. 499–514, 2009.
- [23] M. Bramer, *Principles of Data Mining*, 2nd ed. Springer Publishing Company, Incorporated, 2013.
- [24] I. Rish, “An empirical study of the naïve bayes classifier,” *IJCAI 2001 Workshop On Empirical Methods in Artificial Intelligence*, vol. 3, no. 22, pp. 41–46, 2001.
- [25] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, p. 5–32, Oct. 2001.
- [26] T. K. Ho, “Random decision forests,” in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, 1995, pp. 278–282 vol.1.
- [27] 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*, vol. 7, p. 285, 12 2016.
- [28] J. Thomczik, “Entwicklung eines tools zur erklärungs von datenschutzrichtlinien,” Bachelorarbeit, Leibniz Universität Hannover, Fachgebiet Software Engineering, 2021.
- [29] A. Y. Ng and M. I. Jordan, “On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes,” in *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, ser. NIPS’01. Cambridge, MA, USA: MIT Press, 2001, p. 841–848.
- [30] S. Wilson, F. Schaub, A. A. Dara, F. Liu, S. Cherivirala, P. Giovanni Leon, M. Schaarup Andersen, S. Zimmeck, K. M. Sathyendra, N. C. Russell, T. B. Norton, E. Hovy, J. Reidenberg, and N. Sadeh, “The creation and analysis of a website privacy policy corpus,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1330–1340.
- [31] A. Kulkarni and A. Shivananda, *Converting Text to Features*. Berkeley, CA: Apress, 2019, pp. 67–96.
- [32] S. Robertson, “Understanding inverse document frequency: On theoretical arguments for idf,” *Journal of Documentation*, vol. 60, no. 5, pp. 503–520, 2004.

- [33] S. Oviatt, “Human-centered design meets cognitive load theory: Designing interfaces that help people think,” in *Proceedings of the 14th ACM International Conference on Multimedia*, ser. MM ’06. New York, NY, USA: Association for Computing Machinery, 2006, p. 871–880.
- [34] 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*, ser. AUIC ’03. AUS: Australian Computer Society, Inc., 2003, p. 25–32.
- [35] S. E. Kettner, C. Thorun, and M. Vetter, “Wege zur besseren informiertheit,” https://www.conpolicy.de/data/user_upload/Studien/Bericht_ConPolicy_2018_02_Wege_zur_besseren_Informiertheit.pdf, 2018, Letzter Zugriff: 13.07.2021.
- [36] C. Tijus, J. Barcenilla, B. C. de Lavalette, and J.-G. Meunier, *The Design, Understanding and Usage of Pictograms*. Leiden, Niederlande: Brill, 2007, pp. 17 – 31.
- [37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in python.” *Journal of Machine Learning Research*, vol. 12, no. 10, pp. 2825 – 2830, 2011.
- [38] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. O’Reilly Media, Inc., 2009.
- [39] F. N. A. Al Omran and C. Treude, “Choosing an nlp library for analyzing software documentation: A systematic literature review and a series of experiments,” in *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*, 2017, pp. 187–197.
- [40] C. Nance, *TypeScript essentials : Develop large scale responsive web applications with TypeScript*, ser. Community Experience Distilled. Birmingham, England: Packt Publishing, 2014.
- [41] C. Wohlin, P. Runeson, M. Hst, M. C. Ohlsson, B. Regnell, and A. Wessln, *Experimentation in Software Engineering*. Springer Publishing Company, Incorporated, 2012.
- [42] T. Boren and J. Ramey, “Thinking aloud: reconciling theory and practice,” *IEEE Transactions on Professional Communication*, vol. 43, no. 3, pp. 261–278, 2000.
- [43] J. Brooke, *SUS - A quick and dirty usability scale*. CRC Press, June 1996.

- [44] ———, “Sus: a retrospective,” *Journal of usability studies*, vol. 8, no. 2, pp. 29–40, 2013.
- [45] S. S. Shapiro and M. B. Wilk, “An analysis of variance test for normality (complete samples),” *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.
- [46] J. Sauro and J. R. Lewis, *Quantifying the User Experience: Practical Statistics for User Research*, 1st ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2012.
- [47] J. T. Mordkoff, “The assumption(s) of normality,” <http://www2.psychology.uiowa.edu/faculty/mordkoff/GradStats/part%201/I.07%20normal.pdf>, 2016, Letzter Zugriff: 13.07.2021.
- [48] C.-C. Lin and G. S. Mudholkar, “A simple test for normality against asymmetric alternatives,” *Biometrika*, vol. 67, no. 2, pp. 455–461, 1980.
- [49] A. Bangor, P. Kortum, and J. Miller, “Determining what individual sus scores mean: Adding an adjective rating scale,” *Journal of Usability Studies*, vol. 4, no. 3, p. 114–123, May 2009.
- [50] M. S. Andreassen, H. V. Nielsen, S. O. Schröder, and J. Stage, “What happened to remote usability testing? an empirical study of three methods,” *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, p. 1405–1414, 2007.
- [51] H. Harkous, K. Fawaz, R. Lebre, F. Schaub, K. G. Shin, and K. Aberer, “Polisis: Automated analysis and presentation of privacy policies using deep learning,” in *27th USENIX Security Symposium (USENIX Security 18)*. Baltimore, MD: USENIX Association, Aug. 2018, pp. 531–548.
- [52] 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: USENIX Association, Jun. 2016.
- [53] L. Cranor, “P3p: making privacy policies more useful,” *IEEE Security Privacy*, vol. 1, no. 6, pp. 50–55, 2003.
- [54] L. F. Cranor, P. Guduru, and M. Arjula, “User interfaces for privacy agents,” *ACM Trans. Comput.-Hum. Interact.*, vol. 13, no. 2, p. 135–178, Jun. 2006.
- [55] N. Tomuro, S. Lytinen, and K. Hornsberg, “Automatic summarization of privacy policies using ensemble learning,” in *Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy*, ser. CODASPY ’16. New York, NY, USA: Association for Computing Machinery, 2016, p. 133–135.

