HOCHSCHULE
MITTWEIDA
University of Applied Sciences

BACHELORARBEIT

Herr
Dennis Henke

Automatisierte Analyse von Datenschutzerklärungen

2021

Fakultät **Angewandte Computer- und Biowissenschaften**

BACHELORARBEIT

Automatisierte Analyse von Datenschutzerklärungen

Autor:
Dennis Henke

Studiengang:
IT-Forensik/Cybercrime

Seminargruppe:
CC17w1-B

Matrikelnummer:
46490

Erstprüfer:
Prof. Ronny Bodach

Zweitprüfer:
Dipl.-Inform. Eric Clausing

Mittweida, August 2021

Faculty of **Applied Computer Sciences and Biosciences**

BACHELOR THESIS

Automated analysis of privacy policies

Author:
Dennis Henke

Study Programme:
IT-Forensic/Cybercrime

Seminar Group:
CC17w1-B

Student Number:
46490

First Referee:
Prof. Ronny Bodach

Second Referee:
Dipl.-Inform. Eric Clausing

Mittweida, August 2021

Bibliografische Angaben

Henke, Dennis: Automatisierte Analyse von Datenschutzerklärungen, 75 Seiten, 13 Abbildungen, Hochschule Mittweida, University of Applied Sciences, Fakultät Angewandte Computer- und Biowissenschaften

Bachelorarbeit, 2021

Kurzfassung

Die Datenschutzgrundverordnung hat in der Europäischen Union für ein einheitliches Datenschutzrecht gesorgt. Sie fordert unter anderem eine einfache und verständliche Sprache von Datenschutzerklärungen und benennt umfangreiche, inhaltliche Anforderungen. Im Praxisumfeld ist es aber immer noch nicht selbstverständlich, dass Datenschutzerklärungen sich an formalen und inhaltlichen Festlegungen der Datenschutzgrundverordnung und weiterer geltender deutscher Gesetze orientieren. Das in dieser Arbeit erstellte Python-Programm zur automatisierten Analyse von Datenschutzerklärungen kategorisiert Themenabschnitte mit Hilfe einer Stichwortsuche und prüft die Datenschutzerklärung anhand eines festgelegten Katalogs an Testkriterien. Abschließend wird eine Bewertung auf Basis der Form und des Inhalts vorgenommen.

Abstract

The General Data Protection Regulation has ensured uniform privacy legislation in the European Union. It requires simple and comprehensible language for privacy policies and specifies extensive content requirements. In the practical environment, however, it is still not taken for granted that privacy statements are based on the formal and content specifications of the General Data Protection Regulation and other applicable German laws. The Python program created in this work, Automated analysis of privacy policies, categorizes sections using a keyword search and checks the privacy statement against a defined catalog of test criteria. Finally, an evaluation is made based on form and content.

I. Inhaltsverzeichnis

II. Abbildungsverzeichnis

III. Tabellenverzeichnis

IV. Quelltextverzeichnis

V. Abkürzungsverzeichnis

API ..... Application Programming Interface, Seite 24

Art ..... Artikel, Seite 7

ASL ..... Average Sentence Length, Seite 11

ASW ..... Average number of Syllables per Word, Seite 11

BDSG ..... Bundesdatenschutzgesetz, Seite 8

DDoS ..... Distributed Denial of Service, Seite 46

DoS ..... Denial of Service, Seite 46

DSGVO ..... Datenschutzgrundverordnung, Seite 7

HTML ..... Hypertext Markup Language, Seite 28

HTTP ..... Hypertext Transfer Protocol, Seite 29

ISO ..... International Organization for Standardization, Seite 31

JSON ..... JavaScript Object Notation, Seite 28

Kap ..... Kapitel, Seite 7

MS ..... Anteil der Worte mit mehr als 2 Silben, Seite 13

NLP ..... Natural Language Processing, Seite 14

NLTK ..... Natural Language Toolkit, Seite 31

PDW ..... Percentage of Difficult Words, Seite 12

SL ..... Mittlere Satzlänge, Seite 13

TMG ..... Telemediengesetz, Seite 10

TTDSG ..... Telekommunikations-Telemedien-Datenschutzgesetz, Seite 10

UUID ..... Universally Unique Identifier, Seite 26

YAML ..... YAML Ain't Markup Language, Seite 28

1 Einleitung

Mobile Applikationen sind in zahlreiche Lebensbereiche eingezogen und begleiten Smartphone-Nutzer tagtäglich in vielfältigen Situationen. Im Jahr 2019 wurden allein in Deutschland 1,6 Milliarden Euro mit mobilen Applikationen umgesetzt, für 2020 ist ein Umsatz von etwa 2 Milliarden prognostiziert worden (vgl. Krösmann und Olsok, 2020). Dieser Umsatz wurde mehr oder weniger bewusst, aber direkt von den Nutzern bzw. Käufern verursacht. Allerdings werden mit Hilfe von Applikationen verschiedene personenbezogene Daten gewonnen, welche von den Anbietern ausgewertet werden und erhöhtes Potenzial für eine Monetarisierung bieten. Die Anbieter der Applikationen müssen hierüber zwar in der jeweiligen Datenschutzerklärung informieren, da aber lediglich 13 % der Europäer diese vollständig lesen (vgl. European Commission, 2019), ist dem Nutzer nur selten bewusst, dass die gerade genutzte Applikation persönliche Daten aufzeichnet.

Der Hauptgrund für das nicht- bzw. nur teilweise Lesen von Datenschutzerklärungen liegt darin, dass diese als zu lang empfunden werden. Als zweiter Grund wurde im Rahmen einer 2019 europaweit durchgeführten Umfrage der Europäischen Kommission angegeben, dass Datenschutzerklärungen im Allgemeinen eher schwammig bzw. schwer verständlich formuliert sind. (vgl. ebd.)

Eine automatisierte Analyse von Datenschutzerklärungen schafft an dieser Stelle Abhilfe. Mit ihr wird ohne manuellen Aufwand geprüft, ob die jeweilige Datenschutzerklärung verständlich geschrieben wurde und sie über relevante Themen informiert. Das in dieser Arbeit entwickelte Programm wird Datenschutzerklärungen basierend auf ihrer Lesbarkeit und ihres Inhalts bewerten und dem Nutzer eine erste Einschätzung darüber liefern, welchen Informationsgrad diese hat und interessante Themenabschnitte zugänglich machen. Hiermit wird er in die Lage versetzt, die Qualität der Datenschutzerklärung zu erkennen und fehlende Informationen zu identifizieren.

1.1 Aufgabenstellung und Motivation

Die Datenschutzgrundverordnung hat in Europa die Notwendigkeit für umfassendere und transparentere Datenschutzerklärungen geschaffen. Dennoch sind diese häufig mehrere A4-Seiten lang und geben Informationen nur verklausuliert wieder. Im Rahmen dieser Bachelorarbeit soll ein Verfahren zur automatisierten Analyse von deutsch- und englischsprachigen Datenschutzerklärungen entwickelt werden, mit dem das vollständige Lesen der Datenschutzerklärung nicht mehr oder nur noch in Ausnahmefällen erforderlich ist. Die Analyse wird aufgrund des beruflichen Kontextes des Autors hauptsächlich auf die Analyse von Datenschutzerklärungen mobiler Applikationen ausgerichtet sein, aber im Konzept die Möglichkeit offenhalten, auch Datenschutzerklärungen ohne Applikationsbezug zu untersuchen. Die Analyse soll darlegen, ob

die untersuchte Datenschutzerklärung verständlich formuliert ist und die von der Datenschutzgrundverordnung und gegebenenfalls nationalen Gesetzen geforderten Informationen enthält. Die Datenschutzerklärung wird außerdem aufgrund eines festgelegten Katalogs bewertet. Um die Funktionsweise sicherzustellen, wird das entwickelte Analyseverfahren durch eine Massenanalyse von Datenschutzerklärungen gängiger mobiler Applikationen getestet. Die Ergebnisse der Analysen sollen außerdem in einer Auswertung vorgestellt werden.

1.2 Stand der Technik und Abgrenzung

Im Rahmen einer Vorrecherche konnten generell nur wenige Projekte gefunden werden, die sich mit der automatisierten Analyse von Datenschutzerklärungen befassten. Hierzu gehört Polisis[1], mit dem Datenschutzerklärungen auf Basis eines Machine Learning Ansatzes analysiert werden. Auf den Einsatzzweck trainierte Machine Learning Modelle wurden in der vorhergehenden Recherche allerdings nur für englischsprachige Datenschutzerklärungen gefunden, beispielsweise von der Carnegie Mellon University[2]. Weiterhin waren die meisten der gefundenen Ansätze nicht auf die Datenschutzgrundverordnung hin geeicht, sondern waren vor Erscheinung dieser entwickelt, wie etwa das FrontierProject[3]. Letzterer basiert nicht auf maschinellem Lernen, sondern auf einer statischen Kategorisierung von Textbausteinen. Auch kommerzielle Produkte wie Guard[4] stützen sich nur auf englischsprachige Texte.

In dieser Arbeit wird daher ein Analyseverfahren ausgearbeitet, mit dem sowohl deutsch- als auch englischsprachige Datenschutzerklärungen gleichermaßen untersucht und bewertet werden können. Da für deutschsprachige Datenschutzerklärung kein trainiertes Machine Learning Modell gefunden werden konnte und die Schaffung eines neuen Modells nur auf Basis vieler manuell klassifizierter Datenschutzerklärungen aufgebaut werden kann, wird ein Lösungsansatz auf Basis statischer Kategorisierung entwickelt, mit dem nach relevanten bzw. nach geltenden Bestimmungen notwendigen Bestandteilen der Datenschutzerklärungen gesucht wird.

[1] `https://pribot.org/polisis`
[2] `https://aclanthology.org/C14-1084.pdf`
[3] `https://github.com/adityamarella/frontierproject`
[4] `https://useguard.com/`

2 Grundlagen

Die zum Verständnis dieser Arbeit notwendigen Grundlagen werden in diesem Kapitel vorgestellt. Zunächst wird auf die geschichtliche Entwicklung des Datenschutzes eingegangen und relevante Begriffe, Gesetze und Verordnungen erläutert. Danach wird auf die Entwicklung von Algorithmen der Lesbarkeitsanalyse näher eingegangen und zuletzt relevante Teile der Textanalyse beschrieben.

2.1 Datenschutz

In diesem Abschnitt wird zu Beginn auf einige geschichtliche Aspekte der Entwicklung des Datenschutzes eingegangen. Relevante Begriffe, die in dieser Arbeit Verwendung finden, werden im Anschluss hieran erklärt. Abschließend werden einige Gesetze und Verordnungen näher erläutert und für diese Arbeit relevante Artikel und Paragrafen beschrieben.

2.1.1 Geschichte

Bereits in der Antike Griechenlands gab es im weitesten Sinne erste Entwicklungen zum Datenschutz, die noch heute Bestand haben (vgl. Moos, Schefzig und Arning, 2018). Der sog. Eid des Hippokrates wird heute zwar nicht mehr von Medizinern geleistet, ist aber immer noch Bestandteil der Medizinethik. Er enthält moralische und ethische Aspekte hinsichtlich der Behandlung von Patienten. (vgl. Marschall, 2016) Insbesondere die ärztliche Schweigepflicht ist im Kontext dieser Arbeit hervorzuheben, da mit ihr Ärzte dazu verpflichtet werden, Informationen über Patienten in jedem Fall geheim zu halten. Sie ist auch in der sog. Deklaration von Genf des Weltärztebundes enthalten: „Ich werde die mir anvertrauten Geheimnisse auch über den Tod der Patientin oder des Patienten hinaus wahren.“ (Weltärztebund, 2017)

In der Gesetzgebung begann die Datenschutzthematik ab Ende des 18. Jahrhunderts Einzug zu halten. Zum einen wurde 1776 in Schweden das Gesetz zur Pressefreiheit verabschiedet, weiterhin wurde 1789 das Recht auf freie Meinungsäußerung in der Verfassung der USA etabliert. (vgl. Moos, Schefzig und Arning, 2018) Mit dem Zeitungsartikel „The Right to Privacy“[5] der Juristen Samuel D. Warren und Louis D. Brandeis begannen Ende des 19. Jahrhunderts in den USA erste Debatten um das Thema Datenschutz. Der Artikel zielte hauptsächlich auf die Presse ab und forderte unter anderem das Recht, „in Ruhe gelassen“ zu werden.[6] (vgl. ebd.) Einer der Autoren, Louis D. Brandeis, war der erste Richter am US Supreme Court[7], der ein verfassungsmäßiges Recht auf Privatsphäre auslegte. Zu seinen Ehren wird seit 2012 der Louis D. Brandeis

[5] `https://www.jstor.org/stable/pdf/1321160.pdf`, 1890
[6] „The right to be left alone“
[7] Oberster Gerichtshof der USA

Privacy Award an Führungspersönlichkeiten verliehen, die als Verfechter des Datenschutzes außergewöhnliches leisteten. Peter Schaar, ehemaliger Bundesdatenschutzbeauftragter, wurde 2014 mit diesem Preis gekührt. (vgl. Gropper und Peel, 2017)

In Deutschland führte die Weimarer Reichsverfassung im Jahr 1919 das Fernsprechgeheimnis ein und erlegte unter anderem bei dem Öffentlichmachen von Privatangelegenheiten hohe Strafen auf (vgl. Moos, Schefzig und Arning, 2018). Im Jahr 1948 wurde durch die Generalversammlung der Vereinten Nationen die Allgemeine Erklärung der Menschenrechte veröffentlicht, einer der Grundsteine des internationalen Menschenrechtsschutzes (vgl. Stefanovic, 2020). In Artikel 12, „Schutz der Freiheitssphäre des Einzelnen“ der Erklärung heißt es wie folgt:

> *„Niemand darf willkürlichen Eingriffen in sein Privatleben, seine Familie, sein Heim oder seinen Briefwechsel noch Angriffen auf seine Ehre und seinen Beruf ausgesetzt werden. Jeder Mensch hat Anspruch auf rechtlichen Schutz gegen derartige Eingriffe oder Anschläge.“* (ebd.)

Das erste Datenschutzgesetz in Deutschland wurde im Jahr 1970, mit dem Inkrafttreten des ersten Hessischen Datenschutzgesetzes, verabschiedet. Das Gesetz hatte neben der Stellung eines Datenschutzbeauftragten als Kontrollinstanz vor allem auch den Schutz elektronisch verarbeiteter Daten zum Ziel. Bis 1981 waren alle Bundesländer dem Beispiel Hessens gefolgt und hatten ebenfalls Landesdatenschutzgesetze beschlossen. (vgl. Bundeszentrale für politische Bildung, 2017) Das *Gesetz zum Schutz vor Missbrauch personenbezogener Daten bei der Datenverarbeitung* trat als erste Fassung des heutigen, deutschen Bundesdatenschutzgesetzes, im Jahr 1977 in Kraft. Bis zur jetzigen, angepassten Version, wurde es mehrfach überarbeitet bzw. novelliert. (vgl. ebd.)

Das Europäische Parlement erließ im Jahr 1995 die *Richtlinie zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten und zum freien Datenverkehr* (Richtlinie 95/46/EG) mit dem Ziel der Schaffung eines auf EU-Ebene einheitlichen Datenschutzniveaus (vgl. ebd.).

Mit der im Dezember 2000 unterzeichneten Charta der Grundrechte der Europäischen Union wurde Datenschutz auch als menschliches Grundrecht eingetragen (vgl. Moos, Schefzig und Arning, 2018).

> *„Jede Person hat das Recht auf Schutz der sie betreffenden personenbezogenen Daten.“* (Art. 8 Abs. 1 GRCh)

Aufgrund rascher technologischer Änderungen der zunehmenden Globalisierung wurde es erforderlich, die bisherige europäische Datenschutzrichtlinie abzulösen (vgl. Bundeszentrale für politische Bildung, 2017). Die über einen mehrjährigen Zeitraum entwickelte Datenschutzgrund-

verordnung löste diese daher mit Inkrafttreten im Mai 2016 ab und brachte europäische Länder auf ein einheitliches Datenschutzniveau. Seit dem Stichtag 25. Mai 2018 findet die Datenschutzgrundverordnung in der Europäischen Union Anwendung. (vgl. European Data Protection Supervisor, 2021)

Sollte nationales Recht mit der Datenschutzgrundverordnung in Konflikt stehen, hat die Verordnung stets Vorrang. Vor diesem Gesichtspunkt wurde auch das Bundesdatenschutzgesetz an diese angepasst und trat zum Tag der Anwendung der Datenschutzgrundverordnung zum 25. Mai 2018 in seiner derzeitigen Fassung in Kraft. (vgl. Moos, Schefzig und Arning, 2018)

2.1.2 Relevante Begriffe

Einige wichtige Begriffe, die im Rahmen dieser Arbeit, aber auch in den referenzierten Gesetzen Verwendung finden, werden im Folgenden kurz erläutert. In der Datenschutzgrundverordnung ist hierzu der Artikel 4, „Begriffsbestimmungen“, verankert, in dem viele Begriffe kurz beschrieben werden. Aus diesem Grund wird sich in der Reihenfolge der Erklärung nach diesem gerichtet.

Natürliche/Juristische Personen

Als Träger von Rechten und Pflichten wird ein Mensch mit der Geburt zu einer natürlichen Person. Als juristische Personen werden unter anderem Personenvereinigungen bezeichnet, die durch gesetzliche Anerkennung rechtsfähig sind, also ebenfalls Rechte und Pflichten besitzen. (vgl. Stobitzer, 2021)

Personenbezogene Daten

Gemäß Art. 4 DSGVO sind personenbezogene Daten Informationen, die einen Bezug zu einer identifizierbaren bzw. identifizierten natürlichen Person aufweisen. Beispiele für personenbezogene Daten sind allgemeine Personendaten wie Name, Geburtsdatum, E-Mail-Adresse oder Anschrift, aber auch Kennnummern oder Online-Daten wie die IP-Adresse oder Standortdaten (vgl. Ambros, 2021).

Betroffene Personen

Eine durch ihre Daten identifizierbare, natürliche Person wird im Kontext der Datenschutzgrundverordnung gemäß Art. 4 DSGVO als betroffene Person bezeichnet.

Datenverarbeitung

Unter einer Verarbeitung personenbezogener Daten wird nach Art. 4 DSGVO ein Vorgang bezeichnet, der sämtliche Prozesse von der Erhebung bzw. dem Erfassen bis zu deren Löschung beinhaltet.

Verantwortliche

Laut Art. 4 DSGVO sind Verantwortliche die Instanz, die über die Art und Weise der Datenverarbeitung personenbezogener Daten entscheidet und somit die Verantwortung hierfür trägt.

Auftragsverarbeiter

Nicht immer sind entsprechende Kompetenzen oder Ressourcen für eine Datenverarbeitung beim Verantwortlichen vorhanden. Aus diesem Grund können Daten gemäß Art. 4 DSGVO auch im Auftrag, von sogenannten Auftragsverarbeitern, verarbeitet und gespeichert werden.

Anonymität

Wenn eine Person nicht bzw. nicht mehr durch verarbeitete Daten identifiziert werden kann, findet die Verarbeitung anonym statt. Somit bilden anonyme Daten das Gegenstück zu personenbezogenen Daten. (vgl. Moos, Schefzig und Arning, 2018)

Pseudonymität

Laut Art. 4 DSGVO liegen Informationen in pseudonymer Form vor, wenn diese Informationen einer Person ohne Hinzuziehung weiterer Daten nicht zugeordnet werden können. Direkte Identifizierungsmerkmale werden von den restlichen Informationen getrennt und eine Identifizierung durch eine Zuordnungstabelle möglich. (vgl. ebd.)

Verschlüsselung

Von Verschlüsselung wird gesprochen, wenn Daten in einer Form übermittelt oder gespeichert werden, die nicht für Unberechtigte, sondern nur vom Schlüsselinhaber lesbar sind. Im Kontext der Datenschutzgrundverordnung ist Verschlüsselung eine Schutzmaßnahme, die ähnlich wie die Pseudonymisierung einzustufen ist. (vgl. ebd.)

Tracking

Als Tracking werden Technologien bezeichnet, die das Nutzerverhalten in mobilen Applikationen oder beim Besuch von Webseiten nachverfolgen, beispielsweise um gezielte Werbung zu schalten oder einen Nutzer webseiten- oder applikationsübergreifend zu identifizieren. Hierzu werden unter anderem sog. Cookies genutzt, kleine Textdateien, die im Browser gespeichert werden und eine Wiedererkennung bei späteren Webseitenbesuchen oder Nutzungen der Applikation ermöglichen. (vgl. Schallaböck, 2019)

2.1.3 Relevante Gesetze und Verordnungen

In dieser Arbeit werden Datenschutzerklärungen hinsichtlich ihrer Konformität zu geltenden Gesetzen und Verordnungen bewertet, welche im Folgenden kurz beschrieben werden.

Datenschutzgrundverordnung

Die Datenschutzgrundverordnung (DSGVO) dient dem Schutz der Grundrechte und Grundfreiheiten natürlicher Personen (vgl. Art. 1 DSGVO). Wie in Unterabschnitt 2.1.1 erwähnt, ist sie seit 25. Mai 2018 in Anwendung. Sie legt zahlreiche Grundsätze für die Verarbeitung personenbezogener Daten fest. Gemäß Art. 12 DSGVO muss die betroffene Person über die Art und Weise der Datenverarbeitung in nachvollziehbarer Form informiert werden. Die Zweckbindung in der Verarbeitung sichert gemäß Art. 5 DSGVO, dass erhobene Daten nur für den vorab festgelegten Zweck verarbeitet werden dürfen. Die Datenschutzgrundverordnung trifft viele weitere Auflagen, wie etwa zur Dauer der Speicherung oder zum Schutz der Daten vor Verlust, Zerstörung oder unberechtigter Verarbeitung. Einige, für diese Arbeit relevante Artikel werden im Folgenden kurz in ihre jeweiligen thematischen Kapitel zusammengefasst erläutert.

Kap. 1 - Allgemeine Bestimmungen (Art. 1 - 4)

In den Artikeln 1 bis 4 werden in der Datenschutzgrundverordnung sowohl ihr Anwendungsbereich als auch relevante Begrifflichkeiten definiert. Das erste Kapitel dient demnach primär der Eingrenzung und Erläuterung der Datenschutzgrundverordnung (vgl. Ingelheim und Fünkner, 2021)

Kap. 2 - Grundsätze (Art. 5 - 11)

Die gesamte Datenschutzgrundverordnung betreffende Grundsätze werden in den Artikeln 5 bis 11 beschrieben. Artikel 5 normiert die allgemeinen Grundsätze zum Umgang mit personenbezogenen Daten. Die Bedingungen der Rechtmäßigkeit einer Datenverarbeitung, beispielsweise der Einwilligung der Person, werden in den folgenden Artikeln definiert. Spezielle Regelungen für Minderjährige, aber auch besondere Kategorien personenbezogener Daten, wie etwa biometrische oder gesundheitliche Daten einer Person werden hier ebenfalls aufgestellt. (vgl. ebd.)

Kap. 3 - Rechte der betroffenen Personen (Art. 12 - 23)

Die betroffenen Personen durch die Datenschutzgrundverordnung zugesicherten Rechte werden in Kapitel 3 behandelt. Diese sind in fünf Abschnitte untergliedert. Im Bereich Transparenz gehören hierzu die Informationspflicht des Verantwortlichen und das Recht auf Auskunft von Betroffenen. Weiterhin zählen die Rechte auf Berichtigung bzw. Löschung von Daten und Einschränkung bzw. Widerspruchs der Verarbeitung und automatisierter Entscheidungsfindung zu diesem Bereich. Betroffene Personen müssen detailliert informiert werden, zu welchem Zweck und auf welcher Rechts-

grundlage personenbezogene Daten von ihnen erhoben und verarbeitet werden. (vgl. Ingelheim und Fünkner, 2021)

Kap. 4 - Verantwortlicher und Auftragsverarbeiter (Art. 24 - 43)

Neben der Sicherheit der Verarbeitung wird in Kapitel 4 Organisationen empfohlen, alle Unternehmensprodukte unter dem Gesichtspunkt der Datenschutzgrundverordnung zu betrachten und diese in Prozesse einzuarbeiten. Die von Verantwortlichen und Verarbeitern umzusetzenden Maßnahmen zur Datensicherheit werden ebenfalls behandelt, beispielsweise Schutzmaßnahmen, mit denen der Zugriff auf personenbezogene Daten geregelt wird. Bei Datenschutzverletzungen müssen betroffene Personen unverzüglich und die zuständige Aufsichtsbehörde binnen 72 Stunden nach Entdeckung des Vorfalls benachrichtigt werden. (vgl. Ogden, 2018)

Kap. 5 - Übermittlungen personenbezogener Daten an Drittländer [...] (Art. 44 - 50)

In Kapitel 5 werden die Voraussetzungen für die Übermittlung von personenbezogenen Daten in Drittländer oder an internationale Organisationen festgelegt. Als erster Grundsatz wird hier festgelegt, dass das durch die Datenschutzgrundverordnung geschaffte Schutzniveau hierdurch nicht untergraben werden darf. Als Legitimation der Datenübertragung gilt zum einen ein Angemessenheitsbeschluss der Europäischen Kommission, der dem Drittland oder der Organisation ein entsprechendes Datenschutzniveau attestiert. Weiterhin kann der Verantwortliche bzw. Verarbeiter im Drittland Garantien vorweisen, durch die die Übertragung legitimiert werden kann. (vgl. Ingelheim und Fünkner, 2021)

Bundesdatenschutzgesetz

Das Bundesdatenschutzgesetz (BDSG) wurde, wie in Unterabschnitt 2.1.1 erläutert, in seinem derzeitigen Stand auf die Datenschutzgrundverordnung hin angepasst und wird seit dem 25. Mai 2018 parallel mit ihr angewendet. Mit Öffnungsklauseln wurde der nationalen Gesetzgebung die Möglichkeit gegeben, Sachverhalte durch zusätzliche Gesetze zu konkretisieren, solange den Vorgaben der Datenschutzgrundverordnung nicht widersprochen wird. Daher ist das Bundesdatenschutzgesetz als eine Ergänzung zu selbiger und nicht als eigenständiges Gesetz zu betrachten, da es nur Themen enthält, die von der Datenschutzgrundverordnung explizit offen gelassen wurden. Diese belaufen sich im Wesentlichen auf Straf- und Bußgeldvorschriften, die Stellung des Datenschutzbeauftragten und den Arbeitnehmerdatenschutz. (vgl. Möllers, 2019a) In dieser Arbeit referenzierte Artikel werden im Folgenden beschrieben.

§ 64 BDSG - Anforderungen an die Sicherheit der Datenverarbeitung

Maßnahmen, die die Sicherheit der Verarbeitung personenbezogener Daten gewährleisten sollen, werden in § 64 BDSG behandelt. Dieser Paragraf gilt in Zusammenhang mit Art. 32 DSGVO, in welchem entsprechende technische und organisatorische Maß-

nahmen aufgeführt sind. Diese Liste wird von § 64 BDSG entsprechend erweitert, erwähnt unter anderem Zutritts-, Zugangs- und Zugriffskontrolle und zahlreiche weitere Kontrollmechanismen, die dem Schutz und der Sicherheit der Daten dienen. (vgl. Möllers, 2019b)

§ 71 BDSG - Datenschutz durch Technikgestaltung [...]

Der für die Datenverarbeitung Verantwortliche muss laut Festlegungen in § 71 BDSG bereits in der Konzeption der Verarbeitung geeignete Vorkehrungen treffen, Datenschutzgrundsätze durch diese Verarbeitung einzuhalten und sicherzustellen, dass die Rechte betroffener Personen eingehalten werden. Bei der Gestaltung von Verarbeitungssystemen soll das Ziel sein, so wenig personenbezogene Daten wie möglich zu verarbeiten und diese frühestmöglich zu anonymisieren oder pseudonymisieren, sofern möglich. Weiterhin muss der Verantwortliche sicherstellen, dass dass nur personenbezogene Daten verarbeitet werden können, die für den Verarbeitungszweck erforderlich sind. Dies folgt demnach dem Grundsatz der Datensparsamkeit. (vgl. § 71 BDSG)

Safe Harbor und Privacy Shield

Gemäß Art. 44ff. DSGVO werden an die Datenverarbeitung in Ländern außerhalb der Europäischen Union einige Anforderungen gestellt, die zwingend einzuhalten sind. Beispielsweise kann eine Übertragung personenbezogener Daten in ein Drittland legitim sein, wenn diesem von der Europäischen Kommission nach Art. 45 DSGVO ein angemessenes Schutzniveau attestiert wird.

Um die Übertragung von Daten in die USA zu legitimieren, gab es vor der Datenschutzgrundverordnung das sogenannte Safe Harbor Abkommen, welches im Jahr 2000 durch die Europäische Kommission in Kraft trat. Dieses wurde allerdings im September 2015 vom Europäischen Gerichtshof für ungültig erklärt, da US-Sicherheitsbehörden weitgreifenden Zugriff auf die Daten von EU-Bürgern hatten. (vgl. Hancock und Schmitz, 2020) Um die Datenverarbeitung in den USA wieder zu legitimieren, wurde hierzu am 12.06.2016 das EU-US-Privacy-Shield von der Europäischen Kommission verabschiedet. Durch Selbstzertifizierung der verarbeitenden Organisationen, die gegebenen Datenschutzgrundsätze einzuhalten, wurde die Übertragung und Verarbeitung auch im Rahmen der Datenschutzgrundverordnung zulässig. (vgl. Amtsblatt der Europäischen Union, 2016) Dieser Datenschutzschild wurde aber am 16. Juli 2020 vom Europäischen Gerichtshof ebenfalls für ungültig erklärt. Seither ist es nicht mehr zulässig, auf Basis des EU-US Privacy Shield Abkommens personenbezogene Daten in die USA zu übermitteln. Auf Seiten des Verantwortlichen besteht eine Prüfpflicht, ob der Verarbeiter in den USA hinreichende Garantien bietet, um die Übertragung auf Basis von Standardvertragsklauseln wieder zu legitimieren. (vgl. Gesellschaft für Datenschutz und Datensicherheit e.V., 2020) Die USA gelten nun im Rahmen der Datenschutzgrundverordnung als Drittland, weshalb Akteure ein identisches

Datenschutzniveau in der Übertragung und Verarbeitung nachweisen können müssen (vgl. Hancock und Schmitz, 2020).

Telemediengesetz

Das Telemediengesetz (TMG) wurde mit dem Ziel entwickelt, rechtliche Anforderungen elektronischer Informations- und Kommunikationsdienste zu regeln, behandelt unter anderem auch das Thema Datenschutz. Es wurde am 18.01.2007 verabschiedet. (vgl. Hoeren, 2007) Es wird in dieser Arbeit nicht detaillierter beschrieben, da es mangels Anpassung an die Datenschutzgrundverordnung und des Vorrangs selbiger nur noch in wenigen Paragrafen Relevanz besitzt (vgl. Datenschutzkonferenz, 2018). Diese werden im Rahmen des Konzepts an entsprechender Stelle erläutert. Zum 1. Dezember 2021 wird das Telemediengesetz von dem Telekommunikations-Telemedien-Datenschutzgesetz (TTDSG) abgelöst und richtet sowohl das Telemedien- als auch Telekommunikationsgesetz nach den Vorgaben der Datenschutzgrundverordnung aus (vgl. Jähn-Nguyen, 2021).

Nachdem im vergangenen Kapitel die geschichtliche Entwicklung des Datenschutzes sowie relevante Begriffe, Gesetze und Verordnungen erläutert wurden, wird im folgenden Abschnitt die Lesbarkeitsanalyse näher beschrieben.

2.2 Lesbarkeitsanalyse

Das Ziel einer Lesbarkeitsanalyse ist es, die Schwierigkeit des Lesens eines Textes festzustellen. Die Möglichkeit, diese zu bestimmen, versetzt Lehrkräfte in die Lage, für ihre Schüler geeignete Literatur auszuwählen, oder Autoren, ihre Texte ihrer Zielgruppe entsprechend verständlich auszurichten. (vgl. Zamanian und Heydari, 2012) Die möglichst einfache Lesbarkeit von Datenschutzerklärungen spielt in der Datenschutzgrundverordnung eine zentrale Rolle. Aus diesem Grund werden Methoden der Lesbarkeitsanalyse auch in dieser Arbeit genutzt.

Es werden einige Abschnitte aus der Forschung um die Lesbarkeit herausgegriffen, die zur Entwicklung aktuell genutzter Algorithmen führten. Einige gängige Verfahren zur Bestimmung der Lesbarkeit werden im Anschluss näher erläutert.

2.2.1 Geschichte

Einen Grundstein legte der amerikanische Psychologe Edward Thorndike 1921 mit der Veröffentlichung des Teachers' Work Book. Dieses war das erste Buch, das eine umfassende Liste englischer Worte nach Frequenz ihres Auftretens in allgemeiner Literatur bereitstellte. Er vertrat die Annahme, dass häufiger auftretende Worte aufgrund ihrer Geläufigkeit leichter zu verstehen sind, als seltener auftretende. Eine Formel zur Berechnung der Lesbarkeit wurde von ihm nicht entwickelt, Veröffentlichungen wie diese trugen aber zur Entwicklung selbiger bei. Die erste Les-

barkeitsformel wurde 1923 von Lively und Pressey entwickelt, um die Buchauswahl für die Junior Highschool zu erleichtern. Diese basierte direkt auf der Auftretensfrequenz des Wortes, ähnlich zu der Arbeit von Thorndike. Da keine Skala zur Auswertung des Ergebniswertes veröffentlicht wurde, fand diese Formel keine Nutzung, legte aber einen weiteren Grundstein für künftige Algorithmen. Die bekannteste Formel zur Lesbarkeitsbestimmung wurde von Rudolph Flesch entwickelt. In seiner Dissertation veröffentlichte er 1943 seine erste Formel zur Messung der Lesbarkeit eines Textes. Diese wurde zeitnah von Verlegern genutzt und führte zu einer deutlichen Steigerung der Leserschaft. Im Jahr 1948 veröffentlichte er eine weitere, vereinfachtere Version dieser Formel, die heute unter dem Namen „Flesch Reading Ease Readability Formula“ bekannt ist. (vgl. DuBay, 2004) Diese wird immer noch genutzt und ist beispielsweise in Microsoft Word implementiert (vgl. Zamanian und Heydari, 2012). Die Formel wird in Unterabschnitt 2.2.2 näher beschrieben.

Auf Basis der Entwicklung von Flesch veröffentlichten Dale und Chall im gleichen Jahr die „Dale-Chall Readability Formula“ mit dem Ziel, Defizite seiner Formel auszugleichen (vgl. ebd.). Bis zu Beginn der 80er Jahre wurden auf Basis dieser Entwicklungen bereits 200 Lesbarkeitsalgorithmen für die englische Sprache entwickelt (vgl. DuBay, 2004). Zwei bekannte Indizes bilden der „Gunning Fog Index“ (1952) und die „Flesch-Kincaid Readability Formula“ (1975). (vgl. Zamanian und Heydari, 2012) Beide werden ebenfalls im Folgenden Unterabschnitt 2.2.2 erläutert.

2.2.2 Gängige Verfahren

Im englischsprachigen Raum gibt es, wie in Unterabschnitt 2.2.1 benannt, mehr als 200 Algorithmen zur Bestimmung der Lesbarkeit von Texten. Die bekanntesten, noch heute in Verwendung befindlichen Verfahren werden in diesem Abschnitt beschrieben.

Flesch Reading Ease Readability Formula

Die Reading Ease Formel von Rudolf Flesch ist eine auf englischsprachige Texte ausgelegte Funktion zur Berechnung der Lesbarkeit eines Textes. Sie gibt die Schwierigkeit in einem Wertebereich 1 - 100 aus, wobei 30 und geringer als „sehr schwer“ und alles über 70 als „leicht“ verständlich bedeutet. Die Berechnung erfolgt mit folgender Formel:

$$Score = 206,835 - (1,015 \cdot ASL) - (84,6 \cdot ASW)$$

Wobei *ASL* für die durchschnittliche Satzlänge (engl. Average Sentence Length) und *ASW* für die durchschnittliche Anzahl von Silben je Wort (engl. Average number of Syllables per Word) steht. (vgl. DuBay, 2004)

Dale-Chall Readability Formula

Die von Edgar Dale und Jeanne Chall entwickelte Lesbarkeitsformel baut auf der Entwicklung von Flesch auf, stützt sich aber nicht nur auf Durchschnittswerte. Dale und Chall entwickelten eine Liste („long list“) einfacher Worte, von denen 80 Prozent Viertklässlern geläufig sind.

$$Score = 0,1579 \cdot PDW + 0,496 \cdot ASL + 3,6365$$

Alle nicht auf dieser Liste vorkommenden Worte werden als schwierige Worte definiert und gehen anhand ihres Prozentanteils im Text als Wert *PDW* (engl. Percentage of Difficult Words) in die Berechnung ein. Als zweiter Faktor wird die durchschnittliche Satzlänge, ähnlich zu der Flesch-Formel, als Wert *ASL* genutzt. Im Gegensatz zur Flesch Formel wird hier auf einen engeren Wertebereich gesetzt. Ergebnisse mit einem Wert von 4,9 oder kleiner werden als für Viertklässler angemessen gewertet, Werte größer gleich 10 für College-Absolventen. (vgl. DuBay, 2004)

Gunning Fog Index

Robert Gunning veröffentlichte seinen auf Erwachsene ausgerichteten Fog Index 1952. Sie basiert wie die Flesch- und Dale-Chall-Formeln auf der durchschnittlichen Satzlänge, fügt als weitere Komponente aber den Prozentanteil schwieriger Worte (engl. hard words), seiner Definition nach Worte mit mehr als zwei Silben, hinzu. Ihr in Schuljahren angegebener Wertebereich reicht von 6, sechste Klasse, bis zu 17, College-Absolventen. (vgl. ebd.)

$$Grade\,Level = 0,4 \cdot (ASL + hard\,words)$$

Flesch-Kincaid Readability Formula

Die Flesch-Kincaid Formel wurde im Jahr 1976 im Auftrag der U.S. Navy von Rudolf Flesch und Peter Kincaid entwickelt und gibt auf Basis der Flesch-Formel die Anzahl von Schuljahren aus, die zum Verständnis des Textes, im konkreten Fall Einsatzhandbücher und ähnliche Materialien der Navy, nötig sind. (vgl. ebd.)

$$Grade\,Level = (0,39 \cdot ASL) + (11,8 \cdot ASW) - 15,59$$

Die Formel ist nicht nur eine Umrechnung der Flesch-Skala auf den Wertebereich in Schuljahren, sondern benutzt eine weiterentwickelte Berechnung. Diese basiert zwar ebenfalls auf der durchschnittlichen Satzlänge und Anzahl von Silben je Wort, allerdings wird die Wortlänge deutlich höher gewichtet. (vgl. Fries und Keimig, 2019)

Wiener Sachtextformel

Für deutschsprachige Texte gibt es sehr wenige Lesbarkeitsindizes. Zu den bekannteren gehören die Wiener Sachtextformeln, bzw. im Besonderen die 4. Wiener Sachtextformel als am häufigsten genutzte der Formelsammlung. Ähnlich zur Flesch-Kincaid Lesbarkeitsformel gibt auch diese die Anzahl von Schuljahren aus, die für das Textverständnis erforderlich sind. Sie basiert in ihrer Berechnung auf der mittleren Satzlänge (SL) und dem Anteil an Worten mit mehr als zwei Silben.

$$WSF = 0,2656 \cdot SL + 0,2744 \cdot MS - 1,693$$

Der Wertebereich des Ergebnisses bewegt sich zwischen 4 (sehr leicht/4. Klasse) und 15, wobei Werte größer 12 als sehr schwer verständlich zu bezeichnen sind. (vgl. ebd.)

Für diese Arbeit relevante Bausteine der Lesbarkeitsanalyse wurden in diesem Abschnitt beschrieben. Im folgenden Abschnitt werden benötigte für das Konzept notwendige Grundlagen zur Textanalyse erläutert.

2.3 Textanalyse

In diesem Abschnitt wird auf die Grundlagen der in dieser Arbeit verwendeten Gebiete der Textanalyse näher eingegangen. Zunächst wird ein kurzer Überblick über die geschichtliche Entwicklung derselben gegeben, im Anschluss daran werden Computerlinguistik und Text Mining näher erläutert. Relevante Begriffe aus dem Gebiet der Textanalyse werden abschließend definiert. Die benannten Technologien werden im Rahmen der automatisierten Analyse von Datenschutzerklärungen genutzt, um Texte in ihre Bestandteile zu zerlegen und zu durchsuchen.

2.3.1 Geschichte

Die Computertechnologie der dreißiger und vierziger Jahre des vergangenen Jahrhunderts war vorrangig auf die Lösung numerischer Probleme ausgerichtet. Das hohe Potenzial der Technologie wurde aber auch zu dieser Zeit schon erkannt, unter anderem bei der Dechiffrierung verschlüsselter Funksprüche und Nachrichten im zweiten Weltkrieg. Aufgrund der geringen Leistungsfähigkeit damaliger Hardware, waren die meisten Ansätze nur auf rein symbolischer Basis vorhanden und wurden durch sehr vereinfachte Modelle demonstriert. In den achtziger Jahren entwickelte sich die Erkennung gesprochener Sprache deutlich. Einen entscheidenden Beitrag hierzu stellten Korpusse von Sprachdaten, auch Training bezeichnet, dar. Diese Trainingsverfahren ermöglichten auch die Erkennung der Sprache mit größerem Wortschatz und verschiedenen Sprechern. Mit Hilfe syntaktischer und semantischer Analyse war es zu dieser Zeit erstmals möglich, Wortarten automatisiert zu bestimmen. Verschiedene Herangehensweisen, statistische und stochastische Verfahren, wurden kombiniert. Dies führte zu ersten maschinellen

Übersetzern von Text in verschiedene Sprachen. Mit steigender Verbreitung des Internets und der damit einhergehenden stark anwachsenden Menge verfügbarer Informationen, war auch in der Computerlinguistik ein schneller Anstieg von Forschungsaktivitäten zu verzeichnen. Gerade in Form von Diktieranwendungen fand die automatische Spracherkennung Anfang der neunziger Jahre hohe Verbreitung. Die steigende Leistungsfähigkeit der Hardware, aber auch durch die zunehmende Entwicklung künstlicher Intelligenz prägten das Forschungsfeld der Computerlinguistik. (vgl. Carstensen, 2010) Dieses wird im folgenden Abschnitt näher erläutert.

2.3.2 Computerlinguistik und Text Mining

Das Forschungsfeld Computerlinguistik (engl. Natural Language Processing, NLP) beschäftigt sich mit der Verarbeitung natürlicher Sprache. Es ist eine Kombination mehrerer Forschungsmethoden, beispielsweise nicht nur der Informatik, sondern auch Aspekten der Philosophie. Letztere sind insbesondere bei der Frage nach der Verbindung zwischen Sprache, Denken und Handeln relevant. Die Informatik bietet notwendige Datenstrukturen und Algorithmen zur Lösung der Phänomene natürlicher Sprache. (vgl. ebd.)

In engem Zusammenhang mit der Computerlinguistik steht Text Mining, ein Teilgebiet aus dem Bereich Data Mining. Data Mining umfasst Techniken zur Analyse großer, strukturierter Datenbestände. Text Mining wird zur Extraktion von Wissen aus Texten genutzt. Mit Hilfe dieser Technik können neue, bislang unbekannte Informationen aus Texten extrahiert werden. Ein Teilbereich des Text Minings ist die Informationsextraktion. Diese hat zum Ziel, einzelne Informationen aus einem Text herauszulösen und zur Verfügung zu stellen. Dies kann zum einen der Prüfung auf Relevanz, aber auch der Analyse von Zusammenhängen oder Kategorisierung von Texten dienen. Einige Verfahren der Computerlinguistik, beispielsweise statistische Sprachverarbeitung, werden im Text Mining angewandt. Auch künstliche Intelligenz findet in diesem Bereich Anwendung, unter anderem bei der Suche nach relevanten Texten oder der Kategorisierung des Inhalts. (vgl. Witte und Mülle, 2006)

Im Rahmen der Textanalyse sind einige Teilgebiete der Computerlinguistik, Morphologie, Syntax und Semantik, von Relevanz. Diese sollen im folgenden Abschnitt näher beleuchtet werden. Relevante Begriffe aus dem Bereich Text Mining werden dort ebenfalls definiert.

2.3.3 Relevante Begriffe

Einige Begriffe aus den Gebieten Computerlinguistik und Text Mining werden in diesem Abschnitt erläutert. Zunächst wird auf Grundlagen der Linguistik eingegangen, dann einige Verfahren aus der Computerlinguistik vorgestellt. Die Reihenfolge richtet sich primär nach dem Werk von Witte und Mülle, 2006, die in ihrer Arbeit viele Grundlagen des Text Minings näher beleuchten.

Morphologie

Die Morphologie beschäftigt sich mit der Zusammensetzung von Worten aus Morphemen, den kleinsten Einheiten einer Sprache, denen Bedeutung zugemessen wird. Morpheme werden in Stämme und Affixe[8] eingeteilt. (vgl. ebd.)

Syntax

Die Syntax einer Sprache dient ihrer Strukturbildung. Sie enthält neben Wortarten auch deren Zusammenfassung in Wortgruppen und ihre Abhängigkeiten. (vgl. ebd.)

Semantik

Die einer sprachlichen Äußerung beigemessene Bedeutung wird als Semantik bezeichnet. Sie behandelt die Bedeutung einzelner Worte, Sätze und Abschnitte. Morphologie und Syntax ermöglichen die Ableitung der Struktur und das Treffen einer inhaltlichen Aussage. Verständnis für die Bedeutung wird aber erst im Rahmen einer Analyse der Semantik erhalten. (vgl. ebd.)

Stoppworte

Worte, die nicht problemlösend wirken bzw. dem Text keinen Inhalt bringen, werden als Stoppworte bezeichnet. Zu diesen zählen unter anderem Artikel, Konjunktionen oder Hilfsverben. (vgl. ebd.)

Tokenisierung

Mithilfe der Tokenisierung wird ein Text in sprachlich relevante Einheiten unterteilt und so für eine weitere Verarbeitung vorbereitet. Einheiten (engl. Tokens) können aus Worten, Phrasen oder Sätzen bestehen. (vgl. ebd.)

Stemming

Stemming führt ein Wort auf seinen Wortstamm[9] zurück. Der Stamm muss kein existierendes Wort der Sprache sein, sondern hängt von dem verwendeten Algorithmus ab. (vgl. ebd.)

Lemmatisierung

Das sogenannte Lemma, die Grundform[10] eines Wortes, wird durch die Lemmatisierung erzeugt. Während sie im Englischen vorwiegend aus dem Streichen von Affixen besteht, wird im Deutschen aufgrund komplexer, morphologischer Regeln auf lexikon-basierte Ansätze zurückgegriffen, da sich bei der Lemmatisierung hier auch der Wortstamm selbst ändern kann.[11] (vgl. ebd.)

[8] Beispiel: geh-en
[9] Beispiel Stemming: lachte → *lach*
[10] Beispiel Lemmatisierung: lachte → *lachen*
[11] Beispiel Lemmatisierung mit Wortstammänderung: Mäuse → *Maus*

Klassifikation

Mithilfe der Untersuchung von Wortmustern und vorhandener Themen eines Textes kann er unterteilt und in definierte Gruppen eingeordnet werden. Die Themengebiete werden vordefiniert und durch mehrere Bezeichner im Dokument gefunden. (vgl. Witte und Mülle, 2006)

Machine Learning

Unter maschinellem Lernen, einem Teilgebiet künstlicher Intelligenz, wird die Generierung künstlichen Wissens aus Erfahrung verstanden. Es kann genutzt werden, um beispielsweise relevante Daten in einem Datenbestand zu identifizieren und zusammenzufassen. Das Machine Learning System besteht aus einem Basis-Satz an Algorithmen und kann mit einem bestehenden Datensatz auf einen speziellen Einsatzzweck trainiert werden und findet eigenständig Lösungen für gegebene Probleme. (vgl. Luber und Litzel, 2016)

Entropie

Der Informationsgehalt einer Nachricht wird in Form der Entropie berechnet. Sie bezeichnet die Wahrscheinlichkeit des Auftretens eines Textvorkommens. Worte mit einer niedrigen Auftretenswahrscheinlichkeit tragen nach der Informationstheorie mehr Information in sich als häufig vorkommende. Als Messwert zeigt die Entropie den Durchschnittswert der Wahrscheinlichkeiten. (vgl. Brownlee, 2019)

Reguläre Ausdrücke

Reguläre Ausdrücke (engl. regular expressions bzw. RegEx) ermöglichen es, Zeichenfolgen aufgrund von Zeichenketten mit syntaktischen Regeln zu beschreiben. Durch normale Zeichen oder deren Kombination mit Sonderzeichen (sog. Metazeichen) können Zeichenkonstellationen gefunden werden. (vgl. Behrens, 2019)

Auf Basis der dargestellten Grundlagen wird im folgenden Kapitel ein Konzept für die automatisierte Analyse von Datenschutzerklärungen erarbeitet.

3 Konzeptionelle Überlegungen

Dieses Kapitel zeigt die mit der Entwicklung des Programms zur automatisierten Analyse von Datenschutzerklärungen einhergehenden Vorüberlegungen. Hierzu werden zunächst Anforderungen und Zielsetzungen erläutert. Die für die Analyse zu nutzenden Testkriterien werden im Anschluss behandelt. Abschließend wird der konzipierte Programmablauf dargestellt und auf die zu verwendenden Technologien eingegangen.

3.1 Anforderungen

Ziel dieser Arbeit ist die Entwicklung eines Programms, mit dem die automatisierte Analyse von Datenschutzerklärungen in deutscher und englischer Sprache ermöglicht wird. Die Untersuchung soll aufzeigen, ob angenommen werden kann, dass die jeweilige Datenschutzerklärung konform zur Datenschutzgrundverordnung und weiteren deutschen Gesetzen ist und über die relevanten Bestandteile informiert. Neben der inhaltlichen Analyse soll auch die Lesbarkeit bewertet werden, sodass nachvollziehbar ist, ob die Datenschutzerklärung verständlich formuliert wurde. Auch wenn die Motivation primär auf die Analyse von Datenschutzerklärungen mobiler Applikationen abzielt, soll das Programm auch die Untersuchung von Datenschutzerklärungen ohne den Bezug zu einer Applikation, wie etwa Datenschutzerklärungen von Webseiten, mit wenigen Anpassungen ermöglichen. Eine vollständige, rechtssichere Prüfung kann mangels juristischer Ausbildung nicht geleistet werden, allerdings sollen Datenschutzerklärungen in der automatisierten Analyse aufgrund festgelegter Kriterien umfassend überprüft und fehlende Informationen und potenzielle Mängel aufgedeckt werden. Es soll ein Bewertungsmaßstab erstellt werden, mit dessen Hilfe analysierte Datenschutzerklärungen verglichen werden können, beispielsweise im Rahmen eines Vergleichs zweier mobiler Applikationen identischer Kategorie. Die Einrichtung des Analyseprogramms soll mit wenigen Schritten möglich sein, weiterhin soll es an andere Softwaresysteme angebunden werden können, um diese gegebenenfalls um die Analyse von Datenschutzerklärung ergänzen zu können. Ein modularer Aufbau soll die zukünftige Erweiterbarkeit sicherstellen.

3.2 Testkriterien

Die für diese Arbeit relevanten Artikel der Datenschutzgrundverordnung wurden bereits im Unterabschnitt 2.1.3 näher beschrieben, auch wurden die ergänzend herangezogenen Paragrafen aus dem Bundesdatenschutz- und Telemediengesetz erläutert. Auf Basis der benannten Verordnung und Gesetze werden in diesem Abschnitt Testkriterien mit dem Ziel entwickelt, diese automatisiert beantworten zu können. Die Testkriterien werden in statistisch ermittelbare und inhaltliche Kriterien eingeteilt. Statistische Kriterien werden in Unterabschnitt 3.2.1 näher beleuchtet. Sie behandeln vornehmlich die in Art. 12 DSGVO getroffenen, formalen Festlegungen in Bezug auf

die transparente Information der betroffenen Person. Inhaltlich abzuprüfende Kriterien werden im Unterabschnitt 3.2.2 erläutert. Mit diesen soll geprüft werden, ob die Datenschutzerklärung über wesentliche bzw. erforderliche Punkte informiert.

3.2.1 Statistik

Ob ein Text „in präziser, transparenter, verständlicher und leicht zugänglicher Form in einer klaren und einfachen Sprache“ (Art. 12 DSGVO) formuliert wurde, kann von Individuum zu Individuum unterschiedlich aufgefasst werden. In Bezug auf die automatisierte Analyse einer Datenschutzerklärung muss daher eine Methode gefunden werden, die eine neutrale Bewertung ermöglicht. Aus diesem Grund werden die folgenden drei Prüfkriterien genutzt.

Struktur

Im Prüfpunkt *Struktur* wird das Verhältnis zwischen der Anzahl der Überschriften und der ihnen durchschnittlich zugehörigen Anzahl an Worten ermittelt. Somit ist eine Bewertung der Übersichtlichkeit des Textes möglich und daher auch seiner Zugänglichkeit und Präzision.

Entropie

Die Inhalte der Datenschutzerklärung sollen laut Datenschutzgrundverordnung einfach und präzise sein, ein langer Text mit geringem Informationswert spricht daher gegen diese Auflage. Unter dem Punkt *Entropie* wird daher der Informationsgehalt des Textes berechnet. Um die Genauigkeit in Bezug auf den geforderten Inhalt von Datenschutzerklärungen zu erhöhen, wird der ermittelte Wert mit der Gesamtwertung der im Unterabschnitt 3.2.2 erläuterten Kriterien in Relation gesetzt.

Lesbarkeit

Abschließend wird ein Messwert für die *Lesbarkeit* des Textes gebildet, sodass nachvollziehbar ist, ob die Datenschutzerklärung zielgruppenorientiert erstellt wurde und somit gemäß Art. 12 DSGVO in klarer, einfacher Sprache formuliert ist. Als Zielgruppe wird in diesem Kontext jeder angesehen, der die allgemeine Schulpflicht von 10 Jahren erfüllt. Diese Schwelle wird gewählt, da Jugendliche ab dem vollendeten 16. Lebensjahr in Deutschland vor dem Datenschutz nicht mehr als Kinder gelten (vgl. Milkaite und Lievens, 2019).

3.2.2 Inhalt

Um zu überprüfen, ob die Datenschutzerklärung die laut Datenschutzgrundverordnung erforderlichen Informationen enthält, werden folgende inhaltliche Prüfkriterien genutzt. Diese enthalten auch weitere, mit Hilfe von Punkten des Bundesdatenschutz- und Telemediengesetzes ergänzte Bestandteile. Die Kriterien sind in mehrere Kategorien aufgeteilt, um im Rahmen der Bewertung

Themenkomplexe unterschiedlich gewichten oder gar ausklammern zu können, beispielsweise falls eine Datenschutzerklärung ohne zugehörige mobile Applikation analysiert werden soll.

In der Kategorie *Allgemein* wird geprüft, ob die Datenschutzerklärung übliche bzw. verpflichtende, allgemeine Informationen enthält. Hierzu zählt unter anderem die Erwähnung eines Änderungs- oder Gültigkeitsdatums. Ob die mobile Applikation und ihre Berechtigungen in der Datenschutzerklärung Erwähnung finden, wird in der Kategorie *Mobile Applikation* bearbeitet. Das Teilen von Daten mit Dritten und ob ein Zweck hierzu angegeben ist, wird im Testbereich *Dritte* beantwortet. Wie die aufgezeichneten Daten behandelt werden, beispielsweise hinsichtlich Anonymisierung oder Verschlüsselung, wird im Bereich *Datenbehandlung* erfasst.

Allgemein

Üblicherweise enthalten Datenschutzerklärungen ein Gültigkeits- bzw. Änderungsdatum. Dies ist laut Datenschutzgrundverordnung nicht explizit vorgeschrieben, ermöglicht dennoch der betroffenen Person, Änderungen in der Datenschutzerklärung leichter festzustellen. Auch eine Versionierung des Textes, also die Möglichkeit der Einsichtnahme in frühere Stände des Dokuments, stärkt die Transparenz gegenüber der betroffenen Person. Aus diesem Grund wird im Bereich *Allgemein* auf diese Kriterien geprüft. Nach der Datenschutzgrundverordnung bzw. dem Bundesdatenschutzgesetz erforderliche, allgemeine Angaben wie die Kontaktdaten des Datenschutzbeauftragten und Informationen zum Rechtemanagement zählen ebenfalls zu diesem Bereich.

Änderungsdatum

Damit der Benutzer erkennen kann, ob sich seit dem letzten Lesen in der Datenschutzerklärung etwas geändert hat, wird darauf geprüft, ob ein Änderungs- bzw. Gültigkeitsdatum in der Datenschutzerklärung vermerkt ist.

Versionierung

Auch wenn ein Änderungsdatum in der Datenschutzerklärung vermerkt ist, ist nicht zwingend ersichtlich, an welchen Stellen sich die Datenschutzerklärung geändert hat. Da die Anzeige früherer Versionen oder einer Änderungshistorie einen besseren Überblick für die betroffene Person ermöglicht, wird hierauf ebenfalls geprüft.

Kontaktdaten

Laut Art. 37 DSGVO muss ein Unternehmen mit hauptsächlicher Verarbeitungstätigkeit einen Datenschutzbeauftragten benennen und die Kontaktdaten veröffentlichen. Auch müssen dem Benutzer im Rahmen der Informationspflicht (Art. 13 DSGVO) Kontaktdaten des für die Verarbeitung Verantwortlichen zur Verfügung gestellt werden. In diesem Punkt wird nach entsprechenden Informationen in der Datenschutzerklärung gesucht.

Rechtemanagement

Nach § 64 BDSG muss bei automatisierten Verarbeitungssystemen eine Zugriffskontrolle vorhanden sein, anhand der sichergestellt ist, dass Berechtigte ausschließlich Zugriff auf personenbezogene Daten besitzen, für die sie auch die Berechtigung besitzen. Ob entsprechende Maßnahmen in der Datenschutzerklärung erwähnt werden, wird in diesem Punkt untersucht.

Mobile Applikation

Zu Beginn des Nutzungsvorgangs muss der Benutzer laut § 13 TMG über Art, Umfang, Zweck und Verwendung der Datenaufzeichnung informiert werden. Auch Art. 13 DSGVO benennt dies als Teil der Informationspflichten des Verantwortlichen. Aus diesem Grund wird an dieser Stelle geprüft, ob die mobile Applikation in der Datenschutzerklärung erwähnt wird. Weiterhin wird ermittelt, ob auch Berechtigungen (beispielsweise Standort/GPS bzw. Kamera) benannt sind.

Erwähnung

Sofern die Datenschutzerklärung zu einer mobilen Applikation gehört und diese eine Datenverarbeitung vornimmt, muss diese auch im Rahmen der Datenschutzerklärung benannt werden. In der Analyse wird aus diesem Grund geprüft, ob in der Datenschutzerklärung die mobile Applikation Erwähnung findet.

Berechtigungen

Durch das Erteilen von Berechtigungen können mobile Applikationen personenbezogene Daten aufzeichnen. Als Teil der Informationspflichten des Verantwortlichen nach Art. 13 DSGVO , aber auch nach § 13 TMG, muss darüber informiert werden, welcher Zweck mit der Datenaufzeichnung verfolgt wird.

Dritte

Ob personenbezogene Daten durch den Verantwortlichen mit Partnern bzw. Dritten geteilt werden und ob hierzu ein Zweck vermerkt ist, wird aus der Datenschutzerklärung ausgelesen. Insbesondere durch Art. 13 DSGVO müssen Verantwortliche im Rahmen der Informationspflichten über diesen Sachverhalt informieren. Weiterhin wird ermittelt, ob sog. Tracker benannt werden, die das Nutzerverhalten zu verschiedensten Zwecken analysieren.

Teilen von Daten

Die Empfänger bzw. Kategorien von Empfängern personenbezogener Daten müssen dem Nutzer gemäß Art. 13 DSGVO mitgeteilt werden. Aus diesem Grund wird geprüft, ob die Datenschutzerklärung über das Teilen von Daten informiert.

Zweck des Teilens

Der Zweck der mit der Verarbeitung, auch durch Dritte, einhergeht, muss gemäß der in Art. 13 DSGVO benannten Informationspflichten zu Beginn der Verarbeitung mitgeteilt werden. Daher wird geprüft, ob im Zusammenhang mit dem Teilen von Daten auch der Zweck dessen benannt wird.

Tracker

Der Einsatz von Tracking-Mechanismen, die eine Analyse des Nutzerverhaltens vornehmen, wird nicht explizit durch die Datenschutzgrundverordnung geregelt. Daher fällt diese Form der Verarbeitung ebenfalls unter die generellen Informationspflichten des Verantwortlichen gemäß Art. 13 DSGVO. In der Konferenz der unabhängigen Datenschutzbehörden des Bundes und der Länder vom 26. April 2018 wurde die Anwendbarkeit nationalen Rechts diskutiert und ein Dokument „Zur Anwendbarkeit des TMG für nicht-öffentliche Stellen ab dem 25. Mai 2018“ veröffentlicht. In Bezug auf die Nutzung von Tracking-Technologien wird an dieser Stelle darauf hingewiesen, dass eine informierte Einwilligung vor der Datenverarbeitung, bzw. bevor Cookies gesetzt werden, eingeholt werden muss, da relevante (erlaubende) Paragrafen des Telemediengesetzes nicht mehr anwendbar sind. (vgl. Datenschutzkonferenz, 2018) Aus diesem Grund wird geprüft, ob in der Datenschutzerklärung sog. Tracker erwähnt werden. Ein Abgleich mit der ggf. vorhandenen Applikation erfolgt an dieser Stelle nicht, kann aber als Teil einer späteren Erweiterung erfolgen.

Datenbehandlung

Die Informationspflichten des Verantwortlichen nach Art. 13 DSGVO sind umfangreich. Die Testkriterien im Bereich Datenbehandlung decken diese in den relevanten Punkten ab. Es wird überprüft, ob notwendige Informationen wie der Verarbeitungszweck und die Nutzerrechte angegeben wurden. Zur Datenbehandlung zählen weiterhin, ob Daten anonymisiert bzw. pseudonymisiert werden und ob Verschlüsselung angewendet wird. Ob der Standort der Datenverarbeitung angegeben wurde und ob bei der Datenverarbeitung das Aufzeichnen mancher Daten optional ist, wird ebenfalls abgeprüft.

Zweck der Erhebung

Der mit der Datenverarbeitung verfolgte Zweck muss der betroffenen Person durch die Informationspflichten des Verantwortlichen (Art. 13 DSGVO) mitgeteilt werden. Aus diesem Grund wird geprüft, ob im Rahmen der Datenerhebung auch ein Zweck benannt ist.

Anonymisierung / Pseudonymisierung

Sofern personenbezogene Daten für die Verarbeitung nicht (mehr) erforderlich sind, sollten diese anonymisiert oder pseudonymisiert werden. (Art. 32 DSGVO i.V.m. § 71 BDSG) Daher wird überprüft, ob im Zusammenhang mit der Datenverarbeitung auch diese Themen in der Datenschutzerklärung Erwähnung finden.

Verschlüsselung

Nach Art. 32 DSGVO i.V.m. § 64 BDSG sollen personenbezogene Daten nach einer durch den Verantwortlichen zu treffenden Risikoeinschätzung möglichst in verschlüsselter Form übertragen und gespeichert werden, sofern dies in der Verarbeitung möglich ist. Es wird aus diesem Grund geprüft, ob im Text der Datenschutzerklärung in Zusammenhang mit der Verarbeitung auch eine Verschlüsselung vorkommt.

DSGVO-Rechte

In der Datenschutzgrundverordnung werden den betroffenen Personen in den Artikeln 12 ff. DSGVO umfangreiche Rechte zugesichert. Mit diesen kann beispielsweise Einsicht in die gespeicherten, personenbezogenen Daten genommen werden. Ob in Art. 12 ff. DSGVO benannte Rechte, wie Auskunft, Berichtigung, Löschung, Einschränkung oder Widerspruch erwähnt werden, wird in diesem Punkt geprüft.

Speicherdauer

In Art. 13 DSGVO ist vermerkt, dass bei Erhebung darüber zu informieren ist, wie lange personenbezogene Daten gespeichert werden oder auf welchen Kriterien die Speicherdauer beruht. Daher wird geprüft, ob die Datenschutzerklärung Informationen über die Dauer der Speicherung enthält.

Standort der Datenverarbeitung

Die Artikel 44 - 50 DSGVO treffen umfangreiche Festlegungen für die Datenübertragung in Drittländer. Es wird aus diesem Grund daraufhin geprüft, ob vermerkt ist, in welchem Land die Datenverarbeitung stattfindet, sodass der Benutzer im Zweifel prüfen kann, ob dieses ein angemessenes Datenschutzniveau bietet.

Optionale Daten

Datenverarbeitungen, die nicht primär der Vertragserfüllung, sondern anderen Zwecken dienen, können entweder nach Art. 6 DSGVO aufgrund berechtigter Interessen des Verantwortlichen stattfinden oder aber durch Einwilligung des Benutzers nach Art. 7 DSGVO. Ob in der Datenschutzerklärung Daten erwähnt werden, die erst nach Einwilligung verarbeitet werden und somit optionaler Natur sind, wird in diesem Punkt überprüft.

3.3 Kategorisierung

Um die inhaltlichen Testkriterien beantworten zu können, müssen Rückschlüsse auf den Inhalt der Datenschutzerklärung ermöglicht werden. In Abschnitt 1.2 erwähnte andere Ansätze gehen zum Teil den Weg umfangreich trainierter Machine Learning Modelle. Der Aufwand des Trainings eines eigenen Modells kann in dieser Arbeit nicht geleistet werden. Weiterhin soll die Funktionsweise für deutsch- und englischsprachige Datenschutzerklärungen gleichwertig sein, was

mit zwangsläufig auf unterschiedliche Sprachen trainierten Machine Learning Modellen nicht nachvollziehbar belegt werden könnte. Aus beiden Gründen wird im Rahmen dieser Arbeit auf eine statische Kategorisierung auf Basis einer Stichwortsuche zurückgegriffen. Hierbei werden Textabschnitte durch Auftreten bestimmter Stichworte zu einer Kategorie zugeordnet. Über eine Kombination verschiedener Kategorien eines Absatzes können so Rückschlüsse auf dessen Inhalt gezogen werden. Das in Abschnitt 1.2 benannte Frontierproject teilte Inhalte von Datenschutzerklärungen in sechs Kategorien ein[12]: *COLLECT* (Sammeln), *SHARE* (Teilen), *USE* (Nutzung), *SECURITY* (Sicherheit), *ACCESS* (Zugriff) und *CHOICES* (Wahlmöglichkeiten) ein. Auf Basis dieser Kategorisierung wird in dieser Arbeit aufgebaut und die erkennbaren Kategorien wie in Tabelle 3.1 dargestellt, erweitert, sodass die inhaltlichen Testkriterien aus einer Kombination dieser Kategorien beantwortet werden können.

Kategorie	**Beschreibung**
access	Zugriff auf personenbezogene Daten
anonym	Anonymisierung bzw. Pseudonymisierung
app	Erwähnung einer Applikation
app-permissions	Berechtigungen einer Applikation
authorized	Organisatorische Maßnahmen, nur authorisierter Datenzugriff
choices	Auswahlmöglichkeiten bzw. optionale Daten
collect	Sammeln personenbezogener Daten
contact	Kontaktdaten des Datenschutzbeauftragten
countries	Ländernamen
duration	Speicherdauer
encryption	Verschlüsselung
gdpr-rights	Nutzerrechte nach der Datenschutzgrundverordnung
gpdr	Datenschutzgrundverordnung
location	Standort
personal-data	Personenbezogene Daten
privacy-shield	EU-US Privacy-Shield bzw. Safe Harbor
purpose	Zweck der Datenverarbeitung
security	Sicherheit bzw. Schutz von Daten
share	Teilen von Daten
social	Soziale Netzwerke
third-parties	Drittparteien oder verbundene Unternehmen
use	Nutzung bzw. Verarbeitung von Daten
versioning	Angaben zur Versionierung der Datenschutzerklärung

Tabelle 3.1: Erweiterte Kategorien zur Beantwortung inhaltlicher Testkriterien

3.4 Beschreibung und Ablauf der Analyse

Im Folgenden wird aufgeschlüsselt, wie die entwickelte Analyse ablaufen soll und in welche Schritte diese aufgeteilt ist. Der konzipierte Ablauf ist in Abbildung 3.1 entsprechend dargestellt und wird anhand der in dieser dargestellten Schritte erläutert.

[12] `https://github.com/adityamarella/frontierproject/blob/master/classify.py#L21`

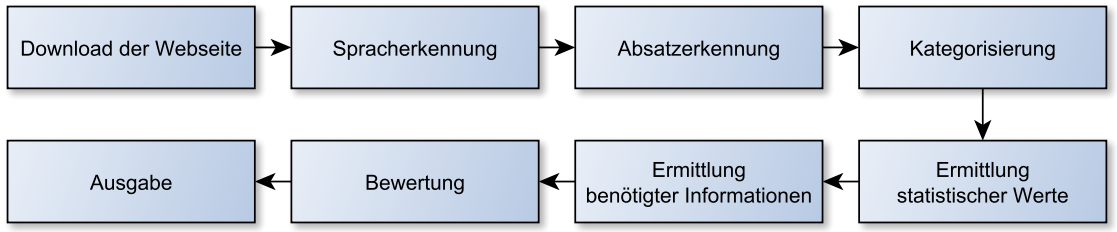

Abbildung 3.1: Konzipierter Programmablauf

Über eine Programmierschnittstelle (API) wird der Link der Datenschutzerklärung an das Programm gesendet. Im Anschluss an die Analyse soll dann das Resultat in einem üblichen, strukturierten Datenformat ausgegeben werden.

Zunächst wird die Datenschutzerklärung über den durch die API bereitgestellten Link heruntergeladen und für die weitere Analyse vor- bzw. aufbereitet. Die Sprache des Textes wird erkannt und der Text anhand seiner Überschriften in Absätze unterteilt. Die erkannten Absätze werden auf Basis des Inhalts kategorisiert. Nach der Ermittlung der für die statistischen und inhaltlichen Prüfkriterien (siehe Abschnitt 3.2) notwendigen Informationen wird aus diesen eine Bewertung gebildet und die Informationen gesammelt über die API ausgegeben.

Der konzipierte Ablauf der Analyse wird in den wichtigsten Punkten folgend näher beschrieben.

Download

Der Download der Datenschutzerklärung erfolgt nach der Bereitstellung der URL automatisiert. Es wird nach Abschluss des Downloads geprüft, ob dieser erfolgreich war und es sich bei den heruntergeladenen Daten um HTML-Quelltext handelt. Sollte dies nicht der Fall sein, ist von einem Fehler auszugehen. Die Analyse wird in diesem Fall mit konkreter Fehlermeldung abgebrochen.

In den Artikeln 44 - 50 DSGVO wird festgelegt, ob und wie die Datenübertragung in Drittländer erfolgen darf. Bei dem Besuch von Webseiten, wie einer Datenschutzerklärung, kann unwissentlich mit Servern außerhalb der Europäischen Union kommuniziert werden, weiterhin wird hierbei zwangsläufig die eigene, öffentliche IP-Adresse an den Server gesendet. Diese ist laut Urteil des Bundesgerichtshofs (Urt. v. 16.05.2017, Az. VI ZR 135/13) als persönliches Datum anzusehen. Aus diesem Grund wird der Serverstandort der Datenschutzerklärung ermittelt.

Für die weitere Analyse wird der Text der Datenschutzerklärung extrahiert und über diesen eine Prüfsumme (engl. Hash) gebildet, anhand der überprüfbar ist, ob bereits eine Analyse für eine Datenschutzerklärung vorliegt. Sollte dies der Fall sein, werden die Resultate der bekannten Analyse ausgegeben.

Spracherkennung

Die Sprache des Textes wird herausgefunden, um in der weiteren Analyse gegebenenfalls andere Programmabschnitte zu aktivieren, beispielsweise bei der Auswahl des zu nutzenden Algorithmus für die Lesbarkeitsanalyse.

Absatzerkennung

Um im Text verschiedene Themenkomplexe erkennen zu können, wird der Inhalt in Abschnitte unterteilt. Es bietet sich hierfür an, die HTML-Formatierung der Webseite zu nutzen, sodass Überschriften als Trenner des Textes genutzt werden und ihnen auf diese Weise der Textabschnitt bzw. Absatz zugeordnet wird. Sollte keine Formatierung erkennbar sein, soll die Analyse abgebrochen werden.

Kategorisierung

In dem nächsten Schritt werden die gefundenen Absätze mit Hilfe einer Kategorisierung (siehe Abschnitt 3.3) entsprechend der erkannten Thematik gekennzeichnet. Dies dient im späteren Verlauf auch der Beantwortung der inhaltlichen Testkriterien. Um die Beantwortung der Kriterien manuell überprüfen zu können, wird auf diese Weise nachvollziehbar gemacht, ob der Absatz der korrekten Thematik zugeordnet wurde. Sollte eine über die automatisierte Analyse hinausgehende, manuelle Prüfung der Datenschutzerklärung erfolgen, wird durch die Kategorisierung auch der hierfür notwendige Aufwand reduziert, da gegebenenfalls nur themenrelevante Absätze gelesen werden müssen.

Ermittlung statistischer Werte

Im Anschluss an die Kategorisierung der Absätze werden die statistischen Kriterien ermittelt. Wie bereits in Unterabschnitt 2.2.2 erläutert, sind für deutsch- und englischsprachige Texte unterschiedliche Algorithmen zur Lesbarkeitsanalyse verfügbar. Anhand der vorher erkannten Sprache der Datenschutzerklärung wird ein für diese geeigneter Algorithmus berechnet. Für die Entropie- und Strukturwertung der Datenschutzerklärung werden in diesem Schritt die benötigten Werte ermittelt. Dies beinhaltet unter anderem die Anzahl von Überschriften und Worten.

Ermittlung benötigter Informationen

Wie bereits beschrieben, wird die Beantwortung der inhaltlichen Testkriterien hauptsächlich auf Basis einer Stichwortsuche erfolgen. Diese wird für einige Testkriterien kombiniert, beispielsweise zur Erkennung, ob ein Verarbeitungszweck in Zusammenhang mit der Verarbeitung durch Dritte benannt wird. Da im Rahmen der Kategorisierung bereits die Themen der Absätze aufgenommen werden, wird die Bewertung der Kriterien im folgenden Schritt durch Kombination der zugeordneten Kategorien möglich.

Daten, die nicht Teil der Kategorisierung sind, wie etwa das Gültigkeitsdatum oder ob Kontaktdaten zum Datenschutzbeauftragten vermerkt sind, sollen in diesem Schritt mit Hilfe regulärer Ausdrücke in Erfahrung gebracht werden.

Bewertung

Die beantworteten Kriterien werden bewertet und fließen in eine Gesamtwertung ein. Als maximal erreichbare Punktzahl werden 100 Punkte genutzt, da dies als genauerer Indikator für die Wertung der Datenschutzerklärung dient, als dies zum Beispiel mit Schulnoten der Fall wäre. Im Rahmen der Umsetzung wird auf Basis der Relevanz der Kriterien eine Gewichtung der einzelnen Kriterien erarbeitet, die die Vergleichbarkeit der Gesamtwertung zwischen verschiedenen Datenschutzerklärungen sicherstellt.

Ausgabe und Speicherung

Um eine spätere Identifikation der erfolgten Analyse vornehmen zu können, wird eine eindeutige Kennzeichnung in Form einer sog. UUID[13] für das Analyseprojekt erstellt. UUIDs sind ein Standard im Bereich der Identifikationsnummern, mit dem als weltweit eindeutig geltende Zeichenfolgen generiert werden können (vgl. Augsten, 2019). Als weiteres Identifikationsmerkmal wird ein Projektname definiert und gespeichert. Sollte dem Programm kein entsprechender Name übergeben worden sein, soll dieser auf Basis des Titels und der Domain der Webseite gebildet werden.

Für die Speicherung werden bewertungsrelevante Daten in einem JSON-Objekt erfasst, sodass für sich ändernde oder hinzukommende Prüfkriterien das Datenmodell nicht verändert werden muss. Neben dem extrahierten Text der Datenschutzerklärung wird auch der HTML-Quelltext in der Datenbank gespeichert, um beispielsweise bei Änderungen am Bewertungsverfahren auf Basis dieser Daten eine neuerliche Analyse vornehmen zu können.

Der Ablauf der Analyse wurde anhand der wichtigsten Schritte vom Herunterladen bis zur Ausgabe der erfassten Daten in diesem Abschnitt näher erläutert. Technologische Voraussetzungen werden im folgenden Abschnitt ausführlich beschrieben.

3.5 Technologien

In dem folgenden Abschnitt wird auf die zu verwendenden Technologien näher eingegangen. Zunächst wird die Basis bzw. Grundlage des Programms erläutert. Die für den Download und die Verarbeitung der HTML-Daten zu nutzenden Module werden im Anschluss hieran definiert. Für die Textanalyse notwendige Bibliotheken werden abschließend näher beleuchtet.

[13] Universally Unique Identifier

3.5.1 Basis des Programms

Den Unterbau bzw. die Grundlage eines Verarbeitungsprogramms bilden unter anderem die Programmiersprache, die Datenspeicherung und die Benutzerschnittstelle. Wenn diese Komponenten feststehen, kann darauf aufbauend das Programm gestaltet und für das aufgestellte Konzept geeignete Technologien gefunden werden. Im Folgenden wird erläutert, auf welcher Basis die Programmierung erfolgt und welche Module für den Unterbau aus welchem Grund genutzt werden.

Programmiersprache

Als Programmiersprache wird dem Programm Python 3[14] zugrunde liegen. Neben Plattformunabhängigkeit, einer umfangreichen Dokumentation und einer Vielzahl integrierter Bibliotheken verfügt Python über einen integrierten Paketmanager, mit dem bei Bedarf Module nachgeladen werden können.[15]

Plattform

Python ist eine plattformunabhängige Programmiersprache, daher ist die Wahl der Plattform irrelevant. Um die Installation weitestgehend einfach zu halten, soll es allerdings ermöglicht werden, die Anwendung innerhalb eines Docker[16] Containers zu starten, sodass gegebenenfalls zu installierende Pakete/Module automatisch bereitgestellt werden. Docker kann Programme in voneinander unabhängigen Umgebungen (sog. Containern) bereitstellen. Prozesse und Anwendungen bleiben voneinander getrennt, weiterhin kann vorhandene Infrastruktur effizienter und sicherer genutzt werden. (vgl. Red Hat Inc., 2021)

Schnittstelle

Um das Programm an andere Softwaresysteme anbinden zu können, wird eine API geschaffen. Damit die Anwendung auch manuell bedient werden kann, ist neben einer Dokumentation der API auch nötig, sie ohne Programmierkenntnisse steuern zu können. Weiterhin soll der Aufwand der Implementierung in andere Softwareumgebungen möglichst gering sein. Ein Standard für eine programmiersprachenunabhängige Beschreibung einer HTTP API ist OpenAPI[17]. Dieser ermöglicht es, sowohl Menschen als auch Computern, die Fähigkeiten eines Dienstes in Erfahrung zu bringen, ohne dass Zugriff auf den Quellcode des Programms erforderlich ist. (vgl. OpenAPI Initiative, 2021)

Der OpenAPI-Standard basiert auf JSON, einem textbasierten Datenaustauschformat, das zwar programmiersprachenunabhängig, aber trotzdem von vielen Programmiersprachen als interne Datenstruktur umgesetzt werden kann (vgl. Ecma International, 2017). Aus diesem Grund bietet es sich auch an, das JSON-Format als Ausgabe für die Analyse-

[14] `https://www.python.org/`
[15] Bspw. `https://pypi.org/`
[16] `https://www.docker.com/`
[17] `https://github.com/OAI/OpenAPI-Specification`

ergebnisse des Programms zu nutzen. Swagger[18], eine bekannte, OpenAPI-kompatible Dokumentationssoftware, stellt neben einer Dokumentation für Mensch und Maschine auch Möglichkeiten bereit, um die API und somit das Programm direkt anzusteuern.

In der Python-Programmierung gibt es verschiedene Bibliotheken, mit denen eine eigene HTTP API aufgebaut werden kann. Zu den bekanntesten zählen Flask[19] oder Django[20], allerdings sind beide Bibliotheken sehr umfangreich und erfordern für eine OpenAPI-Dokumentation dennoch das Einarbeiten in Zusatzpakete. Eine gute Alternative hierzu ist `FastAPI`[21]. Neben kürzerer Entwicklungszeit und höherer Performance bietet FastAPI eine direkte Integration der Swagger-Oberfläche. Weiterhin können übergebene Parameter direkt validiert werden. (vgl. Nafies, 2020) Aus diesem Grund wird `FastAPI` für den Aufbau der Schnittstelle genutzt.

Datenbank

Für viele Datenbanklösungen wie MySQL, PostgreSQL oder Microsoft SQL gibt es herunterladbare Bibliotheken, um diese aus einem Python Programm ansprechen zu können. Die benannten Datenbanklösungen erfordern eine Server-Software bzw. einen laufenden Dienst und teilweise auch die Installation zusätzlicher Client-Software, damit Anfragen an die Datenbank gestellt werden können. (vgl. Makai, 2021) SQLite wird als einziger Datenbanktyp ohne Nachinstallation von Modulen unterstützt und benötigt weder zusätzliche Software, noch einen laufenden Dienst. Sobald die `sqlite3`[22] Bibliothek eingebunden ist, kann die Datenbank angesprochen werden. Weiterhin werden alle Daten in einer Datenbankdatei gespeichert. (vgl. Tao, 2020) Da sich der Aufwand für die Implementierung von SQLite auf ein Minimum beschränkt, wird diese als Datenbanklösung genutzt.

Konfiguration

Damit für Änderungen an Datenpfaden oder weiteren notwendige Einstellungen, wie zum Beispiel des Netzwerk-Ports, unter dem die API erreichbar sein soll, nicht im Programmcode angepasst werden müssen, wird die Konfiguration als separate Datei abgelegt. Hierfür gibt es verschiedene Möglichkeiten, beispielsweise als INI- oder XML-Datei. Ein weiterer Ansatz ist die Speicherung im YAML-Format. Ähnlich zu HTML ist YAML eine Auszeichnungssprache, dient also der Strukturierung und Anzeige von Daten. Im Gegensatz zu vergleichbaren Standards wie JSON dient YAML primär der Speicherung programminterner Datenstrukturen in einer für Menschen lesbaren Form. Die Strukturierung wird ähnlich zu Python selbst durch Einrückungen erstellt. (vgl. Augsten, 2017) Auch im Docker-Umfeld[23] wird auf dieses Format zur Speicherung von Konfigurationsdateien zurückgegriffen. Aus

[18] `https://swagger.io/`
[19] `https://flask.palletsprojects.com/en/2.0.x/`
[20] `https://www.djangoproject.com/`
[21] `https://fastapi.tiangolo.com/`
[22] `https://docs.python.org/3/library/sqlite3.html`
[23] `https://docs.docker.com/compose/`

genannten Gründen wird das YAML-Format für die Speicherung der Konfiguration im Rahmen des Programms genutzt. Das Auslesen und Schreiben von YAML-Dateien ist in Python mit der Bibliothek `pyyaml`[24] möglich.

3.5.2 Download und HTML-Verarbeitung

Die für das Herunterladen und die Vorbereitung bzw. Verarbeitung des HTML-Quelltextes der Datenschutzerklärung zu nutzenden Technologien werden im Folgenden dargestellt.

Download

Für das Herunterladen von Webseiten und Dateien können in Python verschiedene Bibliotheken genutzt werden. Neben dem integrierten `urllib` Modul, sind zahlreiche weitere verfügbar. In der Python Dokumentation zu `urllib` selbst wird das Modul `requests`[25] empfohlen[26]. `Requests` ist eine schlanke, sehr einfach einzurichtende HTTP Bibliothek für Python (vgl. Reitz, 2021). Daher wird diese Bibliothek dem Download der Datenschutzerklärungen dienen.

HTML-Parser

Damit die heruntergeladenen HTML-Daten durchsucht und verarbeitet werden können, ist ein HTML-Parser notwendig. Dieser dient der Umwandlung des HTML-Formats in einen programmtechnisch durchsuchbaren Datensatz. Hierfür ist in Python der `html.parser`[27] integriert. Mit der Bibliothek `beautifulsoup4`[28] steht allerdings ein Modul zur Verfügung, das das Durchsuchen der HTML-Daten weiter vereinfacht. Es baut auf Parsern wie `html.parser` auf, hält die Komplexität für die Implementierung aber auf einem niedrigen Niveau und ermöglicht ebenfalls eine einfache Navigation durch das Dokument. (vgl. Palakollu, 2019) Daher wird dieses Modul für die HTML-Verarbeitung genutzt.

Extrahierung des Textes

HTML-Parser wie `beautifulsoup4` können zwar den gesamten Text einer Webseite ausgeben, allerdings kann dieser für die Analyse nicht notwendigen bzw. unerwünschten Inhalt, beispielsweise der Kopf- und Fußzeile einer Webseite, enthalten.

Für die Extraktion relevanten Texts gibt es eine große Modulauswahl[29]. Die Qualität der verfügbaren Bibliotheken wurde in einem unabhängigen Test verglichen[30]. Das Modul mit

[24] https://pypi.org/project/PyYAML/
[25] https://pypi.org/project/requests/
[26] https://docs.python.org/3/library/urllib.request.html#module-urllib.request
[27] https://docs.python.org/3/library/html.parser.html
[28] https://pypi.org/project/beautifulsoup4/
[29] https://github.com/adbar/trafilatura#evaluation-and-alternatives
[30] https://github.com/currentsapi/extractnet#performance

der höchsten Präzision[31] ist hierbei `trafilatura`[32] und wird im Rahmen dieser Analyse zur Extraktion des relevanten Textes der Datenschutzerklärung implementiert.

3.5.3 Textanalyse

Alle mit der Textanalyse zusammenhängenden, zu nutzenden Technologien werden in den folgenden Punkten näher erläutert.

Spracherkennung

Um die Sprache eines Textes zu erkennen, können in Python verschiedene Herangehensweise und Bibliotheken genutzt werden. Im Rahmen eines Benchmark Tests[33] wurden verschiedene Module miteinander verglichen. Hierbei schnitt `fasttext`[34] von Facebook sowohl im Bereich Geschwindigkeit als auch Genauigkeit am besten ab. Die Bibliothek benötigt allerdings allein ein 126MB großes Machine Learning Modell für die Erkennung der Sprache (vgl. Facebook Inc., 2021). Das Modul `langdetect`[35] weist die zweithöchste Genauigkeit auf, benötigt aber keine zusätzlichen Daten, um die Spracherkennung vornehmen zu können (vgl. Lee, 2020). Aus diesem Grund wird dieses für die Erkennung der Sprache der Datenschutzerklärung genutzt.

Lesbarkeitsanalyse

Gängige Algorithmen zur Lesbarkeitsanalyse wurden im Unterabschnitt 2.2.2 vorgestellt. In der Analyse wird auf das „Flesch-Kincaid Grade Level“ für englisch- und auf die „Wiener Sachtextformel“ für deutschsprachige Datenschutzerklärungen zurückgegriffen, da diese zu den üblichsten Algorithmen gehören und trotz leicht abweichender Herangehensweisen in der Berechnung einen identischen Wertebereich von zum Textverständnis notwendigen Bildungs- bzw. Schuljahren angeben.

Das Python Modul `textstat`[36] kann gängige Lesbarkeitsindizes berechnen und kommt ohne Installation weiterer Zusatzmodule aus. Aus diesem Grund wird es für die Berechnung des „Flesch-Kincaid Grade Level“ genutzt.

Die Berechnung der „Wiener Sachtextformel“ ist in `textstat` nicht implementiert, daher wird auf `wstf`[37] zurückgegriffen, eine nicht im Paketmanager gelistete Python Bibliothek.

[31] Stand 21.06.2021
[32] `https://pypi.org/project/trafilatura/`
[33] `https://towardsdatascience.com/benchmarking-language-detection-for-nlp-8250ea8b67c`
[34] `https://fasttext.cc/`
[35] `https://pypi.org/project/langdetect/`
[36] `https://pypi.org/project/textstat/`
[37] `https://github.com/pablotheissen/wstf`

Absatzerkennung

Das in Abschnitt 1.2 benannte *Frontierproject* enthält Ansätze, die auch in dieser Arbeit Verwendung finden. Einer dieser Ansätze ist die Absatzerkennung der `Sectioner`[38] Bibliothek des Projekts. Diese findet anhand einer Stichwortsuche themenrelevante Überschriften und ordnet den jeweiligen, zugehörigen Absatz zu. Da die Bibliothek nur auf englischsprachige Begriffe hin prüft, wird die Bibliothek für diese Arbeit dahingehend angepasst, dass auch deutschsprachige Datenschutzerklärungen verarbeitet werden können.

Kategorisierung

Ein weiterer Ansatz aus dem *Frontierproject* ist die `Classifier` Bibliothek. Mit Hilfe einer Stichwortsuche kann diese einem gegebenen Text Kategorien (engl. *Tags*) zuordnen. Ähnlich zu der `Sectioner` Bibliothek ist die Klasse nur auf englischsprachige Texte ausgelegt und muss dementsprechend angepasst werden. Im Rahmen der Kategorisierung wird nach gleichem Schema auch auf gegebenenfalls enthaltene Ländernamen geprüft, um den Standort der Verarbeitung in Erfahrung zu bringen. Das Python Modul `pycountry`[39] soll hierzu die Einträge der ISO-Datenbanken, beispielsweise Ländernamen und -kürzel (ISO 3166[40]), in englischer und deutscher Sprache bereitstellen.

Erkennnung von Sätzen und Extraktion weiterer Daten

Im Bereich des Natural Language Processing gibt es für Python verschiedene Bibliotheken, die verschiedene Ansätze verfolgen. Das Natural Language Toolkit (`NLTK`[41]) ist beispielsweise auf Forschungszwecke ausgerichtet und gibt dem Entwickler zahlreiche Möglichkeiten der Anpassung auf seine Zwecke. Im Gegensatz hierzu steht `spaCy`[42], eine auf den Produktiveinsatz ausgelegte Bibliothek, die für jede Sprache die am besten geeigneten Algorithmen vorgibt und dadurch die Einrichtung vereinfacht. (vgl. The Data Incubator, 2016) Im Rahmen dieser Arbeit wird `spaCy` für die Erkennung von Sätzen genutzt. In der programmtechnischen Umsetzung wird eruiert, inwiefern `spaCy` weiterhin geeignet ist, zusätzliche Daten zu extrahieren.

Um gegebenenfalls in der Datenschutzerklärung erwähnte Tracker herauszufinden, ist ein Datensatz mit offiziellen Bezeichnungen, nach denen gesucht werden kann, erforderlich. In der Analyse mobiler Applikationen kann der Dienst von Exodus Privacy[43] über 400 Tracker identifizieren. (Stand: 02.06.2021) Die Datenbank des Dienstes steht öffentlich zur Verfügung und wird in diesem Projekt dazu genutzt, die Namen von in der Datenschutzerklärung erwähnten Trackern herauszufinden.

[38] `https://github.com/adityamarella/frontierproject/blob/master/sectioner.py`
[39] `https://pypi.org/project/pycountry/`
[40] `https://www.iso.org/iso-3166-country-codes.html`
[41] `https://www.nltk.org/`
[42] `https://pypi.org/project/spacy/`
[43] `https://exodus-privacy.eu.org/en/`

Der konzipierte Ablauf der Analyse wurde in diesem Abschnitt um die notwendigen Technologien ergänzt. Neben der Basis bzw. Grundlage des Programms wurde erläutert, welche Programmmodule aus welchem Grund genutzt wurden.

3.6 Massenanalyse

Um die entwickelte Methode zu evaluieren, wird ein Testset an Datenschutzerklärungen ausgewählt, für die eine automatisierte Analyse vorgenommen wird. Das Testset soll verschiedenste Datenschutzerklärungen enthalten, sodass nach der Massenanalyse sichergestellt ist, dass das entwickelte Konzept nicht nur auf einen Themen- oder Produktbereich anwendbar ist.

Nach der automatisierten Analyse wird eine manuelle Bewertung mit Hilfe der Testkriterien (siehe Abschnitt 3.2) an den Datenschutzerklärungen vorgenommen, die im Rahmen der automatisierten Analyse am besten und am schlechtesten abschnitten. Im Anschluss werden die automatisiert und manuell ermittelten Ergebnisse entsprechend verglichen.

4 Programmtechnische Umsetzung

Auf den Grundlagen (siehe Kapitel 2) und der Konzeption (siehe Kapitel 3) aufbauend wird in diesem Kapitel die Umsetzung der Programmierung beschrieben. Zunächst wird auf den Aufbau des Programms eingegangen, im Anschluss hieran die entwickelte Datenbankstruktur beschrieben. Der entwickelte Bewertungsmaßstab und das Format der Ausgabewerte wird im Anschluss hieran erläutert. Abschließend werden einige Herausforderungen in der Implementierung aufgezeigt und der Ablauf der automatisierten Analyse anhand eines Beispiels erläutert.

4.1 Programmaufbau

Auf Basis der im Konzept erläuterten Technologien (siehe Abschnitt 3.5) und dem konzipierten Ablauf der Analyse (siehe Abschnitt 3.4) wurden die im Folgenden näher erläuterten Klassen erstellt. Diese werden in Abbildung 4.1 entsprechend gezeigt. In der Abbildung wurde zur besseren Übersicht auf die Darstellung der `Config`- und `Logger`-Klassen verzichtet, da diese in nahezu allen Klassen integriert sind.

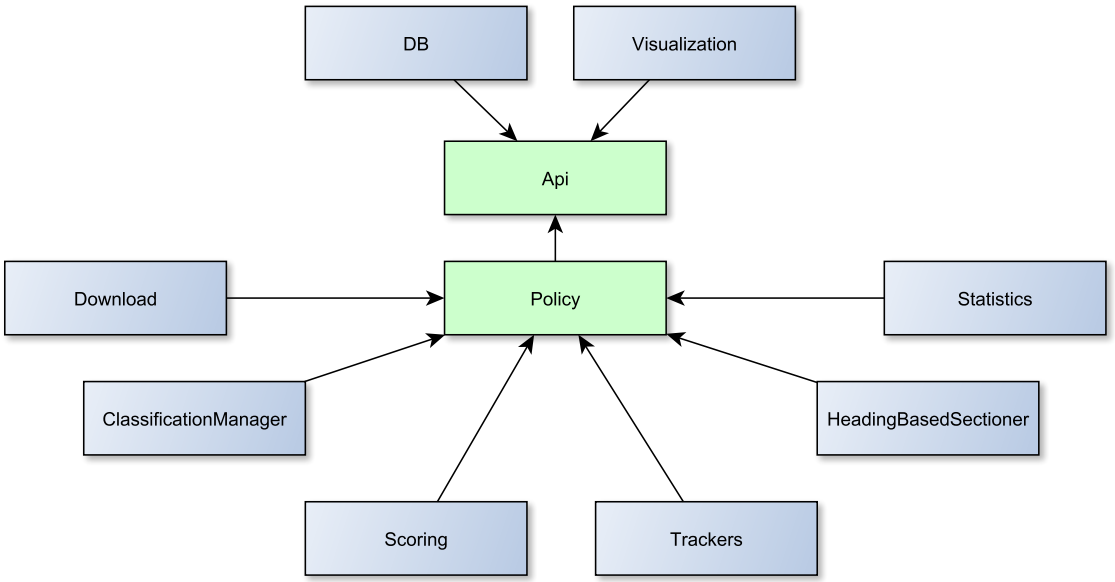

Abbildung 4.1: Übersicht der funktionalen Klassen

Api

Als zentraler Bestandteil bildet die `Api`-Klasse die Benutzer- und Programmschnittstelle. Über sie werden neue Analysen gestartet und Ergebnisse abgerufen. Durch die Implementierung der `FastAPI` Bibliothek ist die Dokumentation der API direkt integriert und mit einem Webbrowser aufrufbar (siehe Abbildung 4.2). Weiterhin kann die API über diese Webseite auch direkt gesteuert werden, ohne dass zusätzliche Software notwendig ist, wie in Unterabschnitt 3.5.1 erläutert. Die `Api`-Klasse bindet aufgrund ihrer zentralen Position im Programm viele Klassen ein.

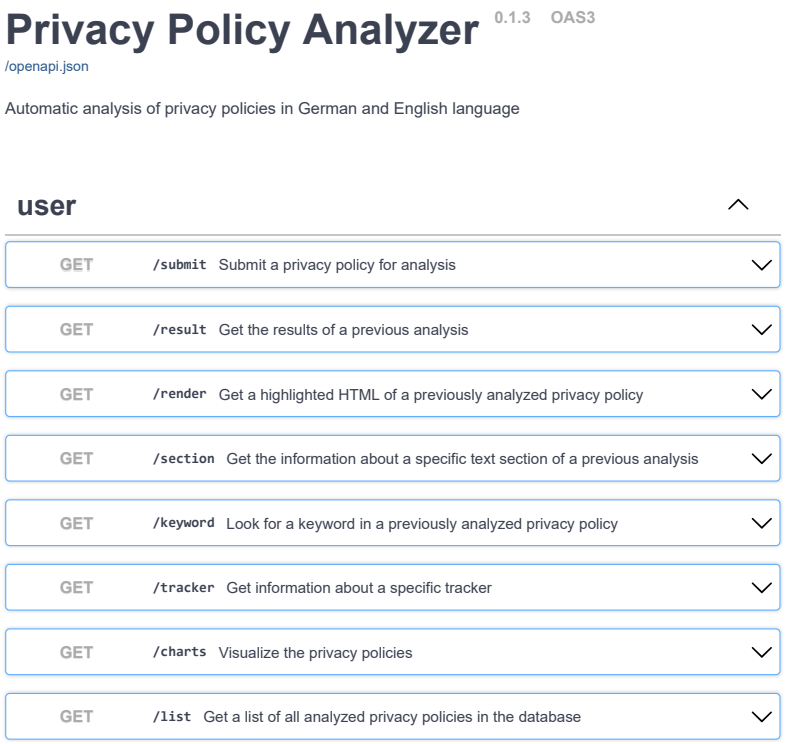

Abbildung 4.2: Integrierte Swagger API-Dokumentation

Config

Mithilfe des in Unterabschnitt 3.5.1 beschriebenen Python Moduls `pyyaml` wird der Inhalt der Konfigurationsdatei (siehe Quelltext 4.1) im YAML-Format gespeichert und über die `Config`-Klasse dem Programm bereitgestellt. Die Konfiguration beinhaltet neben den zu nutzenden Verzeichnispfaden oder des API-Ports auch weitere, die Analyse betreffende Parameter, wie die in der Kategorisierung zu nutzenden Schlüsselworten.

DB

Mit der `DB`-Klasse wird im Programm die Datenbankverbindung hergestellt. Bei der ersten Verbindung (siehe Quelltext 4.2) werden die Konfigurationsparameter aus der Konfigurationsdatei ausgelesen, die Datenbank verbunden und gegebenenfalls die Datenbankstruktur (siehe Abschnitt 4.2) erzeugt. Wie in Unterabschnitt 3.5.1 erwähnt, wird für das Projekt primär SQLite als Datenbanklösung genutzt, allerdings wurde die Datenbank-Klasse so gestaltet, dass andere Datenbanksysteme integriert werden können, sofern erforderlich.

```
root: C:\Privacy-Policy-Analyzer
data: C:\Privacy-Policy-Analyzer\data

projects: $data/projects
tmpFolder: $data/tmp
useragent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (
    KHTML, like Gecko) Chrome/89.0.4389.105 Safari/537.36

api:
  ip: 0.0.0.0
  port: 8000

db:
  type: SQLite
  config:
    path: $data/privacy.db
```

Quelltext 4.1: Ausschnitt aus der Konfigurationsdatei des Programms

```
def getDBConnected() -> DB:
    """
    Factory method that returns connected database.
    """
    config = Config.get_config()
    db = getDB(config["db"]["type"])
    db.connect(config["db"]["config"])
    if db.db_is_empty():
        db.createtables()
    return db
```

Quelltext 4.2: Funktion zum Herstellen einer Datenbankverbindung

Logger

Die `Logger`-Klasse stellt für alle Programmbestandteile eine zentrale Protokollierung (engl. Logging) bereit. Die Ausführlichkeit der Protokollierung (engl. log level) ist konfigurierbar, sodass während der Entwicklung detailliertere Meldungen ausgegeben werden als im Produktivbetrieb.

Download

In der `Download`-Klasse sind Methoden gesammelt, die das Herunterladen der übergebenen URL sowie die Verarbeitung bzw. Vorbereitung des hierbei heruntergeladenen HTML-Quelltextes übernehmen, wie in Unterabschnitt 3.5.2 ausgearbeitet. Die zu akzeptierenden Sprachen werden beim Download mit dem Parameter `Accept-Language` auf Deutsch und Englisch eingestellt. Dies ist in Quelltext 4.3 entsprechend dargestellt. Weiterhin wird zwecks Simulation eines echten Web-Browsers der Parameter `User-Agent` gemäß Konfiguration auf einen aktuellen Google Chrome Browser eingestellt[44], da dieser Stand Juli 2021 der meistgenutzte Browser ist (vgl. Stetic GmbH, 2021).

[44] *User-Agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.124 Safari/537.36*

```
headers = {
    "User-Agent": self.config["useragent"],
    "Accept-Language": "de, en; q=0.8, q=0.7"
}
try:
    self.data = requests.get(url, headers=headers, timeout=self.config["
    download"]["timeout"])
except requests.exceptions.ConnectionError as e:
    self.log.error("Error connecting to url {}. Error: {}".format(url, e))
    raise DownloadException
```

Quelltext 4.3: Ausschnitt aus der Download Funktion

In Quelltext 4.4 ist dargestellt, wie der HTML-Quellcode mithilfe von `BeautifulSoup4` in ein durchsuchbares Objekt umgewandelt wird. Inhalte, die für die weitere Analyse irrelevant sind, beispielsweise JavaScript-Elemente oder Formatierungselemente der Webseite, werden in diesem Schritt entfernt.

```
soup = BeautifulSoup(text, 'html.parser')
for script in soup(["script", "style", "link"]):
    script.extract()
for comments in soup.findAll(text=lambda text: isinstance(text, Comment)):
    comments.extract()
return soup
```

Quelltext 4.4: Ausschnitt aus der HTML-Vorbereitung

Policy

Mit den in der `Policy`-Klasse implementierten Methoden werden alle Analyseschritte vollzogen und gesteuert. In ihr werden mithilfe des in Unterabschnitt 3.5.3 erläuterten `spaCy` Moduls sowohl der Text in Sätze und Worte unterteilt, als auch alle weiteren Verarbeitungsschritte in einem Ausgabeobjekt zusammengeführt

HeadingBasedSectioner

Die Einteilung der Datenschutzerklärung in Überschriften und Absätze (engl. headings bzw. sections) wird mit Hilfe der Klasse `HeadingBasedSectioner` vorgenommen. Die aus einem anderen Projekt übernommene Klasse (siehe Unterabschnitt 3.5.3) wurde für diese Arbeit erweitert, sodass auch deutschsprachige, themenrelevante Überschriften erkannt werden. Weiterhin wurden viele vorher statische Parameter über die Konfigurationsdatei zugänglich gemacht.

ClassificationManager

Um die gefundenen Absätze thematisch zu kategorisieren, wird die `ClassificationManager`-Klasse genutzt. Sie stammt aus demselben Projekt wie die `HeadingBasedSectioner`-Klasse und wurde daher ebenfalls insofern erweitert, dass die Kategorisierung auch deutschsprachige Texte enthält. Weiterhin wurden die Kategorien im Sinne der Testkriterien umfangreich erweitert und in die Konfigurationsdatei eingepflegt. Ein Programmausschnitt der Kategorisierung einer Überschrift und ihres Inhalts ist in Quelltext 4.5 ersichtlich.

Dieser zeigt die Suche mit Hilfe regulärer Ausdrücke in Überschriften und dem zugehörigen Textabschnitt. Weiterhin sind einige der inhaltlichen Kategorien in Tabelle 4.1 samt ihrer Suchworte aufgelistet.

```
for pattern in self.PATTERN_LABELS:
    headinglookup = pattern["regex"].findall(text)
    if headinglookup:
        if pattern["label"] not in categories.keys():
            categories[pattern["label"]] = {"in_heading": headinglookup, "
    in_section": []}
        else:
            for result in headinglookup:
                categories[pattern["label"]]["in_heading"].append(result)

    sectionlookup = pattern["regex"].findall(section.lower())
    if sectionlookup:
        if pattern["label"] not in categories.keys():
            categories[pattern["label"]] = {"in_heading": [], "in_section"
    : sectionlookup}
        else:
            for result in sectionlookup:
                categories[pattern["label"]]["in_section"].append(result)
```

Quelltext 4.5: Ausschnitt aus der Kategorisierung einer Überschrift bzw. ihres Absatzes

Kategorie	Suchworte
gdpr-rights	ihre rechte|ihr recht|auskunftsrecht|berichtigung|lösch ung|einschränkung|...
gpdr	dsgvo|ds-gvo|verordnung
location	standort
personal-data	persönlich|person|beziehbar|individu|bezogen
privacy-shield	shield|schild|harbor|hafen
purpose	zweck
security	ssl|sicher|schutz|schütz
share	send|teil|transfer|weitergeb|weitergab

Tabelle 4.1: Ausschnitt aus den genutzten inhaltlichen Kategorien

Statistics

Die statistischen Testkriterien (siehe Unterabschnitt 3.2.1) werden mit den in der Klasse `Statistics` implementierten Funktionen erfasst. Dies beinhaltet strukturelle Informationen wie die Anzahl von Sätzen, Worten und Überschriften, aber auch Durchschnittswerte, wie beispielsweise die durchschnittliche Anzahl von Worten oder Sätzen je Absatz. Die Lesbarkeitsanalyse und Berechnung der Entropie werden ebenfalls im Rahmen dieser Klasse vorgenommen. Die hierzu verwendeten Technologien werden in Unterabschnitt 3.5.3 beschrieben.

Trackers

In der `Trackers`-Klasse wird gegebenenfalls die aktuell verfügbare Liste an Trackern der Exodus Privacy Datenbank (siehe Unterabschnitt 3.5.3) heruntergeladen und abgespeichert. Weiterhin sind in ihr Methoden zur Suche nach Trackernamen in der Datenschutzerklärung integriert. Wie in Quelltext 4.6 ersichtlich, wird für das Herunterladen bzw. Aktualisieren der gespeicherten Tracker die gleiche Methode genutzt, wie für den Download der Datenschutzerklärungen. In der Konfigurationsdatei ist eine automatische Aktualisierung der Tracker aktivierbar.

```
d = Download()
d.get_url(self.config["trackers"]["exodusapi"])
trackers = json.loads(d.get_raw_data())
if trackers["trackers"]:
    for tracker in trackers["trackers"].keys():
        # Remove comments in tracker names
        trackers["trackers"][tracker]["name"] = trackers["trackers"][
    tracker]["name"].split("(")[0].lower()
    self.log.info("{} trackers found in Exodus API".format(len(trackers["
    trackers"])))
    self.db.write_trackers(trackers["trackers"])
```

Quelltext 4.6: Funktion zum Herunterladen der Tracker-Datenbankeinträge

Scoring

Die Bewertung der Datenschutzerklärung und die Zusammenfassung aller hierzu notwendigen Daten findet in der `Scoring`-Klasse statt. Der Bewertungsmaßstab (siehe Abschnitt 4.3) kann in der Konfigurationsdatei angepasst werden. Die zum Analysezeitpunkt festgelegte Punkteverteilung wird ebenfalls gespeichert, sodass sich Veränderungen am Bewertungsmaßstab nicht auf bestehende Analysen auswirken. In Quelltext 4.7 ist ersichtlich, wie mithilfe der Kategorisierung ermittelt wird, ob der Zweck des Teilens von Daten mit Dritten erwähnt wurde. Wenn im selben Absatz sowohl das Teilen von Daten als auch Dritte erwähnt werden, wird dies entsprechend positiv gewertet.

```
def third_party_purpose(self) -> dict:
    score = {
        "score": self.score_config["content"]["third-parties"]["purpose"],
        "max": self.score_config["content"]["third-parties"]["purpose"]
    }
    third = self.categories.get("third-parties", {})
    purpose = self.categories.get("purpose", {})
    common_sections = commonvalues(list(third.keys()), list(purpose.keys()
    ))
    if not common_sections:
        score["score"] = 0
        score["reason"] = "No third-party purpose information found."
    return score
```

Quelltext 4.7: Ermittlung und Bewertung, ob Zweck des Teilens mit Dritten erwähnt ist

Visualization

Um den Vergleich der analysierten Datenschutzerklärungen zu erleichtern, wurde mit Hilfe des Python-Moduls `Plotly`[45] die Möglichkeit integriert, über die API auch Diagramme abzurufen. Die Klasse `Visualization` stellt alle hierzu notwendigen Methoden bereit und fasst Ergebnisse bisheriger Analysen in diesen zusammen. Die in Kapitel 5 gezeigten Auswertungen wurden ebenfalls hiermit generiert.

4.2 Datenbankstruktur

Die zu einem Analyseprojekt (siehe Abbildung 4.3) zugehörigen Daten werden in der Datenbank persistent gespeichert. Im Folgenden wird die Struktur der SQLite-Datenbank erläutert, die in diesem Projekt Anwendung findet. Sie gliedert sich, wie in Abbildung 4.4 zu sehen, in zwei Tabellen auf.

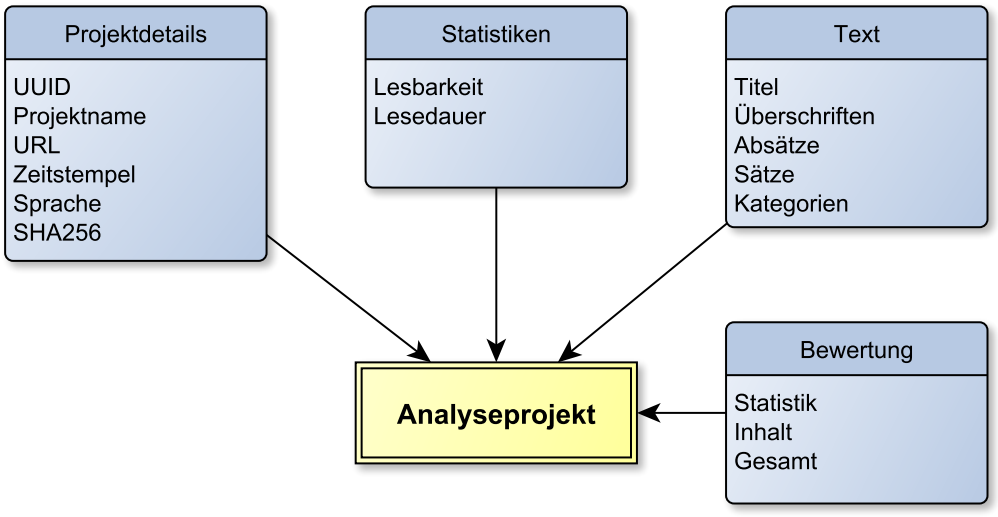

Abbildung 4.3: Daten eines Analyseprojekts

[45] `https://plotly.com/python/`

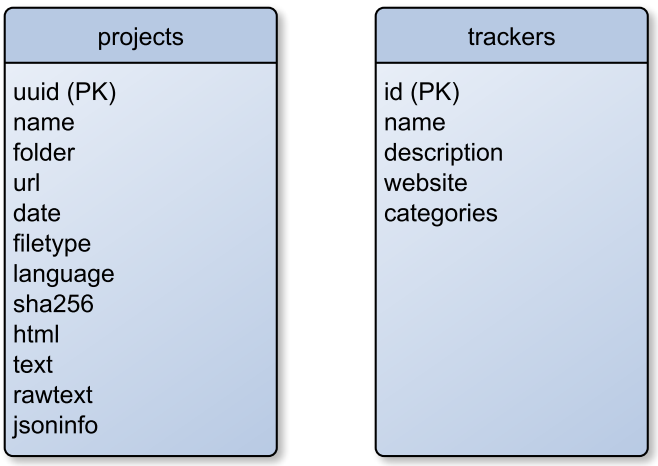

Abbildung 4.4: Datenbankstruktur

projects

Die Tabelle *projects* enthält alle Daten zu bereits getätigten Analysen. Diese werden als Projekt abgespeichert, um sie später erneut abrufen und gegebenenfalls neu analysieren zu können. Neben der UUID als Identifikationsmerkmal wird zum Ende des Analyseablaufs (siehe Abschnitt 3.4) auch der Name des Projekts sowie ein gegebenenfalls vorhandener Datenordner gespeichert. Die URL der Datenschutzerklärung sowie ihr Abrufzeitpunkt werden ebenfalls festgehalten. Auch wenn das Programm derzeit nur HTML-Quelltext unterstützt, wird der Dateityp des Downloads ebenfalls gespeichert, unter dem Gesichtspunkt zukünftiger Erweiterungen, beispielsweise hinsichtlich PDF- oder Word-Dokumenten. Neben dem heruntergeladenen HTML-Quelltext wird auch der hieraus extrahierte Text abgelegt und über diesen ein SHA256 Hash als Prüfsumme gebildet. Das Feld *jsoninfo* enthält alle bewertungsrelevanten Daten im JSON-Format und kann ohne Änderungen an der Datenbank um weitere Prüf- oder Datenpunkte erweitert werden.

trackers

Ähnlich zu der Suche nach Ländernamen wird bei der Suche nach Trackernamen auch ein Datensatz benötigt, auf dessen Basis die Suche erfolgt. Aus der von Exodus Privacy Datenbank wird, wie in Unterabschnitt 3.5.3 erwähnt, eine Liste von Trackern heruntergeladen und die Namen der Tracker samt ihrer Beschreibung, zugehöriger Verarbeitungskategorien und der URL ihrer Webseite abgelegt, damit nicht bei jeder Analyse API-Anfragen gestellt werden müssen.

4.3 Bewertungsmaßstab

Im Rahmen des Konzepts (siehe Abschnitt 3.4) wurde die maximal erreichbare Punktzahl auf 100 festgelegt. Wie diese auf die einzelnen Testkriterien verteilt wird, wird im folgenden Abschnitt beschrieben. Dass dem Benutzer datenschutzrelevante Informationen auf eine möglichst zugängliche Art und Weise nähergebracht werden sollen, ist in Art. 12 DSGVO klar definiert. Die statistischen Kriterien setzen an dieser Stelle an und bewerten die Struktur, Entropie und Lesbarkeit des Textes. Da eine für die Zielgruppe nicht verständliche oder unübersichtliche Datenschutzerklärung nicht der DSGVO-gegebenen Vorgabe entspricht, werden diesen Kriterien daher im Rahmen der Wertung 30 Punkte zukommen. Die übrigen 70 Punkte werden auf die inhaltlichen Testkriterien aufgeteilt.

4.3.1 Statistik

Tabelle 4.2 zeigt die Aufteilung der Punkte auf die drei Kriterien der Kategorie Statistik (siehe Abschnitt 3.2). Als wichtigstes Maß dieser Kategorie werden in der Lesbarkeitsanalyse maximal 15 Punkte vergeben, da auf diese Weise sichergestellt wird, dass unverständlich formulierte Texte entsprechend abgewertet werden. Beide Lesbarkeitsalgorithmen geben, wie in Unterabschnitt 2.2.2 benannt, die Anzahl der zum Textverständnis benötigten Schuljahre aus. Da die allgemeine Schulpflicht in Deutschland bis zu 10 Jahre beträgt, wird bei einer Überschreitung dieses Werts ein Punkt pro Jahr abgezogen. Der Bewertung der Entropie werden 10 Punkte vergeben, um inhaltsarme Datenschutzerklärungen entsprechend abzuwerten. Sie bildet sich aus dem Produkt der Inhaltswertung (siehe Unterabschnitt 4.3.2) und dem Informationsgehalt des Textes. Da die Inhaltswertung einen Wertebereich von 0 - 70 Punkten besitzt, die Entropie aber in der Theorie einen unendlichen Wertebereich hat, muss letzterer hierfür zunächst normalisiert werden. Die Entropie von Texten liegen sprachenübergreifend bei einem Mittelwert von etwa 6 (vgl. Bentz u. a., 2017). Für den Wertebereich der Entropie wird daher bei einem Wert von 6 eine obere Schranke gesetzt. Das Produkt der Inhaltswertung und des Entropiewerts wird auf einen Wertebereich von 0 - 10 Punkten normalisiert. Im Bereich der Strukturwertung werden fünf Punkte vergeben. Damit Fachtexte übersichtlich und verständlich sind, sollten diese im Durchschnitt maximal sechs Sätze je Absatz messen (vgl. Ultimate Proofreader, 2021). Falls dieses Maß im Durchschnitt überschritten wird, wird jeweils ein Punkt pro Satz Überschreitung abgezogen.

Kriterium	Punkte
Struktur	5
Entropie	10
Lesbarkeit	15
Summe	**30**

Tabelle 4.2: Übersicht des Bewertungsschemas der statistischen Kriterien

4.3.2 Inhalt

Die verbliebenen 70 Punkte werden, wie in Tabelle 4.3 dargestellt, auf die folgenden Kategorien an Testkriterien (siehe Abschnitt 3.2) aufgeteilt. Der umfangreichste Teilbereich ist die Datenbehandlung. Dieser beinhaltet die wichtigsten Kriterien in Hinblick auf die Datenerhebung und -speicherung. Diese Kategorie fließt daher mit 30 Punkten in die Wertung ein. Die Kategorien Allgemein und Dritte gehen mit jeweils 15 Punkten in die Wertung ein. Sie enthalten generelle Prüfkriterien wie das Änderungsdatum, behandeln aber auch das Teilen von Daten mit Dritten. Die Prüfung, ob die mobile Applikation und ihre Berechtigung Erwähnung in der Datenschutzerklärung finden, wird mit 10 Punkten in der Wertung berücksichtigt. Wie die für die Kategorien vergebenen Punkte auf die einzelnen Kriterien aufgeteilt werden und wie die hierzu notwendigen Daten erhoben werden, wird im Folgenden erläutert.

Kriterium	Punkte
Allgemein	15
Mobile Applikation	10
Dritte	15
Datenbehandlung	30
Summe	**70**

Tabelle 4.3: Übersicht des Bewertungsschemas der inhaltlichen Kriterien

Allgemein

Die Verteilung der Punkte auf die Kategorie Allgemein ist in Tabelle 4.4 ersichtlich. Da durch genaue Verwaltung der Zugriffsrechte Datenschutzvorfällen[46] vorgebeugt werden kann, wird der Prüfung, ob ein Rechtemanagement erwähnt wird, mit fünf Punkten die höchste Priorität dieser Kategorie zuteil. Die Überprüfung, ob ein Gültigkeits- oder Änderungsdatum in der Datenschutzerklärung benannt wird, wird mit vier Punkten bewertet, da der Benutzer durch eine Datumsangabe nachvollziehen kann, ob sich in der Datenverarbeitung des Produkts etwas verändert hat. Ebenfalls mit vier Punkten bewertet wird die Prüfung auf vorhandene Kontaktdaten. Nur auf diese Weise kann der Ansprechpartner für datenschutzrelevante Fragen wie z.B. das Löschen des Kontos direkt ermittelt werden. Eine weitere Möglichkeit, Änderungen an der Datenschutzerklärung nachvollziehen zu können, ist eine Versionierung. Dem Nutzer frühere Stände der Datenschutzerklärung anzuzeigen ist keine Pflichtangabe, daher geht dieses Kriterium mit zwei Punkten in die Wertung ein.

[46] Beispiel: `https://www.golem.de/news/amazon-tochter-ring-mitarbeiter-konnten-in-kundenwohnungen-blicken-1901-138682.html`

Kriterium	Punkte
Änderungsdatum	4
Versionierung	2
Kontaktdaten	4
Rechtemanagement	5
Summe	**15**

Tabelle 4.4: Übersicht des Bewertungsschemas der Kriterien im Bereich Allgemein

Mobile Applikation

Wie in Tabelle 4.5 dargestellt, unterteilt sich die Prüfung im Bereich mobiler Applikationen auf die reine Erwähnung derselben und die Erwähnung üblicher, von diesen anforderbarer Berechtigungen. Die Prüfung auf benannte Berechtigungen wird mit acht Punkten gewertet, da die reine Nennung der Applikation weniger Informationsgehalt bietet, als die Beschreibung der durch ihre Berechtigungen, wie etwa Standort- oder Kontaktzugriff, erfassbaren Daten.

Kriterium	Punkte
Erwähnung	2
Berechtigungen	8
Summe	**10**

Tabelle 4.5: Übersicht des Bewertungsschemas der Kriterien im Bereich Mobile Applikation

Dritte

Wie aus Tabelle 4.6 hervorgeht, teilen sich die 15 Punkte der Kategorie Dritte hauptsächlich auf das Kriterium Tracker auf. Da mit Trackingmethoden zahlreiche personenbezogene Daten aufgezeichnet werden, die mit der Nutzung anderer Applikationen verknüpft werden können, wird dieses Kriterium mit neun Punkten bewertet. Für jeden in der Datenschutzerklärung gefundenen Tracker wird ein Punkt von dieser Wertung abgezogen. Falls Tracker laut Exodus Privacy mehreren Kategorien, zum Beispiel nicht nur der Analyse, sondern auch der Werbung, angehören, wird für zusätzliche Kategorien ebenfalls Punktabzug vorgenommen. Dass das Teilen von Daten mit Dritten und der Zweck dessen benannt wird, wird jeweils mit drei Punkten bewertet. Falls Informationen zu Dritten und dem Zweck des Teilens von Daten fehlen, wird die im Kriterium Tracker erreichbare Punktzahl auf null gesetzt, da eine inhaltsarme Datenschutzerklärung sonst mit neun Punkten belohnt werden würde.

Kriterium	Punkte
Teilen von Daten	3
Zweck des Teilens	3
Tracker	9
Summe	**15**

Tabelle 4.6: Übersicht des Bewertungsschemas der Kriterien im Bereich Dritte

Datenbehandlung

Die in der Kategorie Datenbank verfügbaren Punkte werden, wie aus Tabelle 4.7 entnommen werden kann, auf sieben Testkriterien aufgeteilt. Die größte Priorität kommt hier der Verschlüsselung und Anonymisierung/Pseudonymisierung mit acht bzw. sieben Punkten zu. Eine verschlüsselte Übertragung bzw. Speicherung verhindert das Ausspähen der Daten durch Unbefugte, erhält daher die höhere Punktzahl. Wenn personenbezogene Daten anonymisiert bzw. pseudonymisiert werden, sind diese dem Benutzer nicht mehr bzw. nicht ohne weiteres zuordenbar. Da dies ebenfalls eine wertvolle Schutzmaßnahme ist, kommt diesem Kriterium die zweithöchste Punktzahl zu. Wenn für die Datenverarbeitung ein entsprechender Zweck angegeben wird, werden hierfür sechs Punkte vergeben, da diese Angabe nach Art. 13 DSGVO zwingend erforderlich ist. Die durch die Datenschutzgrundverordnung gegebenen Rechte müssen ebenfalls zwingend benannt werden. Die Ausübung der Rechte ist für den Benutzer allerdings manueller Aufwand, weshalb dieses Kriterium mit drei Punkten geringer gewertet wird als allgemeingültige Schutzmaßnahmen wie Anonymisierung bzw. Verschlüsselung. Die Prüfung, ob auf die Thematik der Speicherdauer eingegangen wird, wird ebenfalls mit drei Punkten bewertet. Auch die Benennung dieser ist eine Pflichtangabe und zeigt auf, wie lange bestimmte personenbezogene Daten, z.B. der Standortverlauf, gespeichert werden. Ob der Verarbeitungsstandort benannt ist, wird mit zwei Punkten gewertet. Wenn in der Datenschutzerklärung optionale Daten benannt sind, beispielsweise im Rahmen der Nutzungsanalyse, wird dies mit einem Punkt gewertet.

Kriterium	Punkte
Zweck der Erhebung	6
Anonymisierung	7
Verschlüsselung	8
DSGVO-Rechte	3
Speicherdauer	3
Standort Datenverarbeitung	2
Optionale Daten	1
Summe	**30**

Tabelle 4.7: Übersicht des Bewertungsschemas der Kriterien im Bereich Datenbehandlung

4.4 Ausgabewerte

Die in der Analyse ermittelten Daten werden, wie in Abbildung 4.5 dargestellt, im Rahmen der Ausgabe in einem JSON-Objekt (siehe Unterabschnitt 3.5.1) zusammengefasst und ausgegeben. Auf diese Art und Weise ist sichergestellt, dass die Daten im Rahmen anderer Programme weiter genutzt werden können, aber auch bei manueller Bedienung in einem strukturierten, lesbaren Format vorliegen.

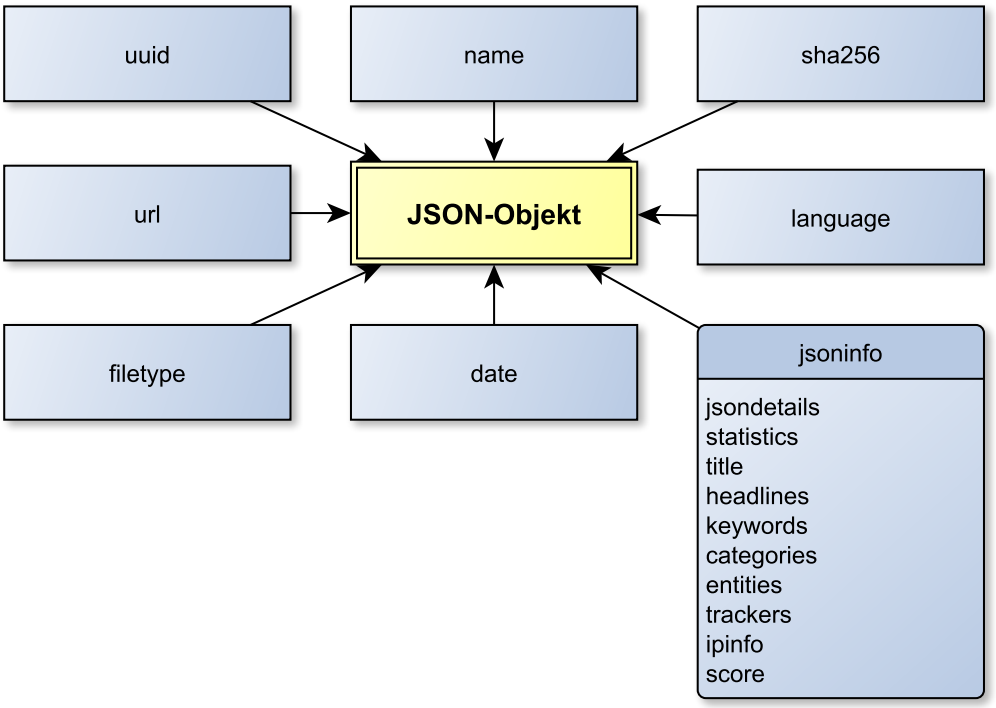

Abbildung 4.5: Aufbau des JSON-Objektes der Ausgabe

Die Struktur der Ausgabe folgt im Wesentlichen der in Abschnitt 4.2 gezeigten Datenbankstruktur. Demnach werden neben Identifikationsmerkmalen wie der UUID, dem SHA256-Hash und dem Projektnamen alle relevanten Daten der Analyse ausgegeben. Diese beinhalten weiterhin ein Datenfeld `jsondetails`, welches bei der Eingabe der zu analysierenden URL ebenfalls an das Programm gesendet werden kann. In diesem können weitere Angaben zum Analyseprojekt getroffen werden, beispielweise die Kategorie der App als Unterscheidungsmerkmal. Dies wird im Rahmen der Massenanalyse in Kapitel 5 genutzt werden. Konkrete Ausgabewerte werden im Abschnitt 4.6 an einem Beispiel gezeigt.

4.5 Herausforderungen in der Implementierung

Bei der Programmierung eines vordefinierten Konzepts treten unweigerlich Herausforderungen auf, die im Rahmen der Umsetzung gelöst werden müssen. Im Folgenden wird auf Probleme eingegangen, die während der Umsetzung des Projekts auftraten. Weiterhin wird erläutert, wie diese gelöst wurden.

DDoS-Schutz

Viele Webseiten werden mit Diensten wie Cloudflare[47] vor Denial of Service- (DoS) bzw. Distributed Denial of Service-Attacken (DDoS) geschützt. Diese Attacken führen durch viele wiederholte Anfragen an die Webseite dazu, dass der Webserver den Dienst einstellt. Mit dem Cloudflare-Dienst geschützte Webseiten können mit der Python Bibliothek `requests` nicht heruntergeladen werden, da dieser vom Webbrowser das Lösen einer in JavaScript geschriebenen Rechenaufgabe (engl. Challenge) erfordert. Da die Python Bibliothek keinen JavaScript Code ausführt, wird die Rechenaufgabe daher nicht bestanden und die Webseite nicht erfolgreich heruntergeladen. Sollte der Download fehlschlagen, wird dieser über selbst gehosteten Google Chrome Dienst[48] heruntergeladen. Der Download über einen „echten“ Webbrowser benötigt mehr Zeit, kann aber die JavaScript Challenges gängiger Dienste lösen.

HTML-Struktur

Mithilfe der Überschrift-Elemente (h1, h2, ...) wird in HTML eine rein visuelle Struktur aufgebaut. Innerhalb des Quelltextes einer Webseite wird hierdurch aber keine Hierarchie geschaffen. Um den zu einer Überschrift zugehörigen Textabschnitt zu identifizieren, nutzt die in Unterabschnitt 3.5.3 erwähnte `Sectioner` Bibliothek reguläre Ausdrücke, sodass etwa der Inhalt zwischen Überschrift A und B auch korrekt zu Überschrift A zugeordnet wird. Da die Bibliothek aber in Tests die Struktur der Webseite nicht immer erfolgreich extrahiert hat, wird, wie in Quelltext 4.8 gezeigt, im Fehlerfall auf eine eigene Implementierung, ebenfalls auf Basis regulärer Ausdrücke, zurückgegriffen.

Zuverlässigkeit Machine Learning

Mit `spaCy` ist ein mächtiges NLP-Toolkit implementiert. Dieses bietet für zahlreiche Sprachen bereits trainierte Modelle an. Da die Suche nach Entitäten (Firmen, Länder, Namen etc.) im Rahmen von Tests nur in englischsprachigen Texten zuverlässig zu funktionieren schien, wurde davon abgesehen, diese Funktionalität weiter auszubauen und auch im Rahmen der Bewertung zu nutzen. `spaCy` bietet zwar auch die Möglichkeit, ein eigenes Machine Learning Modell zu trainieren, um beispielsweise ein auf Datenschutzerklärungen geeichtes Textverständnis zu implementieren, mit dem womöglich eine Erweiterung der Testkriterien auf Basis des Inhalts möglich wäre. Dies übersteigt allerdings den leistbaren Aufwand in dieser Arbeit.

[47] `https://www.cloudflare.com/de-de/`
[48] `https://github.com/browserless/chrome`

```
1  headlines = self.soup.find_all(["h1", "h2", "h3", "h4", "h5"])
2  sections = []
3  for headline in headlines[:]:
4      regex = "<{}.*?>.*?{}.*?<\/{}.*?>(.*?)<h\d.*?>".format(headline.name,
       headline.text.strip(), headline.name)
5      try:
6          pattern = re.compile(regex)
7      except re.error:
8          headlines.remove(headline)
9          self.log.warn("Headline Compile Error. Skipping Entry.")
10         continue
11     content = re.search(pattern, self.unescaped.replace("\r", " ").replace
       ("\n", " "))
12     if content:
13         section = BeautifulSoup(content.groups()[0].strip(), features="
       lxml").text.strip()
14         sections.append(section)
15     else:
16         sections.append(None)
```

Quelltext 4.8: Ausschnitt aus der Zuordnung von Inhalt zu Überschriften

4.6 Analyseablauf anhand eines Beispiels

Um die Logik bzw. den Ablauf des Programms genauer aufzuzeigen, wird dieser im Rahmen eines Beispiels durchlaufen. Als Beispiel dient die Datenschutzerklärung des bekannten Instant Messengers Discord[49]. Die Analyse wird mit der API-Anfrage *http://localhost:8000/submit?url=https://-discordapp.com/privacy/* begonnen. Der Ablauf wird im Folgenden anhand der Protokollierung des Programms beschrieben.

Download

Die für die Analyse eingesendete URL wird, wie in Quelltext 4.9 aufgelistet, zunächst heruntergeladen. Da der HTTP Status Code[50] den Wert *200 OK* hat, wird von einem erfolgreichen Download ausgegangen und mit der Analyse fortgefahren. Um Whois-Informationen des Servers abzufragen, wird die IP für die Domain *discordapp.com* ermittelt und für die IP *162.159.129.233* die entsprechende Abfrage gesendet. Als Serverstandort wird San Francisco, USA herausgefunden. Als nächstes wird der Dateityp der heruntergeladenen Daten als *html*, eine Webseite, identifiziert. Weiterhin werden alle HTML-spezifischen Zeichen ersetzt. Die HTML-Daten werden durchsuchbar gemacht und der Titel der Webseite, *Datenschutzerklärung | Discord*, extrahiert. Der Textinhalt der Webseite wird ausgelesen, weiterhin wird über diesen ein SHA256-Hash gebildet. Die Prüfung, ob unter der URL und dem gebildeten Hash bereits eine Analyse in der Datenbank vorhanden ist, findet keine Ergebnisse. Daher wird mit der Analyse fortgefahren. Da kein Projektname übergeben wurde, wird dieser aus der Domain und dem Titel der Webseite gebildet: *discordapp.com - Datenschutzerklärung | Discord*

[49] https://discordapp.com/privacy/
[50] https://developer.mozilla.org/de/docs/Web/HTTP/Status

```
27-06-2021 10:44:45 Api          INFO     URL submitted for analysis:
    https://discordapp.com/privacy/
27-06-2021 10:44:45 Policy       INFO     Policy Init done.
27-06-2021 10:44:45 Api          INFO     Beginning analysis.
27-06-2021 10:44:45 Download     INFO     Download Init done.
27-06-2021 10:44:45 Download     INFO     Downloading URL https://
    discordapp.com/privacy/
27-06-2021 10:44:48 Download     INFO     Download successful.
27-06-2021 10:44:49 Download     INFO     Downloading URL https://ipinfo.
    io/162.159.129.233/json
27-06-2021 10:44:49 Download     INFO     Download successful.
27-06-2021 10:44:49 Policy       INFO     Server location: San Francisco,
    US
27-06-2021 10:44:49 Policy       INFO     Website detected. (Extension: .
    html)
27-06-2021 10:44:49 Download     INFO     Created Soup object.
27-06-2021 10:44:49 Download     INFO     Replacing all HTML characters
    using estimated encoding utf-8
27-06-2021 10:44:49 Download     INFO     Extracted title: Datenschutzerkl
    ärung | Discord
27-06-2021 10:44:49 Policy       INFO     SHA256 generated: 3
    c4e100fb9c59c59862ac9e8d210ae2f2605cfd5c6ed549a802fd641faba97be
27-06-2021 10:44:49 Policy       WARN     No project name provided. Using
    Domain + Title: discordapp.com - Datenschutzerklärung | Discord
27-06-2021 10:44:49 Policy       INFO     Download function finished.
```

Quelltext 4.9: Protokoll des Downloads

Spracherkennung

Da der Download erfolgreich war, wird, wie in Quelltext 4.10 gezeigt, mit der Analyse fortgefahren. Die Sprache des Textes wird ermittelt und das passende `spaCy` Machine Learning Modell geladen. Die heruntergeladene Discord Datenschutzerklärung wird als deutschsprachiger Text identifiziert, daher wird das Modell *de_core_news_lg* aktiviert.

```
27-06-2021 10:44:50 Policy       INFO     Detecting content language.
27-06-2021 10:44:50 Policy       INFO     German language detected.
27-06-2021 10:44:50 Policy       INFO     Loading spaCy model
    de_core_news_lg
```

Quelltext 4.10: Protokoll der Spracherkennung

Absatzerkennung

In Quelltext 4.11 wird der nächste Analyseschritt aufgeführt. Es werden 24 hervorgehobene Wortgruppen identifiziert. Da themenrelevante HTML-Überschriften gefunden wurden, werden ebenfalls gefundene, als *fett* markierte Zeichenketten nur als normaler Text angesehen. Der zu den Überschriften gehörende Text wird zugeordnet und im nächsten Analyseschritt, der Kategorisierung, weiter genutzt.

```
27-06-2021 10:44:55 Policy       INFO     Looking for interesting
    headlines with sectioner
27-06-2021 10:44:55 HeadingBasedSectioner INFO     24 bold heading
    candidates found.
27-06-2021 10:44:55 HeadingBasedSectioner INFO     Sectioning complete. 18
     sections identified
27-06-2021 10:44:55 Policy       INFO     Combining all headline data
27-06-2021 10:44:55 Policy       INFO     Found 18 headlines containing a
    text section.
27-06-2021 10:44:55 Policy       INFO     Headlines finished
```

Quelltext 4.11: Protokoll der Absatzerkennung

Kategorisierung

Die gefundenen 18 Textsektionen werden in diesem Schritt kategorisiert. Der Ablauf ist in Quelltext 4.12 ersichtlich. Das im Rahmen der Spracherkennung geladene spaCy Machine Learning Modell wird hierbei zum einen für die Zerlegung der Textbausteine in einzelne Sätze genutzt, zum anderen werden mit Hilfe seiner *Named Entity Recognition*[51] Namen von Firmen, Städten, Ländern und ähnlichem extrahiert. Da das deutschsprachige Modell, wie in Abschnitt 4.5 erläutert, in diesem Bereich nicht zuverlässig arbeitet, werden die extrahierten Entitäten nur ausgegeben, aber nicht für die weitergehende Analyse genutzt. In einem weiteren Schritt werden Stoppworte entfernt. Für jeden übrigen Begriff wird das Lemma ermittelt und separat zum Satz abgelegt. Die so gebildeten Grundformen der Worte werden im Anschluss zur Kategorisierung genutzt, sodass nur sinngebende Worte durchsucht werden müssen. Für die Kategorisierung werden alle Überschriften und Absätze mit Hilfe regulärer Ausdrücke durchsucht und auf diese Weise erkannt, welcher Abschnitt welches Thema enthält. Die erkannten Begriffe der Gruppe *Third-Parties* (Drittparteien) sind in Quelltext 4.13 beispielhaft dargestellt.

```
27-06-2021 10:44:55 Policy       INFO     Tagging sentences/tokens in
    headline 0.
27-06-2021 10:44:55 Policy       INFO     Trying to find entities in
    section 0
27-06-2021 10:44:55 Policy       INFO     Tagging sentences in section 0
[..]
27-06-2021 10:44:55 Policy       INFO     Tagging sentences/tokens in
    headline 17.
27-06-2021 10:44:55 Policy       INFO     Trying to find entities in
    section 17
27-06-2021 10:44:55 Policy       INFO     Tagging sentences in section 17
27-06-2021 10:44:55 Policy       INFO     Tagging raw text. This may take
    a while.
27-06-2021 10:44:58 Policy       INFO     Tagging sentences/tokens
    finished.
27-06-2021 10:44:58 Policy       INFO     Categorizing sections
```

Quelltext 4.12: Protokoll der Kategorisierung

[51] https://spacy.io/usage/linguistic-features#named-entities

```
"third-parties":{
  "2":{
     "in_heading":[],
     "in_section":[
        "gesellschaften",
        "dienstleister",
        "gesellschaften",
        "partner",
        "partner",
        "dritt"]
  },
  [...]
}
```

Quelltext 4.13: JSON-Ausgabe der Kategorie third-parties

Ermittlung statistischer Werte

In Quelltext 4.14 werden die Protokolleinträge bei der Ermittlung der statistischen Werte dargestellt. Zunächst wird die Anzahl von Worten, Sätzen und entsprechende Durchschnittswerte, beispielsweise der Anzahl von Worten und Sätzen je Überschrift, berechnet. Im Fall der Discord Datenschutzerklärung sind es 173,11 Worte pro Abschnitt. Im Anschluss werden Lesbarkeitsindizes berechnet. Da die für die Berechnung notwendige Zeit sehr gering ist, werden auch weitere, von der Bibliothek `textstat` unterstützte Messwerte für die Lesbarkeit berechnet. Für die Bewertung genutzt wird die Wiener Sachtextformel mit einem Wert von 14,3 Bildungsjahren, die für das Verstehen des Textes erforderlich sind. In Deutschland beträgt das Mindestalter zur Nutzung 16 Jahre[52], daher ist der Text nicht verständlich genug formuliert. Die in der Datenschutzerklärung am häufigsten vorkommenden Worte werden im Anschluss auf Basis der im Rahmen der Kategorisierung gespeicherten Grundformen der Worte gebildet. Im Fall von Discord sind dies *Datum*, *Dienst* und *Information*. Um den Informationsgehalt des Textes zu bestimmen, wird dessen Entropie berechnet.

Ermittlung benötigter Informationen

Im letzten Schritt vor der Bewertung werden, wie in Quelltext 4.15 sichtbar, zunächst in der Datenschutzerklärung benannte Trackern erkannt. In der Discord Datenschutzerklärung wurde Google Analytics identifiziert. In einem weiteren Schritt wird nach Zeitstempeln in Zusammenhang mit dem Gültigkeits- bzw. Änderungsdatum der Datenschutzerklärung gesucht und gegebenenfalls vorhandene E-Mail-Adressen extrahiert.

[52] `https://support.discord.com/hc/de/articles/360040724612-Warum-fragt-Discord-nach-meinem-Geburtsdatum-`

```
27-06-2021 10:44:58 Policy       INFO     Getting general statistics
27-06-2021 10:44:58 Statistics   INFO     Statistics Init done.
27-06-2021 10:44:58 Statistics   INFO     Generating Flesch Reading Ease
27-06-2021 10:44:59 Statistics   INFO     Generating Flesch Kincaid Grade
    Level
27-06-2021 10:44:59 Statistics   INFO     Generating SMOG Index
27-06-2021 10:44:59 Statistics   INFO     Generating Coleman Liau Index
27-06-2021 10:44:59 Statistics   INFO     Generating Automated Readability
27-06-2021 10:44:59 Statistics   INFO     Generating Linsear Write Formula
27-06-2021 10:44:59 Statistics   INFO     Generating Dale Chall
    Readability Score
27-06-2021 10:45:00 Statistics   INFO     Generating Text Standard Score
27-06-2021 10:45:02 Statistics   INFO     Generating Wiener Sachtext Score
27-06-2021 10:45:02 Statistics   INFO     Generating Reading Time
27-06-2021 10:45:02 Statistics   INFO     Gathering Top words
27-06-2021 10:45:03 Statistics   INFO     Gathering Top words finished.
27-06-2021 10:45:03 Statistics   INFO     Gathering entropy
27-06-2021 10:45:04 Statistics   INFO     Gathering entropy finished.
```

Quelltext 4.14: Protokoll der Ermittlung statistischer Werte

```
27-06-2021 10:45:04 Policy       INFO     Looking for trackers in each
    section
27-06-2021 10:45:04 Trackers     INFO     405 trackers found in database.
27-06-2021 10:45:04 Trackers     INFO     Trackers Init done.
27-06-2021 10:45:04 Policy       INFO     Found 1 trackers in the privacy
    policy.
27-06-2021 10:45:04 Policy       INFO     Searching for criteria-matching
    information.
27-06-2021 10:45:04 Policy       INFO     Finding dates
27-06-2021 10:45:04 Policy       INFO     Finding Email addresses
```

Quelltext 4.15: Protokoll der Ermittlung benötigter Informationen

Bewertung

In der Bewertung können maximal 100 Punkte erreicht werden. Wie diese auf die Testkriterien verteilt sind, wurde in Abschnitt 4.3 näher beleuchtet. Zunächst werden die inhaltlichen Kriterien bewertet, da die Wertung dieser in die Entropie-Wertung der statistischen Prüfkriterien eingeht. Ein Ausschnitt aus dem Protokoll dieses Schritts ist in Quelltext 4.16 dargestellt. Am Beispiel des Kriteriums „Dritte: Teilen von Daten“ wird gezeigt, wie auf Basis der Kategorisierung die Bewertung erfolgt. Der entsprechende Programmabschnitt ist unter Quelltext 4.17 aufgeführt. Weiterhin sind die Ergebnisse der Bewertung der Discord Datenschutzerklärung in Tabelle 4.8 aufgeschlüsselt.

In Summe werden 62 von 100 möglichen Punkten von ihr erreicht. Größere Abzüge gibt es bei den Prüfkriterien Entropie, Lesbarkeit, Rechtemanagement, Verschlüsselung und Speicherdauer. Hier fehlen entsprechende Informationen in der Datenschutzerklärung, weiterhin sind Anpassungen im Bereich der Lesbarkeit erforderlich.

```
27-06-2021 10:45:04 Scoring      INFO     Scoring Init done.
27-06-2021 10:45:04 Scoring      INFO     Date: 4/4
27-06-2021 10:45:04 Scoring      INFO     Versioning: 0/2
27-06-2021 10:45:04 Scoring      INFO     Contact: 4/4
27-06-2021 10:45:04 Scoring      INFO     Authorized: 0/5
[...]
27-06-2021 10:45:04 Scoring      INFO     Total scores:
27-06-2021 10:45:04 Scoring      INFO     Statistics: 20/30
27-06-2021 10:45:04 Scoring      INFO     Content: 42/70
27-06-2021 10:45:04 Scoring      INFO     Total: 62/100
```

Quelltext 4.16: Protokoll der Bewertung

Kriterium	**Punkte**	**Maximum**
Statistik		
Struktur	3	5
Entropie	6	10
Lesbarkeit	11	15
Inhalt		
Allgemein		
Änderungsdatum	4	4
Versionierung	0	2
Kontaktdaten	4	4
Rechtemanagement	0	5
Mobile Applikation		
Erwähnung	2	2
Berechtigungen	0	8
Dritte		
Teilen von Daten	3	3
Zweck des Teilens	3	3
Tracker	8	9
Datenbehandlung		
Zweck der Erhebung	6	6
Anonymisierung	7	7
Verschlüsselung	0	8
DSGVO-Rechte	2	3
Speicherdauer	0	3
Standort Datenverarbeitung	2	2
Optionale Daten	1	1
Summe	**62**	**100**

Tabelle 4.8: Beispielhafte Bewertung der Discord Datenschutzerklärung

```
def third_party_sharing(self) -> dict:
    score = {
        "score": self.score_config["content"]["third-parties"]["sharing"],
        "max": self.score_config["content"]["third-parties"]["sharing"]
    }
    third = self.categories.get("third-parties", {})
    sharing = self.categories.get("share", {})
    common_sections = commonvalues(list(third.keys()), list(sharing.keys()
    ))
    if not common_sections:
        score["score"] = 0
        score["reason"] = "No third-party sharing information found."
    self.log.info("3rd Sharing: {}/{}".format(score["score"], score["max"
    ]))
    return score
```

Quelltext 4.17: Funktion zur Bewertung des Teilens von Daten mit Dritten

Ausgabe und Speicherung

Wie in Quelltext 4.18 ersichtlich, wird für die persistente Speicherung des Analyseprojekts eine UUID generiert, anhand welcher auch später die Analyseergebnisse abgefragt werden können. Für die Speicherung in der Datenbank werden alle eruierten Daten in Form eines Projekts zusammengefasst. Das hierbei für die Speicherung erzeugte JSON-Objekt wird auch in der Ausgabe genutzt.

```
27-06-2021 10:45:04 SQLite       INFO     Created project discordapp.com -
    Datenschutzerklärung | Discord with UUID
    b27131d45bab4a76949d6c20e5ade243
```

Quelltext 4.18: Protokoll der Ausgabe und Speicherung

5 Massenanalyse und Auswertung

Im Rahmen dieser Arbeit wurde eine Massenanalyse von mehr als 400 Datenschutzerklärungen erfolgreich vorgenommen. Die Analyseergebnisse werden in diesem Kapitel vorgestellt. Zunächst wird hierzu die Datengrundlage beschrieben. In den nachfolgenden Abschnitten werden die Ergebnisse präsentiert und Auswertungen getroffen.

5.1 Datengrundlage

Um eine möglichst große Bandbreite verschiedenster Datenschutzerklärungen zu testen, werden die „Top"-Kategorien des Google Play Stores[53] genutzt und zu jeder dieser Kategorien die angezeigten Android Applikationen extrahiert. Seit Anfang 2017 ist es für im Google Play Store neu eingestellte Android Applikationen Pflicht, eine Datenschutzerklärung zu verlinken (vgl. Diercks, 2017). Neben dem Anzeige- und Package-Namen[54] und dem Link zur Datenschutzerklärung der Applikation wurde auch ihr Genre (z.B. Health, Social) notiert, um sie gegebenenfalls in Auswertungen zu verwenden.

Aus den Kategorien *Top-Apps*, *Bestseller-Apps* und *Apps mit dem höchsten Umsatz* sind jeweils 200 Android Applikationen für die deutschsprachige Region gesammelt worden. Da 30 Applikationen Teil mehrerer der drei Kategorien sind, werden die Datenschutzerklärungen von 570 individuellen Applikationen analysiert. Die Analyseergebnisse werden im folgenden Abschnitt ausführlich erläutert.

5.2 Auswertung

Im Zuge der Massenanalyse wurden die Datenschutzerklärungen von 570 Android Applikationen untersucht. Für die Datenschutzerklärungen von 12 Applikationen wurde keine Analyse vorgenommen, da diese nur auf die allgemeine Datenschutzerklärung von Google[55] verweisen, aber nicht zu Google zugehörig sind. Der Download von 35 Datenschutzerklärungen schlug fehl. Dies scheiterte entweder an nicht existenten Webseiten oder aber an einem DDoS-Schutz der Webseite an sich. Bei 37 Datenschutzerklärungen konnte keine Formatierung extrahiert werden, weil unter anderem Überschriften nicht entsprechend formatiert bzw. gekennzeichnet waren. Demnach wäre im Rahmen der Kategorisierung keine korrekte Prüfung auf Themenabschnitte möglich gewesen, weshalb die Analyse abgebrochen wurde. 15 Datenschutzerklärungen lagen in einem nicht unterstützten Format vor, beispielsweise TXT oder PDF. Aus beiden Formaten ließe sich keine Formatierung extrahieren, demnach wäre hier ebenfalls keine Kategorisierung

[53] `https://play.google.com/store/apps/top`
[54] Eindeutige Identifikation einer Applikation im Google Play Store und auf dem Smartphone
[55] `https://policies.google.com/privacy`

von Abschnitten möglich, die für die Bewertung eine zentrale Rolle spielt. Wegen einer nicht unterstützten Sprache wurde die Analyse von fünf Datenschutzerklärungen abgebrochen. Diese lagen nur in französischer bzw. russischer Sprache vor. Im Rahmen der Kontrolle der in der Massenanalyse erfassten Daten fiel auf, dass trotz umfangreicher Prüfungen beim Download dennoch Texte analysiert wurden, die keine Datenschutzerklärungen waren und demnach im Google Play Store fälschlicherweise als Datenschutzerklärung verlinkt wurden. Aus diesem Grund wurde manuell geprüft, bei welchen Analysen dies der Fall war und hierdurch nochmals 46 Datenschutzerklärungen aus der Datenbank entfernt.

Die 420 „Top“-Applikationen, für die eine erfolgreiche Analyse der Datenschutzerklärung vorgenommen wurde, setzen sich aus 29 verschiedenen Genres des Google Play Stores zusammen. Der Anteil der jeweiligen Genres ist in Abbildung 5.1 ersichtlich.

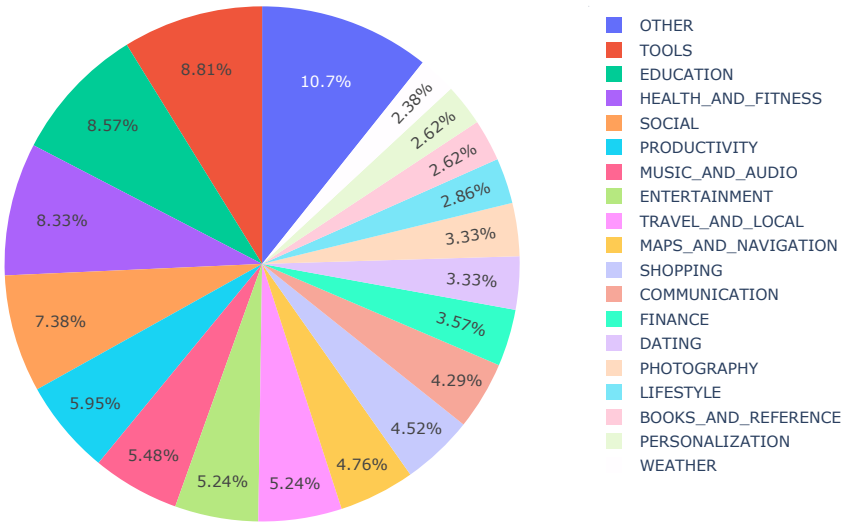

Abbildung 5.1: Genre der Applikationen im Google Play Store

Auch wenn die Analyse auf Basis der „Top“-Applikationen in Deutschland stattfand, so lag nur etwa ein Drittel der getesteten Datenschutzerklärungen in deutscher Sprache vor. Zum Teil veröffentlichen die Anbieter nur einen international einheitlichen Link zu selbiger, oder haben auf der Webseite keine Spracherkennung integriert, sodass nur die englische Version angezeigt wird, auch wenn eine deutschsprachige Version der Webseite verfügbar wäre. Da die Analyse auf beide Sprachen gleichermaßen ausgelegt ist, hat dies aber keinen weiteren Einfluss auf die Wertung.

5.2.1 Einzelauswertung

Um nachzuweisen, dass der konzipierte Analyseansatz korrekt funktioniert, werden manuelle Analysen der zwei laut automatisierter Analyse besten und schlechtesten Datenschutzerklärungen vorgenommen und manuell gegen die gleiche Liste von Prüfkriterien validiert. Im Anschluss werden manuell ermittelte Analyseergebnisse mit denen der automatisierten Analyse verglichen. Die Ergebnisse der automatisierten Analyse sind in Tabelle 5.1 ersichtlich.

Kriterium	DRK	Fonts Keyboard	Oje, ich wachse	KWGT	Max
Statistik					
Struktur	5	5	0	5	5
Entropie	8	9	1	1	10
Lesbarkeit	13	9	15	15	15
Inhalt					
Allgemein					
Änderungsdatum	4	4	0	4	4
Versionierung	0	0	0	0	2
Kontaktdaten	4	4	0	4	4
Rechtemanagement	5	5	0	0	5
Mobile Applikation					
Erwähnung	2	2	2	2	2
Berechtigungen	8	8	0	0	8
Dritte					
Teilen von Daten	3	3	0	0	3
Zweck des Teilens	3	3	0	0	3
Tracker	9	8	0	0	9
Datenbehandlung					
Zweck der Erhebung	6	6	6	0	6
Anonymisierung	7	7	0	0	7
Verschlüsselung	8	8	0	0	8
DSGVO-Rechte	2	2	0	0	3
Speicherdauer	3	3	0	0	3
Standort Datenverarb.	2	2	0	0	2
Optionale Daten	0	1	0	0	1
Summe	**92**	**89**	**24**	**31**	**100**

Tabelle 5.1: Bewertung der zwei am besten und schlechtesten bewerteten Datenschutzerklärungen

Erste Hilfe DRK

Die deutschprachige Datenschutzerklärung der Applikation „Erste Hilfe DRK“[56] erreicht in der automatisierten Analyse eine Wertung von 92 Punkten und somit die Bestwertung der analysierten Datenschutzerklärungen, wie in Tabelle 5.1 ersichtlich.

Die Applikation dient hauptsächlich der Anzeige von Anleitungen im Rahmen der Ersten Hilfe. Die dazugehörige Datenschutzerklärung ist mit durchschnittlich sechs Sätzen je Überschrift gemäß Bewertungsmaßstab korrekt strukturiert. Weiterhin besitzt der Text mit einem Entropiewert von 5,3 (in Tabelle 5.1, wie in Abschnitt 4.3 erläutert, normalisiert) einen guten Informationsgehalt. Mit einem Lesbarkeitsindex von 12,4 Bildungsjahren (Wiener Sachtextformel) ist die Datenschutzerklärung eher entsprechend Abiturniveau formuliert, weshalb in der Wertung zwei Punkte abgezogen werden. Als Änderungsdatum ist Mai 2018, weiterhin der Versionsstand 1.0 angegeben. Neben Kontaktdaten des Verantwortlichen (DRK-Service GmbH, Berlin) samt Telefonnummer und E-Mail-Adresse sind auch technische und organisatorische Maßnahmen erwähnt, die dem Schutz der Daten bzw. Rechte von Betroffenen dienen. Die Erste Hilfe DRK Applikation wird samt ihrer anforderbaren Standort-Berechtigung ausführlich in der Datenschutzerklärung beschrieben. Diese Berechtigung ist optional und dient nicht dem Erstellen von Bewegungsprofilen, sondern wird nur in anonymisierter Form verarbeitet. Personenbezogene Daten werden mit einem namentlich erwähnten Dienstleister geteilt, welcher IT-Dienstleistungen für den Verantwortlichen erbringt. Es werden in der Datenschutzerklärung keine Tracker erwähnt. Weiterhin wird detailliert beschrieben, welche Daten bei dem Aufruf der Applikation übermittelt werden und zu welchem Zweck die Ermittlung erfolgt, beispielsweise die IP-Adresse des Geräts, die technisch für die Kommunikation erforderlich ist. Diese wird nach der Nutzung gelöscht bzw. anonymisiert. Die Übertragung der Daten findet verschlüsselt statt, von einer verschlüsselten Speicherung wird nicht gesprochen. Der Anbieter speichert aber laut Datenschutzerklärung keinerlei personenbezogene Daten, ohne sie vorher zu anonymisieren. Dem Betroffenen werden Rechte in Bezug auf die Verarbeitung zugestanden, unter anderem die Rechte auf Auskunft, Berichtigung, Löschung und Widerspruch. Es wird keine exakte Speicherdauer angegeben, aber benannt, dass persönliche Daten nach Zweckerfüllung bzw. gesetzlichen Pflichten gelöscht werden. Der Standort des Verantwortlichen befindet sich in Deutschland, es kann bei Nutzung der optionalen Kartendienste von Google bzw. Apple aber eine Kommunikation in die USA erfolgen.

Im Vergleich der manuellen mit der automatisierten Analyse zeigt sich, dass es in den inhaltlichen Testkriterien nur in der Versionierung, den DSGVO-Rechten und Optionalen Daten Abzüge gab. In der Datenschutzerklärung ist vermerkt, dass es sich bei ihr um die erste Version handelt. Demnach kann es keine anzeigbaren Vorversionen geben. Dem Nutzer werden in der Datenschutzerklärung einige Rechte zugesichert, allerdings wird hierbei kein Bezug zu entsprechenden Artikeln in der Datenschutzgrundverordnung hergestellt, weshalb nur zwei von drei möglichen Punkten vergeben wurden. Dies ist so

[56] `https://play.google.com/store/apps/details?id=de.bitsz.android.drkapp`

konfiguriert, da automatisiert mit einem Verweis auf die DSGVO eher sichergestellt ist, dass alle Betroffenenrechte enthalten sind. Der manuellen Analyse nach entsprechen die dargelegten Rechte aber im Wesentlichen dem gegebenen Standard. Freiwillige bzw. optionale Angaben oder Daten werden in der Datenschutzerklärung nicht erwähnt, weshalb hierfür in der automatisierten Analyse keine Punkte vergeben wurden. Der manuellen Analyse nach wurden die Testkriterien korrekt bewertet und alle inhaltlichen Themen zutreffend bestimmt.

Fonts Keyboard

Wie in Tabelle 5.1 dargestellt, erhält die englischsprachige Datenschutzerklärung der Applikation „Fonts Keyboard“[57] mit 89 Punkten die zweitbeste Platzierung in der automatisierten Analyse.

Die Datenschutzerklärung der Tastatur-Applikation verfügt in Struktur und Entropie über nahezu identische Eigenschaften wie die Datenschutzerklärung der DRK Applikation, ist aber mit einem Lesbarkeitsindex von 15,9 Bildungsjahren (Flesch-Kincaid) und dahin einhergehenden 9 von 15 Punkten, sehr kompliziert bzw. schwer verständlich formuliert. Sie wurde zum 01.04.2021 zuletzt geändert und verfügt über keine Versionierung. Die Kontaktdaten des in Italien ansässigen Verantwortlichen und des Datenschutzbeauftragten werden korrekt angegeben, weiterhin sind vorgenommene, organisatorische Maßnahmen zum Schutz der Daten benannt. Die Fonts Keyboard Applikation wird samt notwendiger Berechtigungen erwähnt. Wenn der Tastatur die Berechtigungen gewährt werden, werden laut Datenschutzerklärung gegebenenfalls alle Informationen, die eingegeben werden bzw. zuvor eingegeben wurden, an den Verantwortlichen übermittelt. Als Zweck wird hierfür die Notwendigkeit zur Nutzung der Applikation angegeben. Ob und wie diese personenbezogenen Daten weiterverarbeitet werden, wird nicht erwähnt. In der Datenschutzerklärung hat die automatische Analyse als Tracker „Adjust “ identifiziert. Dieser ist laut statischer Analyse der Applikation[58] durch Exodus Privacy zwar tatsächlich in die Applikation integriert, allerdings wurde das Wort im Kontext (dt. einstellen bzw. anpassen) verwendet und repräsentiert in der Datenschutzerklärung daher nicht den Tracker. Neben Adjust wurden in der Analyse durch Exodus Privacy aber noch drei weitere Tracker erkannt, Facebook Analytics, Google CrashLytics und Google Firebase Analytics, welche ebenfalls in der Datenschutzerklärung Erwähnung finden müssen, sofern hierüber eine Verarbeitung persönlicher Daten stattfindet. Eine Überprüfung ist im Rahmen dieser Arbeit nicht möglich, bietet sich in Hinblick auf fehlende Inhalte in dieser Datenschutzerklärung aber an. Hierauf wird in Kapitel 6 weiterführend eingegangen.

[57] `https://play.google.com/store/apps/details?id=com.fontskeyboard.fonts`
[58] `https://reports.exodus-privacy.eu.org/de/reports/191495/`

Die automatisierte Analyse hat die betreffenden Testkriterien bis auf die fälschliche Identifikation eines Trackers korrekt bewertet, allerdings fiel in der manuellen Prüfung der Datenschutzerklärung auf, dass diese mit einer hohen Wahrscheinlichkeit Informationen zur Analyse des Nutzungsverhaltens, beispielsweise durch weitere Tracking-Module, vermissen lässt.

Oje, ich wachse!

Mit einer Wertung von 24 Punkten wird die englischsprachige Datenschutzerklärung der Baby-Applikation „Oje, ich wachse!“[59] in der automatisierten Analyse am schlechtesten bewertet. Sie ist nur sehr wenig strukturiert und hat keinen erkennbaren Bezug zur benannten Applikation, sondern ist auf zwei andere Applikationen („The Wonder Weeks“ und „Back to you“) bezogen. Inhaltlich greift die Datenschutzerklärung nicht auf übliche, sondern eher einfache Begriffe zurück. Dies macht sich auch im Lesbarkeitsindex bemerkbar, der mit 10,4 Bildungsjahren die volle Punktzahl dieses Kriteriums erreicht. Die Entropie spricht mit einem Wert von 5,5 für einen hohen Informationsgehalt, im Datenschutzkontext ist dem nach der automatischen Analyse aber nicht so, weshalb diesem Kriterium nur ein Punkt vergeben wurde.

In der Datenschutzerklärung wird in wenigen Worten dargelegt, dass Daten grundsätzlich nur auf dem Smartphone des Benutzers und nicht auf Servern des Verantwortlichen gespeichert werden. Neben drei Analyse- bzw. Trackingmodulen (Google Firebase, Fabric und Crashlytics) werden demnach sonst nur sehr wenige Informationen zur Datenverarbeitung gegeben. Viele weitere, nach der Datenschutzgrundverordnung notwendige Informationen fehlen, wie auch in der automatisierten Analyse korrekt dargestellt wird. Weiterhin fehlen Informationen zu den Rechten der Betroffenen, da in jedem Fall durch die Tracker eine Aufzeichnung personenbezogener Daten stattfindet, für welche entsprechende Rechte zur Verfügung gestellt werden müssen.

Da die Datenschutzerklärung keinen Bezug zur entsprechenden Applikation „Oje, ich wachse!“ aufweist, fällt die manuelle Analyse negativ aus. Auch wenn die Datenschutzerklärung zunächst den Anschein einer datenschutzfreundlichen Lösung erweckt, wird dennoch durch Drittanbieter das Nutzerverhalten analysiert. Im Unterschied dazu hat die automatisierte Analyse aufgrund der fehlenden Inhalte nur 24 Punkte vergeben, da mangels Textverständnis nicht erkannt wird, dass die Datenverarbeitung laut Datenschutzerklärung lokal auf dem Smartphone verbleibt. Die Datenschutzerklärung könnte aber durch einen höheren Grad der Transparenz, wie bei der zuerst benannten DRK Applikation, eine höhere Punktzahl erreichen.

[59] `https://play.google.com/store/apps/details?id=org.twisevictory.apps`

KWGT Kustom Widget Creator

Die Applikation „KWGT Kustom Widget Creator“[60] dient der Personalisierung des Erscheinungsbilds des Android Homescreens. Die englischsprachige Datenschutzerklärung der Applikation hat, wie in Tabelle 5.1 ersichtlich, in der automatisierten Analyse mit einer Punktzahl von 31 die zweitschlechteste Platzierung in der Massenanalyse erhalten. Sie ist mit einem Lesbarkeitsindex von 10,3 Bildungsjahren auf Realschulniveau und erhält daher in diesem Kriterium die maximal mögliche Punktzahl. Die Applikation kann auf dem Smartphone umfangreiche Berechtigungen anfordern, beispielsweise Kontakte, Kalendereinträge oder den aktuellen Standort erhalten. Demnach ist von der dazugehörigen Datenschutzerklärung zu erwarten, dass sie über eine Verarbeitung dieser Daten informiert. Wie Tabelle 5.1 entnehmbar ist, enthält die Datenschutzerklärung aber nur sehr wenige relevante Informationen. Dies bestätigt auch die manuelle Analye derselben. Die Datenschutzerklärung benennt in der Applikation integrierte Tracker wie Google Admob, Firebase Analytics und CrashLytics, aber nicht, welche Daten von diesen aufgezeichnet werden. Es werden keine Details zu verarbeiteten Daten gegeben oder darüber, welche dieser Daten mit Dritten geteilt werden. Gerade in Hinblick auf die durch die Berechtigungen der Applikation erfassbaren, personenbezogenen Daten, erforderlich wäre. Die automatisierte Analyse zeigt daher korrekt, dass zwar eine strukturierte und verständlich formulierte Datenschutzerklärung vorliegt, aber wichtige Informationen fehlen.

Die Einzelauswertung der vier automatisierten Analysen von Datenschutzerklärungen hat gezeigt, dass die auf Basis des Konzepts erstellte Lösung erfolgreich die Inhalte von Datenschutzerklärungen klassifiziert und aus den gewonnenen Daten eine zutreffende Bewertung ihrer Inhalte erstellt. Im folgenden Abschnitt werden Auswertungen gezeigt, die für die gesamte Datenbasis erstellt wurden.

5.2.2 Gesamtauswertung

Wie anfänglich in Abbildung 5.1 dargelegt, wurden im Rahmen der Massenanalyse Datenschutzerklärungen von vielfältigen Genres an Applikationen geprüft. Die im Durchschnitt erreichten Punkte jedes Genres sind in Abbildung 5.2 entsprechend dargestellt. Applikationen aus dem Bereich Sport, Lifestyle und sozialer Netzwerke erreichen im Durchschnitt die höchsten Bewertungen. Dies liegt zum Teil daran, dass größere Unternehmen des Datenverarbeitungssektors auch auf eigene Rechtsabteilungen mit entsprechendem Fachwissen zurückgreifen können. Applikationen mit geringerer Datenaufzeichnung, beispielsweise aus den schlechter gewerteten Genres wie Video-Playern oder der Personalisierung des Geräts, weisen teilweise auch eine geringere Datenaufzeichnung auf und informieren aus diesem Grund in ihrer Datenschutzerklärung nur über wenige Details zur Datenverarbeitung. Dass auch Applikationen mit wenigen aufgezeichneten, personenbezogenen Daten eine hohe Transparenz aufweisen können, hat die Einzelauswertung der DRK Applikation in Unterabschnitt 5.2.1 gezeigt. Im Durchschnitt umfassen die analysierten

[60] `https://play.google.com/store/apps/details?id=org.kustom.widget`

Datenschutzerklärungen etwa 4400 Worte, umgerechnet etwa acht bis neun A4-Seiten. Die durchschnittliche Lesbarkeit der Texte beträgt 13,3 erforderliche Bildungsjahre. Die getesteten Datenschutzerklärungen sind demnach im Mittel etwas über Abiturniveau geschrieben und demnach schwer verständlich.

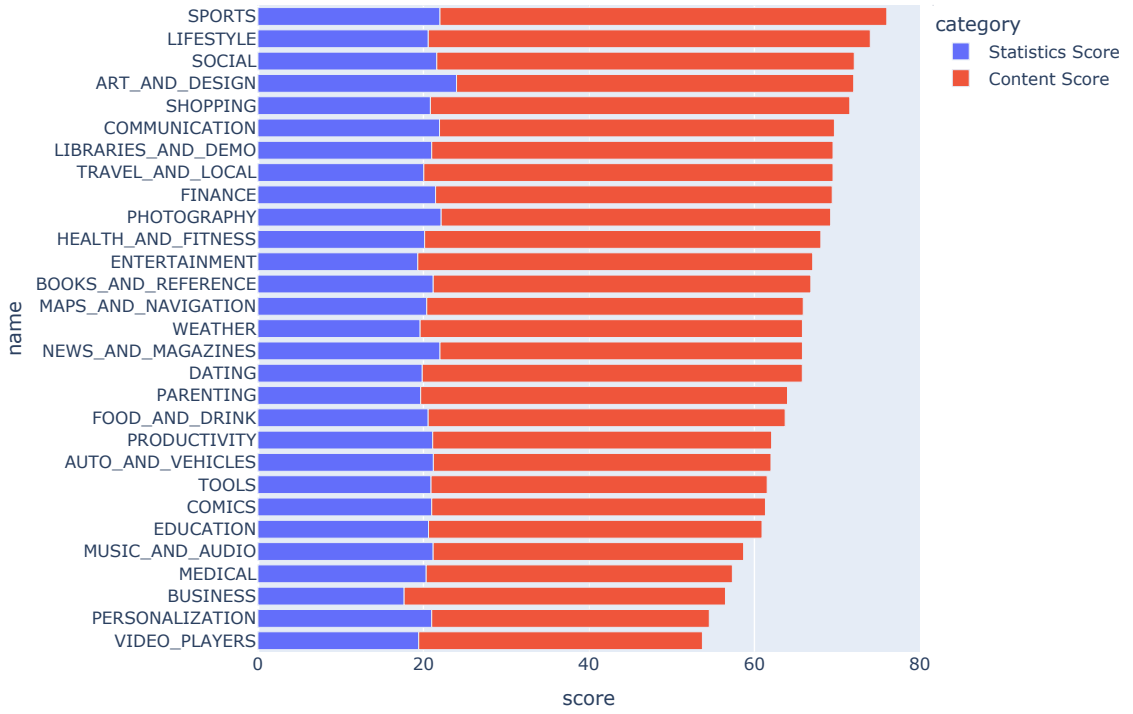

Abbildung 5.2: Durchschnittliche Punktzahl der Datenschutzerklärungen nach analysierten Genres der Applikation

Wie in Abschnitt 5.1 erwähnt, sind einige Applikationen Teil mehrerer Kategorien (Top, Bestseller, Höchster Umsatz). Diese werden im Rahmen von Gesamtauswertungen nur einmal einbezogen, in Vergleichen der drei Kategorien aber für jede Kategorie gezählt, in der sie vorkommen.

Serverstandorte

Wie in Abschnitt 3.4 beschrieben, wird der Standort des Servers, auf dem die Datenschutzerklärung gehostet wird, im Rahmen der Analyse erfasst. Dies ist kein Teil der Wertung, liefert aber einen Indikator für die Herkunft der Applikation. Etwa die Hälfte der deutsch- und ein Viertel der englischsprachigen Datenschutzerklärungen werden von deutschen Servern bereitgestellt. Dass deutschsprachige Datenschutzerklärungen, wie in Abbildung 5.3 ersichtlich, auch zu 29,2 % in den Vereinigten Staaten gehostet werden, begründet sich zumindest bedingt in DDoS-Schutz-Anbietern wie CloudFlare, auf die die Domains weitergeleitet werden.

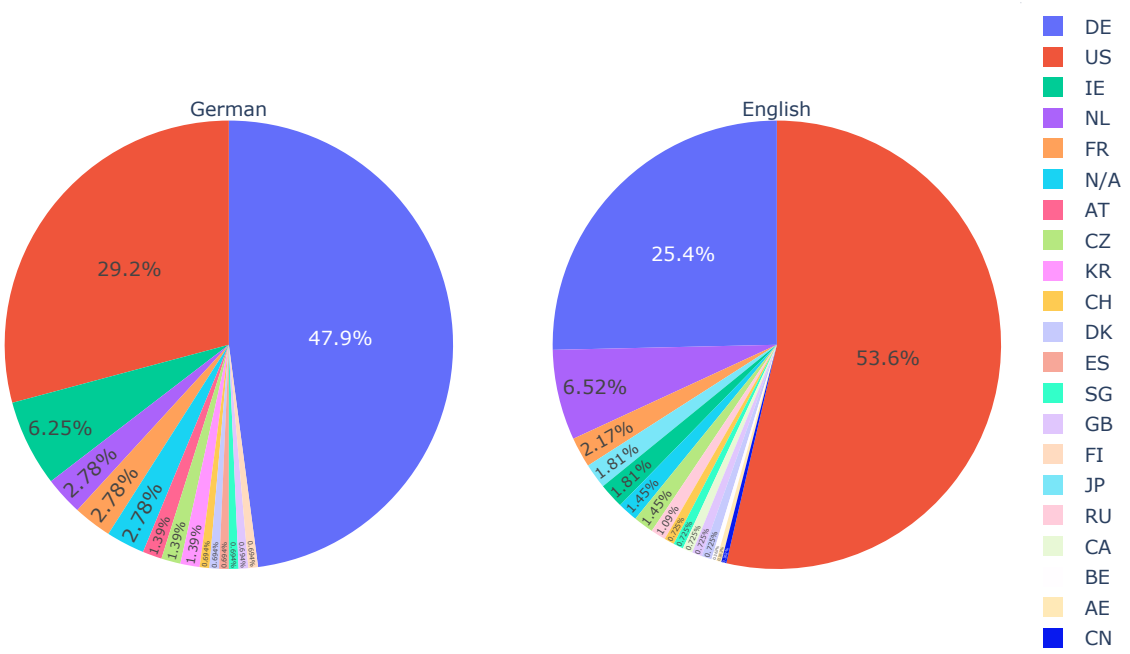

Abbildung 5.3: Serverstandort der Datenschutzerklärungen

Tracker

Welches Applikationsgenre durchschnittlich wie viele Tracker in der Datenschutzerklärung erwähnt, wird in Abbildung 5.4 gezeigt. Eine Analyse zu tatsächlich in den Applikationen integrierten Tracking-Modulen ist in dieser Arbeit nicht erfolgt und wurde in den Einzelauswertungen nur beispielhaft vorgenommen. Im Fazit und Ausblick (siehe Kapitel 6) wird dieser Punkt aber aufgegriffen. In der Grafik fällt auf, dass Wetter-Applikationen mit hohem Abstand die meisten Tracking-Module einzusetzen scheinen. Dies liegt vornehmlich an der Applikation „Wetter Online“[61]. Hier werden in der Datenschutzerklärung[62] mehr als 900 Werbepartner erwähnt, von denen 74 als Tracking-Anbieter in der Exodus Privacy Datenbank gelistet sind und daher durch die Analyse identifiziert wurden. Applikationen aus dem Bereich Business und Medizin haben laut ihrer Datenschutzerklärungen die wenigsten Tracker integriert. Dies kann ein Indikator für das unterschiedliche Geschäftsmodell sein, dass Applikationen verschiedener Genres aufweisen.

[61] `https://www.wetteronline.de/datenschutz/`
[62] Geprüft am 12.07.2021

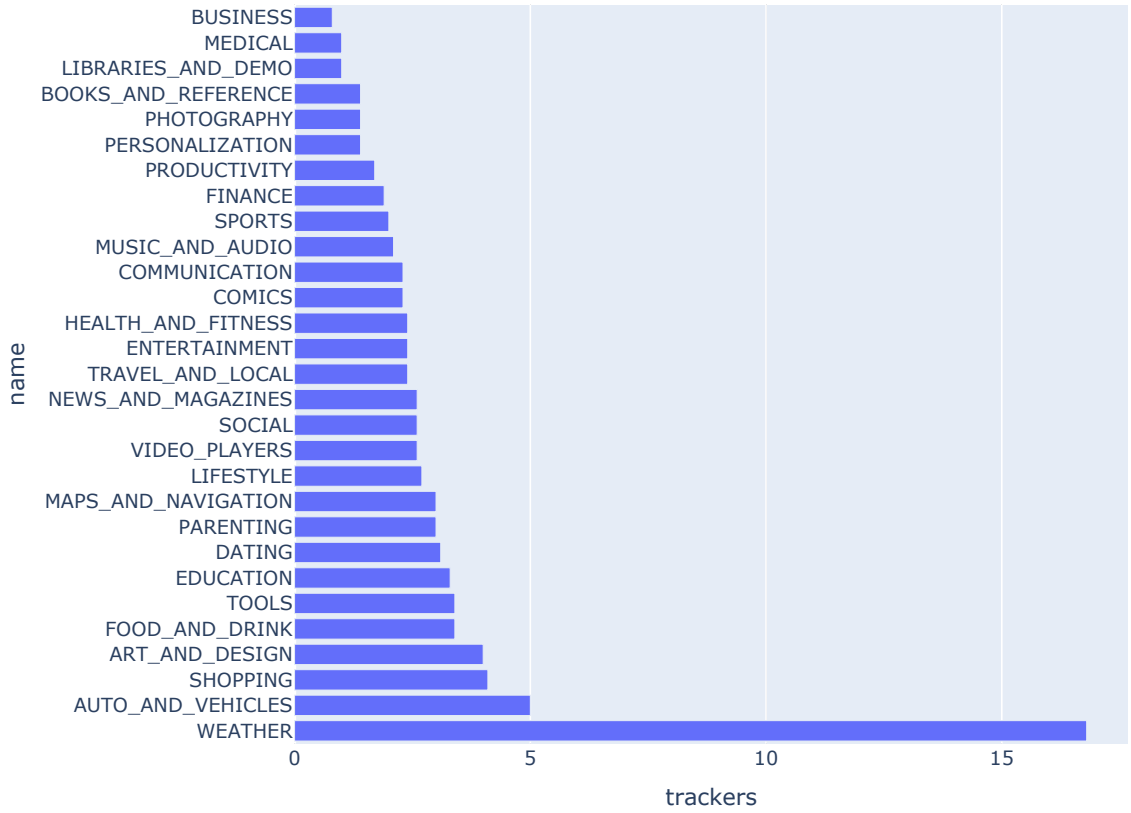

Abbildung 5.4: Durchschnittlich je Genre erwähnte Tracker

Im Vergleich kostenloser mit kostenpflichtigen Applikationen (siehe Abbildung 5.5) zeichnet sich ab, dass kostenlose Applikationen im Durchschnitt 3,3 Tracker erwähnen und damit im Schnitt ein Tracking-Modul mehr enthalten, als kostenpflichtige. Dies ist ein Indiz dafür, dass kostenlose Applikationen eher ein Interesse an der Nutzungs- und Verhaltensanalyse und damit einhergehender Monetarisierung der erfassten Daten haben, als dies bei kostenpflichtigen Produkten der Fall ist.

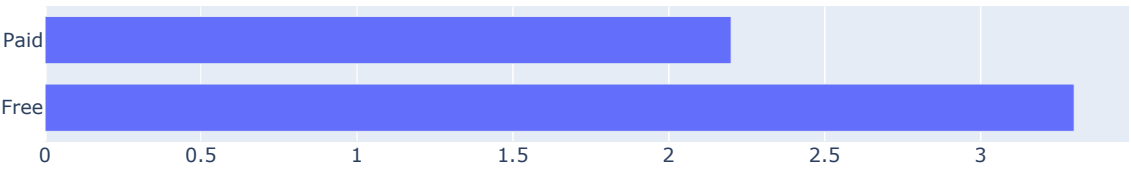

Abbildung 5.5: Durchschnittliche Anzahl erwähnter Tracker bezahlter und kostenloser Applikationen

Bewertung

In Abbildung 5.2 wurde bereits dargestellt, dass sich die erreichten Punktzahlen der Genres von Applikationen teilweise sehr stark unterscheiden. Hierbei wird ersichtlich, dass der Einsatzzweck der Applikation maßgeblichen Einfluss auf die Vollständigkeit ihrer Datenschutzerklärung hat und beispielsweise Applikationen aus dem Bereich der Sozialen Netzwerke, wie Discord (siehe Abschnitt 4.6), detaillierter informieren, als dies bei Applikationen im Bereich der Personalisierung, wie KWGT (siehe Unterabschnitt 5.2.1), der Fall ist. Allerdings werden in ersterer Kategorie auch deutlich mehr personenbezogene Daten verarbeitet.

Dass die erreichte Punktzahl sich auch zwischen kostenlosen und kostenpflichtigen Applikationen sowie den verschiedenen Kategorien (Top, Bestseller, Höchster Umsatz) unterscheidet, ist in Abbildung 5.6 und 5.7 ersichtlich. In der Monetarisierung von Applikationen gibt es verschiedene Strategien. Kostenlose Applikationen mit Abonnements bzw. In-App-Käufen zählen zu den verbreitetsten Ansätzen, da Benutzer häufig nicht dazu bereit sind, eine Applikation ohne vorheriges Ausprobieren zu kaufen (vgl. Torres, 2018). Dies ist auch in den analysierten Kategorien ersichtlich. Sowohl in der Kategorie Höchster Umsatz als auch Top sind nahezu ausschließlich kostenlose Applikationen enthalten. Gründe für die Unterschiede in der erreichten Punktzahl können darin liegen, dass kostenpflichtige Applikationen eher weniger Daten verarbeiten, daher aber auch Inhalte in der Datenschutzerklärung fehlen. Dass Applikationen mit nur geringer Datenverarbeitung auch eine Datenschutzerklärung mit hohem Informationsgrad besitzen können, hat die DRK Applikation in Unterabschnitt 5.2.1 gezeigt. Datenschutzfreundliche Verarbeitung und hoher Informationsgehalt schließen sich daher nicht aus.

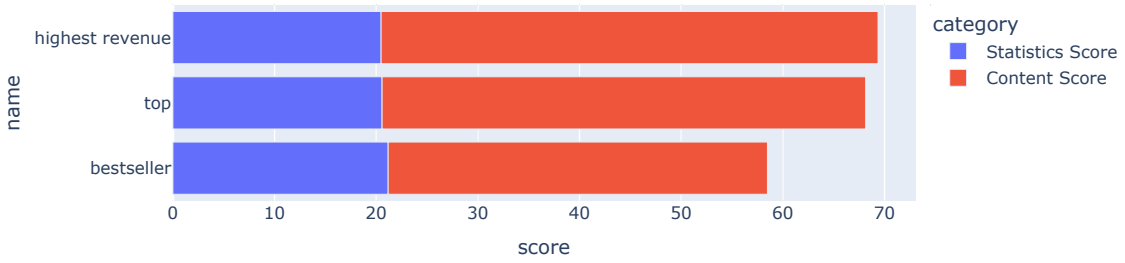

Abbildung 5.6: Durchschnittliche Punktzahl nach „Top“-Kategorie der Applikationen

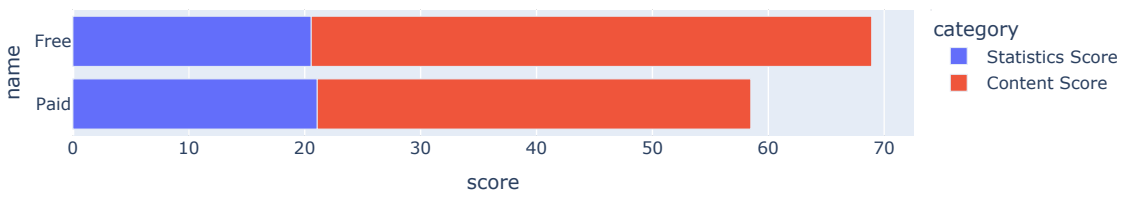

Abbildung 5.7: Durchschnittliche Punktzahl bezahlter und kostenloser Applikationen

6 Fazit und Ausblick

Nur ein geringer Teil der EU-Bürger ist, wie in Kapitel 1 erwähnt, dazu bereit, Datenschutzerklärungen zu lesen. Gründe hierfür liegen neben der Länge auch in der Komplexität der Texte. Die in dieser Arbeit entwickelte automatisierte Analyse hat zum Ziel, die Komplexität auf ein angemessenes Niveau herunterzubrechen. Es wird geprüft, ob die analysierte Datenschutzerklärung verständlich formuliert ist und über alle nach der Datenschutzgrundverordnung und geltender deutscher Datenschutzgesetze erforderlichen Sachverhalte informiert. Das Programm misst statistische Werte wie die Lesbarkeit, Struktur und den relevanten Informationsgehalt der analysierten Datenschutzerklärungen. Die Texte werden auf Basis vorhandener Formatierung in Themenabschnitte unterteilt und diesen mit einer Kategorisierung Stichworte zugeordnet. Mit einer Kombination der Stichworte und weiteren regulären Ausdrücken werden die Thematiken der einzelnen Absätze erkannt und die einzelnen Testkriterien beantwortet. Die Datenschutzerklärungen werden außerdem auf benannte Tracker hin untersucht und diese entsprechend benannt. Abschließend wird eine Bewertung der Datenschutzerklärung auf einer Skala von 0 - 100 Punkten vergeben. Hierdurch wird dem Benutzer eine erste Einschätzung der Datenschutzerklärung ermöglicht und in der Ausgabe fehlende Informationen aufgezeigt.

Die für die Entwicklung notwendigen Grundlagen aus den Themengebieten Datenschutz, Lesbarkeitsanalyse und Computerlinguistik wurden in Kapitel 2 ausführlich erläutert. Hierauf aufbauend wurde in Kapitel 3 das Konzept der automatisierten Analyse entwickelt. Hierzu wurden Testkriterien aufgestellt, deren Beantwortung in einer Bewertung der Datenschutzerklärung mündet. Nützliche Technologien wurden vorgestellt und in Kapitel 4 in einer entwickelten Programmstruktur implementiert. Weiterhin wurde in diesem Kapitel ein Bewertungsmaßstab erstellt, der aufgrund gegebenenfalls anpassbarer Gewichtung auf verschiedene Einsatzszenarien und Themenbereiche hin anpassbar und daher beispielsweise für Datenschutzerklärungen mit und ohne Applikationsbezug nutzbar ist. Für die Datenschutzerklärungen von mehr als 400 „Top“-Applikationen verschiedener Genres des Google Play Stores wurde eine automatisierte Analyse erfolgreich vorgenommen und die Ergebnisse der Massenanalyse in Kapitel 5 vorgestellt. Durch die Auswertung einzelner Datenschutzerklärung wurde die korrekte Funktionsweise des Programms punktuell validiert. Weiterhin wurden in Gesamtauswertungen Erkenntnisse über die gesamte Datenbasis gewonnen.

Die analysierten Datenschutzerklärungen waren im Durchschnitt acht bis neun A4-Seiten lang und erforderten zum Verständnis des Textes mit ca. 13 Bildungsjahren mehr als Abiturniveau. Demnach ist die Annahme der EU-Bürger, dass Datenschutzerklärungen lang und kompliziert formuliert sind, für die Datengrundlage der in dieser Arbeit vorgenommenen Massenanalyse zutreffend. Es wurden die Datenschutzerklärungen von Android Applikationen aus 29 Genres des Google Play Stores analysiert. Hierbei wurde festgestellt, dass die Bewertung unterschiedlicher Genres unterschiedlich hoch ausfiel, beispielsweise Applikationen aus dem Social Media Bereich

besser abschnitten als Video-Player. Eine Begründung hierfür wurde unter anderem in der Unternehmensgröße vermutet, sodass Social Media Unternehmen in der Regel auch über eine eigene Rechtsabteilung und somit eine höhere Kompetenz auf diesem Gebiet verfügen. Im Hinblick auf die in den Datenschutzerklärungen benannten Trackern wurde festgestellt, dass auch hier je nach Genre der Applikation eine unterschiedliche Anzahl erwähnt wurde. Demnach sind Medizin- und Business-Applikationen am wenigsten daran interessiert, das Nutzerverhalten zu analysieren. Auch zwischen bezahlten und kostenlosen Applikationen lagen in dem Bereich Unterschiede vor.

Die Ergebnisse der Massenanalyse zeigen, dass das entwickelte Konzept geeignet ist, Datenschutzerklärungen automatisiert zu analysieren. Durch die Analyse werden fehlende Informationen hervorgehoben und eine transparente und sprachlich einfach gehaltene Darstellung relevanter Informationen belohnt. Der gewählte Ansatz offenbarte in der Umsetzung und Massenanalyse aber auch Schwächen. Durch die Stichwortsuche ist die Analyse der Datenschutzerklärungen statisch und erfordert mangels Textverständnis eine gepflegte Liste an Suchworten der entsprechenden Kategorie. Wenn Datenschutzerklärungen, wie die der in Unterabschnitt 5.2.1 benannten Applikation „Oje, ich wachse“, nicht auf übliche Standardbegriffe zurückgreifen, wird einem möglicherweise guten Produkt eine schlechte Datenschutzerklärung vorgeworfen, obwohl nur eine geringe Datenverarbeitung vorliegt. Weiterhin dauerte eine Analyse im Durchschnitt mehr als fünf Minuten. Insbesondere vor dem Gesichtspunkt einer geplanten Mehrbenutzertauglichkeit ist hier eine Performancesteigerung erforderlich, beispielsweise durch erhöhte Nutzung des in Teilen integrierten Machine Learning Frameworks `spaCy`. Weiterhin bietet sich an, voneinander unabhängige Analyseschritte parallel auszuführen. Da in der aktuell gewählten Datenbanklösung keine parallelen Schreibzugriffe möglich sind, wird in Zukunft eine mehrbenutzerfähige Alternative angebunden werden. In der Massenanalyse zeigte sich, dass die unkomprimierte Speicherung aller Texte der Datenschutzerklärungen in verschiedenen Datenstrukturen einen hohen Speicherverbrauch aufweist. Dies wird in Zukunft vor dem Gesichtspunkt der Datenbankgröße durch Auslagerung auf das Dateisystem gelöst werden. Bei den vorgenommenen Analysen wurde nur auf in der Datenschutzerklärung benannte Tracking-Module hin geprüft, aber kein Abgleich mit der Applikation vorgenommen. Durch die Integration in ein Analysesystem für Android- und iOS-Applikationen wird dies angegangen. Hierbei wird die Analyse um tatsächlich von der Applikation bzw. integrierten Trackern erfasste Daten erweitert und eine erweiterte, inhaltliche Analyse erfolgen.

In Unterabschnitt 2.1.3 wurde bereits erwähnt, dass das Telemediengesetz zum 01.12.2021 durch das Telekommunikations-Telemedien-Datenschutzgesetz abgelöst wird. Sobald der finalisierte Gesetzesentwurf vorliegt, wird geprüft werden, ob die Testkriterien angepasst oder erweitert werden müssen. Als letzter Erweiterungspunkt ist geplant, das Programm um eine grafische Ausgabe bzw. einen PDF-Report zu ergänzen, in dem wesentliche Informationen in Kürze zusammengefasst sind, beispielsweise die Gründe für eine positive oder negative Wertung samt hierzu relevanter Textstellen.

Während der Umsetzung entstand auch die Idee für ein zukünftiges Forschungsprojekt: Ein Programm, das automatisiert Appstores durchsucht und nicht existierende bzw. Datenschutzerklärungen mit mangelhaftem Inhalt an die Betreiber, beispielsweise Google oder Apple, meldet.

Literaturverzeichnis

Ambros, Michael (2021). *Was sind personenbezogene Daten?* URL: https://www.datenschutz.org/personenbezogene-daten/ (besucht am 07. 07. 2021).

Amtsblatt der Europäischen Union, Hrsg. (2016). *DURCHFÜHRUNGSBESCHLUSS (EU) 2016 / 1250 DER KOMMISSION*. DOI: 10.1515/9783110924992-003. URL: https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32016D1250&from=DE (besucht am 07. 07. 2021).

Augsten, Stephan (2017). *Was ist YAML?* URL: https://www.dev-insider.de/was-ist-yaml-a-665391/ (besucht am 18. 06. 2021).

Augsten, Stephan (2019). *Was ist eine UUID?* URL: https://www.dev-insider.de/was-ist-eine-uuid-a-788491/ (besucht am 11. 06. 2021).

Behrens, Jan (2019). *Reguläre Ausdrücke*. URL: https://www.ionos.de/digitalguide/websites/webseiten-erstellen/regulaere-ausdruecke/ (besucht am 05. 07. 2021).

Bentz, Christian u. a. (2017). „The Entropy of Words—Learnability and Expressivity across More than 1000 Languages“. In: *Entropy* 19.6. DOI: 10.3390/e19060275.

Brownlee, Jason (2019). *A Gentle Introduction to Information Entropy*. URL: https://machinelearningmastery.com/what-is-information-entropy/ (besucht am 05. 07. 2021).

Bundeszentrale für politische Bildung (2017). *27. Januar 1977: Das Bundesdatenschutzgesetz wird verabschiedet*. URL: https://www.bpb.de/politik/hintergrund-aktuell/241406/bundesdatenschutzgesetz (besucht am 12. 06. 2021).

Carstensen, Kai-Uwe (2010). *Computerlinguistik und Sprachtechnologie: Eine Einführung*. 3., überarbeitete und erw. Aufl. Spektrum Lehrbuch. Heidelberg: Spektrum Akademischer Verlag. ISBN: 9783827422248. DOI: 10.1007/978-3-8274-2224-8. URL: http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10351853.

Datenschutzkonferenz (2018). *Zur Anwendbarkeit des TMG für nicht-öffentliche Stellen ab dem 25. Mai 2018*. Düsseldorf. URL: https://www.datenschutzkonferenz-online.de/media/ah/201804_ah_positionsbestimmung_tmg.pdf (besucht am 02. 06. 2021).

Diercks, Nina (2017). *Google Play verlangt jetzt Datenschutzerklärung für Apps bis zum 15.03! (was nach dem Gesetz und den bestehenden Entwicklerrichtlinien an sich keine neue Anforderung ist)*. URL: https://diercks-digital-recht.de/2017/02/google-play-verlangt-jetzt-datenschutzerklaerung-fuer-apps-bis-zum-15-03-was-nach-dem-gesetz-und-den-bestehenden-entwicklerrichtlinien-an-sich-keine-neue-anforderung-ist/ (besucht am 20. 07. 2021).

DuBay, William H. (2004). „The Principles of Readability“. In: URL: https://files.eric.ed.gov/fulltext/ED490073.pdf (besucht am 02. 07. 2021).

Ecma International (2017). *ECMA-404 - Ecma International: The JSON data interchange syntax*. URL: https://www.ecma-international.org/publications-and-standards/standards/ecma-404/ (besucht am 18. 06. 2021).

European Commission, Hrsg. (2019). *The General Data Protection Regulation: Special Eurobarometer 487a*. DOI: 10.2838/43726.

European Data Protection Supervisor (2021). *Entwicklungsgeschichte der Datenschutz-Grundverordnung*. URL: https://edps.europa.eu/data-protection/data-protection/legislation/history-general-data-protection-regulation_de (besucht am 15. 06. 2021).

Facebook Inc. (2021). *Language identification*. URL: https://fasttext.cc/docs/en/language-identification.html (besucht am 24. 07. 2021).

Fries, Maximilian und Orvil Keimig (2019). *Umfassende Lesbarkeitsanalyse von Datenschutzerklärungen für Apps*. DOI: 10.13140/RG.2.2.29799.68006.

Gesellschaft für Datenschutz und Datensicherheit e.V. (2020). *Handlungsempfehlungen der GDD: EuGH EU-US Privacy Shield und EU-Standard-vertragsklauseln*. URL: https://www.gdd.de/eu-us-privacy-shield-schrems-ii-urteil/handlungsempfehlungen-eugh-eu-us-privacy-shield-und-eu-standardvertragsklauseln (besucht am 07. 07. 2021).

Gropper, Adrian und Deborah Peel (2017). *Brandeis Privacy Award – Patient Privacy Rights*. URL: https://patientprivacyrights.org/brandeis-award/ (besucht am 12. 06. 2021).

Hancock, Richard und Peter Schmitz (2020). „EU-US Privacy Shield – was nun?“ In: *Security-Insider*. URL: https://www.security-insider.de/eu-us-privacy-shield-was-nun-a-959439/ (besucht am 07. 07. 2021).

Hoeren, Thomas (2007). „Das Telemediengesetz“. In: *Neue Juristische Wochenschrift*, S. 801–864. URL: https://web.archive.org/web/20091007171841/http://128.176.101.170/hoeren_veroeffentlichungen/telemediengesetz.pdf (besucht am 16. 06. 2021).

Ingelheim, Alexander und Dominik Fünkner (2021). *EU-DSGVO Gesetzestext im Wortlaut*. URL: https://www.datenschutzexperte.de/gesetzestext-eu-dsgvo/ (besucht am 24. 07. 2021).

Jähn-Nguyen, Jennifer (2021). *TTDSG – Gesetz mit Auswirkungen auf den Datenschutz? Ein erster Überblick*. URL: https://www.datenschutz-notizen.de/ttdsg-gesetz-mit-auswirkungen-auf-den-datenschutz-ein-erster-ueberblick-3630378/ (besucht am 23. 07. 2021).

Krösmann, Christoph und Anja Olsok (2020). *App-Boom setzt sich fort*. URL: https://www.bitkom.org/Presse/Presseinformation/App-Boom-setzt-sich-fort (besucht am 24. 07. 2021).

Lee, Jenny (2020). *Benchmarking Language Detection for NLP - Towards Data Science*. URL: https://towardsdatascience.com/benchmarking-language-detection-for-nlp-8250ea8b67c (besucht am 21. 06. 2021).

Luber, Stefan und Nico Litzel (2016). *Was ist Machine Learning?* URL: https://www.bigdata-insider.de/was-ist-machine-learning-a-592092/ (besucht am 05. 07. 2021).

Makai, Matt (2021). *Databases*. URL: https://www.fullstackpython.com/databases.html (besucht am 18. 06. 2021).

Marschall, Ursula (2016). *Was ist eigentlich der ”Eid des Hippokrates”?* URL: https://www.barmer.de/presse/infothek/newsletter-gesundheit-im-blick/presse-newsletter-archiv/archiv-2016/eid-des-hippokrates-40048 (besucht am 12. 06. 2021).

Milkaite, Ingrida und Eva Lievens (2019). *Status quo regarding the child's article 8 GDPR age of consent for data processing across the EU*. URL: https://www.betterinternetforkids.eu/practice/awareness/article?id=3017751 (besucht am 23. 07. 2021).

Möllers, Nils (2019a). *DSGVO und BDSG: Neuerungen und Zusammenhänge zum Datenschutz in 2020*. URL: https://keyed.de/blog/bdsg-und-dsgvo/ (besucht am 07. 07. 2021).

Möllers, Nils (2019b). *Technisch organisatorische Maßnahmen (TOM) – Datenschutz gemäß DSGVO*. URL: https://keyed.de/blog/tom-dsgvo/ (besucht am 24. 07. 2021).

Moos, Flemming, Jens Schefzig und Marian Arning (2018). *Die neue Datenschutzgrundverordnung: Mit Bundesdatenschutzgesetz 2018*. DOI: 10.1515/9783110338577.

Nafies, Ahmed (2020). *Why did we choose FAST API over Flask and Django for our RESTFUL Micro-services*. URL: https://ahmed-nafies.medium.com/why-did-we-choose-fast-api-over-flask-and-django-for-our-restful-micro-services-77589534c036 (besucht am 18. 06. 2021).

Ogden, Jacqueline von (2018). *GDPR Chapter Summaries: Part 2*. URL: https://www.cimcor.com/blog/gdpr-chapter-summaries-part-2 (besucht am 24. 07. 2021).

OpenAPI Initiative (2021). *OAI/OpenAPI-Specification*. URL: https://github.com/OAI/OpenAPI-Specification (besucht am 18. 06. 2021).

Palakollu, Sri Manikanta (2019). „Scrapy Vs Selenium Vs Beautiful Soup for Web Scraping“. In: *Medium*. URL: https://medium.com/analytics-vidhya/scrapy-vs-selenium-vs-beautiful-soup-for-web-scraping-24008b6c87b8 (besucht am 21. 06. 2021).

Red Hat Inc. (2021). *Was ist Docker?* URL: https://www.redhat.com/de/topics/containers/what-is-docker (besucht am 18. 06. 2021).

Reitz, Kenneth (2021). *Requests: HTTP for Humans*. URL: https://docs.python-requests.org/en/master/ (besucht am 11. 06. 2021).

Schallaböck, Jan (2019). *Was ist und wie funktioniert Webtracking?* URL: https://irights.info/artikel/was-ist-und-wie-funktioniert-webtracking/23386 (besucht am 22. 06. 2021).

Stefanovic, Valentina (2020). *Artikel 12 AEMR*. URL: https://www.humanrights.ch/de/ipf/grundlagen/rechtsquellen-instrumente/aemr/ (besucht am 12. 06. 2021).

Stetic GmbH (2021). *Browser Statistik: Marktanteile aller Browser*. URL: https://www.stetic.com/de/market-share/browser/ (besucht am 26. 07. 2021).

Stobitzer, Christian (2021). *Rechtsfähigkeit - Natürliche und Juristische Personen*. URL: https://www.wirtschafts-lehre.de/rechtsfaehigkeit.html (besucht am 16. 06. 2021).

Tao, Christopher (2020). „Do You Know Python Has A Built-In Database?“ In: *Towards Data Science*. URL: https://towardsdatascience.com/do-you-know-python-has-a-built-in-database-d553989c87bd (besucht am 18. 06. 2021).

The Data Incubator (2016). *NLTK vs. spaCy: Natural Language Processing in Python*. URL: https://blog.thedataincubator.com/2016/04/nltk-vs-spacy-natural-language-processing-in-python/ (besucht am 22. 06. 2021).

Torres, Magda (2018). *App Monetization Stats: Freemium vs Premium vs Paymium*. URL: http://thinkapps.com/blog/post-launch/paid-vs-freemium-app-monetization-statistics/ (besucht am 18. 07. 2021).

Ultimate Proofreader (2021). *Paragraph length in dissertations, essays: Ideal length of paragraph in academic writing*. URL: https://www.ultimateproofreader.co.uk/blog/paragraph-length-in-dissertations-essays (besucht am 08. 07. 2021).

Weltärztebund (2017). *Deklaration von Genf*. URL: https://www.bundesaerztekammer.de/fileadmin/user_upload/downloads/pdf-Ordner/International/Deklaration_von_Genf_DE_2017.pdf (besucht am 12. 06. 2021).

Witte, René und Jutta Mülle (2006). „Text Mining: Wissensgewinnung aus natürlichsprachigen Dokumenten“. Interner Bericht. Universität Karlsruhe. DOI: 10.5445/IR/1000005161.

Zamanian, Mostafa und Pooneh Heydari (2012). „Readability of Texts: State of the Art“. In: *Theory and Practice in Language Studies* 2.1. ISSN: 1799-2591. DOI: 10.4304/tpls.2.1.43-53.

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe. Sämtliche Stellen der Arbeit, die im Wortlaut oder dem Sinn nach Publikationen oder Vorträgen anderer Autoren entnommen sind, habe ich als solche kenntlich gemacht. Diese Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt oder anderweitig veröffentlicht.

Barleben, 2. August 2021 Dennis Henke

Nutzungs- und Verwertungsrechte

Barleben, 2. August 2021 Dennis Henke