Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Bachelorarbeit

Niklas Hagemann

Eine Untersuchung zum User-Tracking im deutschsprachigen Internet

Fakultät Technik und Informatik
Studiendepartment Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Niklas Hagemann

Eine Untersuchung zum User-Tracking im deutschsprachigen Internet

Bachelorarbeit eingereicht im Rahmen der Bachelorprüfung

im Studiengang Bachelor of Science Angewandte Informatik
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr.-Ing. Olaf Zukunft
Zweitgutachter: Prof. Dr. Klaus-Peter Kossakowski

Eingereicht am: 25. Februar 2019

Niklas Hagemann

Thema der Arbeit
Eine Untersuchung zum User-Tracking im deutschsprachigen Internet

Stichworte
Privatsphäre im Internet, User-Tracking, Datenschutz-Grundverordnung, Datenanalyse

Kurzzusammenfassung
Diese Arbeit stellt eine quantitative Untersuchung des User-Tracking auf deutschsprachigen Webseiten dar. Zu diesem Zweck wird eine Erhebung von 10 000 Webseiten durchgeführt und analysiert. Diese ergibt, dass deutschsprachige Webseiten nicht weniger Tracking einsetzen als nicht-deutschsprachige Webseiten. Webseiten öffentlich-rechtlicher Medien scheinen nicht grundsätzlich über weniger Tracker zu verfügen als Seiten privater Medien. Deutschsprachige Webseiten setzen zur Information des Nutzers über stattfindendes Tracking bisher selten auf eine Opt-in-Lösung. Zuletzt ist die Verbreitung spezifischer Tracking einsetzender Unternehmen auf deutschsprachigen und nicht-deutschsprachigen Webseiten ähnlich.

Niklas Hagemann

Title of the paper
A study of user tracking on German-language websites

Keywords
Web privacy, user tracking, General Data Protection Regulation, data analysis

Abstract
This thesis presents a quantitative study of user tracking on German-language websites. To this end a survey of 10 000 websites is conducted and analysed. It is found that German-language websites do not employ less tracking than websites not of German language. Public media websites do not generally seem to employ less tracking than private media websites. German-language websites rarely use the opt-in model to inform users of tracking. Lastly the prevalence of specific companies employing tracking is similar on German-language websites and those not of German language.

Inhaltsverzeichnis

1 Einleitung

In einer im Dezember 2017 in Deutschland durchgeführten Umfrage gaben 93 % der Befragten an, dass ihnen der Schutz ihrer persönlichen Daten wichtig sei. Gleichzeitig sagten 55 %, sie hätten das Gefühl, keine Kontrolle über ihre Daten im Internet zu haben. [31]

Diese Sorge ist nicht gänzlich unbegründet. So sind beispielsweise die Risiken einer allzu arglosen Nutzung von Social-Media-Plattformen an vielen Stellen diskutiert worden. Doch auch ohne eine aktive Nutzung sozialer Medien hinterlassen Internetnutzer Spuren. Auf nahezu allen Webseiten sammeln Unternehmen Daten über den Nutzer, ohne dass dieser das direkt mitbekäme. Unternehmen interessieren sich dafür, nach welchen Informationen ein Nutzer im Internet sucht, welche Webseiten er betrachtet, wie lange er dies tut, auf welche Verweise und Werbeanzeigen er klickt. Auf Basis dieser Aktivitätsverfolgung werden automatisiert Nutzerprofile erstellt, welche zum Beispiel für zielgerichtete Werbung eingesetzt werden können.

Diese Aktivitätsverfolgung im Internet, im Folgenden User-Tracking genannt, ist Gegenstand von Forschungsbemühungen. Untersucht werden zum einen die Techniken, die eingesetzt werden, um Nutzer im Internet zu verfolgen. Diese Techniken sind, wie das Internet selbst, im stetigen Wandel. Zum anderen wird untersucht, welche Konsequenzen das User-Tracking für Internetnutzer hat. Denn auch wenn die erhobenen Profile bisher primär für personalisierte Werbung eingesetzt werden, sind auch andere Verwendungszwecke denkbar. Dies gilt umso mehr, wenn es möglich ist die Profile konkreten Personen zuzuordnen.

Ergänzt werden diese Forschungsarbeiten durch quantitative Untersuchungen. Erfasst wird das Ausmaß des User-Tracking auf Webseiten unter verschiedenen Gesichtspunkten, wie etwa dem Verbreitungsgrad spezieller Tracking-Mechanismen oder der Abhängigkeit des Tracking von der Webseitenkategorie [8].

Bisher anscheinend unberücksichtigt geblieben ist die Frage, wie es um das User-Tracking auf Webseiten bestellt ist, die sich an ein deutschsprachiges Publikum richten. Die vorliegende Arbeit soll hierzu einen Beitrag leisten.

1.1 Zielsetzung

In dieser Arbeit soll das Ausmaß des User-Tracking auf deutschsprachigen Webseiten untersucht werden. Dazu sollen Webseiten automatisiert besucht und die erhobenen Daten derart analysiert und mit weiteren Informationen verknüpft werden, dass folgende Fragestellungen beantwortet werden können:

1. *Unterscheidet sich das User-Tracking auf deutschsprachigen Webseiten quantitativ von dem auf nicht-deutschsprachigen Webseiten?*

   Englehardt und Narayanan [8] konnten bei einer Untersuchung, in deren Rahmen die 1 000 000 beliebtesten Webseiten besucht wurden, feststellen, dass auf Webseiten durchschnittlich 19 nachverfolgende Drittanbieter eingebunden sind. Spätestens seitdem die Datenschutz-Grundverordnung anzuwenden ist, könnte vermutet werden, dass Webseiten, welche sich an ein deutschsprachiges Publikum wenden, weniger User-Tracking einsetzen als andere Webseiten. Als Arbeitshypothese wird formuliert, dass das User-Tracking auf deutschsprachigen Webseiten geringer ausfällt als auf nicht-deutschsprachigen Webseiten.

2. *Wird auf Webseiten öffentlich-rechtlicher Medien weniger User-Tracking eingesetzt als auf Webseiten privater Medien?*

   In der gleichen Untersuchung stellten Englehardt und Narayanan [8] fest, dass die Webseiten der Kategorie *News* durchschnittlich die meisten Tracker einbinden, während auf Webseiten aus Kategorien, welche primär staatliche und öffentliche Anbieter umfassen, nur wenige Tracker zu finden waren. Deutschland unterhält mit seinem öffentlich-rechtlichen Rundfunk Webseiten, die sich in beide Kategorien einsortieren ließen. Wo lassen sich die Webseiten öffentlich-rechtlicher Medien einordnen? Findet auf ihnen weniger Tracking statt als auf Webseiten privater Medien? Als Arbeitshypothese wird formuliert, dass das User-Tracking auf öffentlich-rechtlichen Webseiten geringer ausfällt als auf Webseiten privater Medien.

3. *Wie informieren Webseiten ihre Nutzer über den Einsatz von User-Tracking?*

   Die Datenschutz-Grundverordnung setzt der Verarbeitung personenbezogener Daten enge Grenzen. Dies wirkt sich auch auf die Zulässigkeit von User-Tracking aus. So gehen Voigt und von dem Bussche [35] davon aus, dass die Nachverfolgung des Nutzerverhaltens ohne dessen ausdrückliche Zustimmung nicht mehr rechtmäßig sei. Wie gestalten Webseiten den Hinweis auf die Nutzung von User-Tracking? Erhält der Nutzer tatsächlich die Möglichkeit zu widersprechen, wird er lediglich informiert oder findet Tracking gänzlich im Hintergrund statt? Setzen Webseiten, welche vorgeben, die Zustimmung des Nutzers einzuholen, vielleicht doch schon vorher identifizierende Cookies?

4. *Welche Unternehmen erheben auf deutschsprachigen Webseiten Daten über Nutzer?*

   Englehardt und Narayanan [8] konnten feststellen, dass eine Vielzahl von Unternehmen das Verhalten von Nutzern im Internet verfolgen. Gleichzeitig sind aber nur eine kleine Menge an Unternehmen auf vielen Webseiten vertreten. So waren lediglich sechs Unternehmen auf mindestens $10\,\%$ der besuchten Webseiten vertreten, wobei Google mit über $70\,\%$ mit weitem Abstand die weiteste Verbreitung hatte. Wie verbreitet sind die Unternehmen auf deutschsprachigen Webseiten? Gibt es Unternehmen, welche ausschließlich auf deutschsprachigen Webseiten vertreten sind?

1.2 Aufbau der Arbeit

In Kapitel 2 wird in die Thematik des User-Tracking eingeführt. Dabei wird sowohl beleuchtet, welche Ziele mit User-Tracking verfolgt werden und wie dieses realisiert wird, als auch die sich ergebende Datenschutzproblematik angesprochen und rechtliche Rahmenbedingungen zusammengefasst. Abschließend wird ein Überblick über aktuelle verwandte Arbeiten gegeben.

Darauf folgt in Kapitel 3 eine Darstellung der Untersuchungsmethodik. Dies beginnt mit der Gewinnung der Rohdaten, geht über die Transformation zur Verdichtung der enthaltenen Informationen und der Anreicherung der Daten mit Hilfe weiterer Informationsquellen und schließt mit den Methoden zur Aufbereitung und Analyse der Ergebnisse.

Nachfolgend werden in Kapitel 4 die Untersuchungsergebnisse dargestellt. Einer Diskussion der Ergebnisse widmet sich Kapitel 5. Dabei wird sowohl auf die methodischen Einschränkungen eingegangen als auch die Bedeutung der Ergebnisse für die formulierten Fragestellungen bewertet.

Abschließend werden in Kapitel 6 die Ergebnisse zusammengefasst und ein Ausblick auf mögliche zukünftige Arbeiten gegeben.

2 Datenschutz und Privatsphäre im Internet

In diesem Kapitel werden die Grundlagen für die nachfolgende Untersuchung dargelegt.

Abschnitt 2.1 führt den Begriff des User-Tracking grundlegend ein. Es werden die Ziele erörtert welche mit User-Tracking verfolgt werden sowie gängige Mechanismen zur Realisierung dargestellt. Anschließend werden in Abschnitt 2.2 die Datenschutzprobleme der Tracking-Praxis diskutiert und dargestellt, welchen rechtlichen Beschränkungen der Einsatz von Tracking unterliegt. Das Kapitel schließt in Abschnitt 2.3 mit einer Darstellung verwandter Arbeiten.

2.1 Ziele und Mechanismen des User-Tracking

2.1.1 Nutzerdaten als ökonomisches Gut

Online-Shopping hat in den letzten Jahren beständig an Relevanz gewonnen. Dem Statistischen Bundesamt zufolge haben im Jahr 2018 67 % der befragten deutschen Internetnutzer innerhalb der letzten drei Monate vor der Befragung das Internet zum Kauf von Waren oder Dienstleistungen verwendet [32]. Im Jahr 2008 lag der Wert noch bei 53 %, im Jahr 2003 erst bei 33 % [5]. In dem Maße in dem sich Einkäufe und Beschaffung von Produktinformationen in das Internet verlagern, steigt auch die Bedeutung von Online-Werbung. Eurostat gibt an, dass im Jahr 2016 89 % aller befragten deutschen Unternehmen über eine Webseite verfügt hätten. 47 % der Unternehmen würden soziale Medien zur Kundenkommunikation verwenden und 28 % würden Online-Werbung einsetzen [9].

Im Vergleich zu klassischen Werbeträgern wie dem Fernsehen oder Zeitungen können Anzeigen im Internet sehr viel zielgerichteter positioniert werden. Während bei der Fernsehwerbung

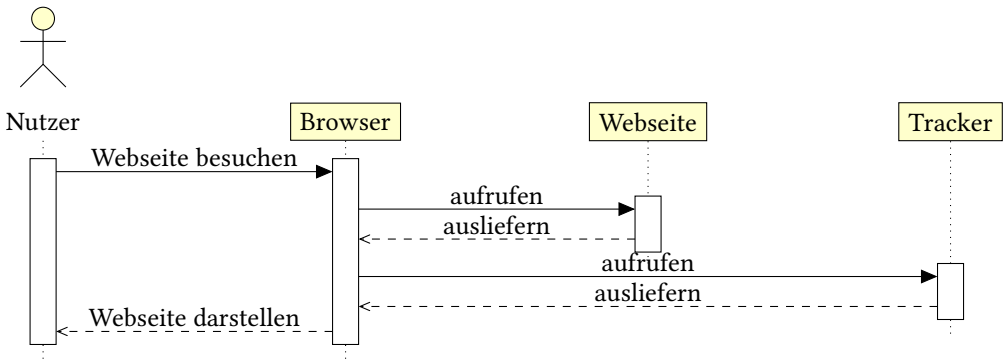

Abbildung 2.1: Grundprinzip des User-Tracking

der Werbetreibende lediglich über die Wahl von Parametern wie Sender und Sendezeit versuchen kann, das Werbepublikum in möglichst große Übereinstimmung mit seiner Zielgruppe zu bringen, kann er bei Werbeanbietern im Internet oftmals diverse Filteroptionen kombinieren. So bietet Facebook seinen Werbekunden die Möglichkeit das Werbepublikum nach Kriterien wie Standort, Alter, Geschlecht, Bildungsstand, Beziehungsstatus, Beruf, Interessen, Hobbys, Gerätenutzung oder Kaufverhalten auszuwählen [11]. Die dafür benötigten Daten stellen Facebook-Nutzer dem Unternehmen durch die Nutzung der Plattform einerseits direkt zur Verfügung. Auf der anderen Seite binden Werbetreibende einen sogenannten Tracker (Facebook bezeichnet diesen als *Facebook Pixel* [10]) in ihre Webseiten ein, mit dem Facebook Nutzeraktivitäten auch außerhalb ihrer Plattform verfolgen kann.

Abbildung 2.1 stellt das Grundprinzip dieser Aktivitätsverfolgung durch Dritte (engl. *third-party tracking*) dar. Ruft ein Nutzer eine Webseite auf, so enthält das gelieferte Dokument häufig Verweise zu weiteren Ressourcen wie Bildern. Diese werden durch den Browser automatisiert nachgeladen. Im Falle des Tracking sind dies in der Regel Skripte, welche dem Nutzer, sofern noch nicht vorhanden, eine eindeutige Kennung zuweisen und dem Tracking einsetzenden Unternehmen die besuchte Seite und Informationen über die Interaktion mit dieser mitteilen.

Aktivitätsverfolgung durch Dritte ist inzwischen weit verbreitet. Es sind eine Vielzahl von Unternehmen (im Folgenden als Tracking-Unternehmen bezeichnet) entstanden, die sich auf das Anlegen von Nutzerprofilen, das Vermarkten dieser Daten und von Werbeflächen auf Webseiten spezialisiert haben. Insbesondere große Internet-Unternehmen wie Facebook,

Google, AOL, Microsoft und Yahoo haben durch Übernahme kleinerer Unternehmen ihre Datensammlungen vergrößern und ihre Reichweite steigern können. [29]

Es ist im ökonomischen Interesse der Tracking-Unternehmen, dass ihre Tracker auf möglichst vielen und insbesondere auf reichweitenstarken Webseiten eingebunden werden. Jede weitere Webseite, die den Tracker des Unternehmens einbettet, erhöht die Chance das Surfverhalten eines Nutzers auch über Seitengrenzen hinweg verfolgen zu können (engl. *cross-domain tracking*). Dazu gehen Tracking-Unternehmen geschäftliche Kooperationen mit Webseiten ein. Für das Einbetten ihrer Tracker und das Bereitstellen von Werbeflächen entschädigen die Unternehmen Webseitenanbieter finanziell oder durch das Bereitstellen von Dienstleistungen wie Analytics-Diensten. Beispiele sind der bereits erwähnte *Facebook Pixel* oder *Google Analytics* [18].[29]

2.1.2 Techniken des User-Tracking

Entscheidend für das Funktionieren des User-Tracking ist das zuverlässige Wiedererkennen eines Nutzers. Das regelmäßige Aufkommen neuer Technologien zur Ausgestaltung des World Wide Web wie Cookies, JavaScript, Flash oder HTML5 hat in der Vergangenheit auch zur einer ständigen Weiterentwicklung der Tracking-Techniken beigetragen. Eine umfassende Darstellung bekannter Techniken findet sich bei Bujlow u. a. [3]. Um ein grundlegendes Verständnis zu vermitteln, soll es an dieser Stelle genügen, die zwei grundsätzlichen Varianten des Tracking, das *zustandsbehaftete* Tracking und das *zustandslose* Tracking, darzustellen und an Beispielen zu erläutern [29].

Zustandsbehaftetes Tracking

Beim zustandsbehafteten Tracking wird ein eindeutiger Identifikator auf dem Computer des Nutzers gespeichert. Dieser wird vom Tracker beim ersten Kontakt mit diesem vorgegeben und bei jedem weiteren Kontakt mitübertragen. Das bekannteste Beispiel für diese Form des Tracking ist die Verwendung von HTTP-Cookies.[29]

HTTP-Cookies sind domaingebundene Schlüssel-Wert-Paare. Sie können per HTTP-Kopfzeilen (Antwortkopfzeile `Set-Cookie` respektive Anfragekopfzeile `Cookie`) sowie per JavaScript-Aufruf gesetzt und ausgelesen werden. Cookies können mit einem Ablaufdatum versehen

werden. Wird kein Ablaufdatum gesetzt, so gelten sie nur für die laufende Sitzung und werden beim Schließen des Browsers gelöscht. [3]

Lange Lebenszeiten sind charakteristisch für Tracking-Cookies. In Experimenten konnte festgestellt werden, dass 90 % der Tracking-Cookies über Lebenszeiten von mehr als einem Tag verfügen, während dies nur bei 20 % der Non-Tracking-Cookies der Fall war. Darüber hinaus verfügen 80 % der Tracking-Cookies von Längen über 35 Zeichen, während dies nur bei 20 % der Non-Tracking-Cookies der Fall ist. Durch Kombination beider Heuristiken, ergänzt durch die Betrachtung der Gesamtlänge aller auf einer Domain gesetzten Cookie-Werte, welche durch das potentielle Aufteilen des Identifikators auf mehrere Cookies notwendig ist, konnten sehr gute Erkennungsraten erreicht werden. [3]

HTTP-Cookies können vom Nutzer gelöscht werden. Der Tracker würde einen neuen Identifikator vergeben, das bisherige Nutzerprofil könnte aber nicht fortgeschrieben werden. Durch zusätzlichen, zunächst redundanten Einsatz anderer Möglichkeiten der Speicherung des Identifikators ist es jedoch möglich, den HTTP-Cookie wiederherzustellen. Dass der Einsatz solcher als Evercookies oder Supercookies bezeichneten Techniken nicht bloß theoretisch ist, zeigt eine Studie von 2009, bei der 54 der 100 beliebtesten Webseiten Flash-Cookies verwendet haben, um gelöschte HTTP-Cookies wiederherzustellen. [3]

Zustandsloses Tracking

Beim zustandslosen Tracking, auch als Fingerprinting bezeichnet, vergibt der Tracker selbst keinen Identifikator, sondern versucht, den Computer anhand von Konfigurationsaspekten zu erkennen. Da ein solcher Wert allein in der Regel nicht eindeutig sind, werden verschiedene Aspekte gleichzeitig betrachtet. Verwendet werden können sowohl passiv aufgezeichnete als auch aktiv abgefragte Aspekte. [3]

Passiv aufgezeichnete Aspekte, die sich zur Identifizierung eignen, sind unter anderem die IP-Adresse, welche auch eine ungefähre Standortbestimmung des Nutzers ermöglicht, Browser, Browserversion und Betriebssystem, welche in der Regel mit der HTTP-Anfragekopfzeile `User-Agent` mitgeteilt werden, und die bevorzugten Sprachen, welche die Anfragekopfzeile `Accept-Language` enthält. [3]

Mit der Verwendung von JavaScript-Funktionen können auch aktiv zum Fingerprinting geeignete Informationen abgefragt werden. Ein aktuelleres Beispiel stellt die Verwendung der

HTML5-Audio-API dar. Diese kann verwendet werden, um ein Audiosignal zu erzeugen, von welchem anschließend ein Hash generiert wird. Auf verschiedenen Systemen erzeugte Signale können, abhängig von Hard- und Software, geringfügig voneinander abweichen, sodass der generierte Hash als Teil eines Fingerprints verwendet werden kann. [8]

Zustandsloses Tracking kann schwer zu erkennen sein und ist nicht gänzlich zu vermeiden. Während Cookies grundsätzlich löschbar sind, kann ein Nutzer dem zustandslosen Tracking nur durch das teilweise oder vollständige Deaktivieren von JavaScript oder durch Verschleierung, etwa durch das Setzen der Kopfzeile `User-Agent` auf einen häufig vorkommenden Wert, entgegentreten.

2.2 Datenschutzprobleme und rechtliche Rahmenbedingungen

2.2.1 Konsequenzen des User-Tracking

Das Wiedererkennen von Nutzern zur Anzeige personalisierter Werbung ist profitabel. Potentiell noch profitabler sind die erhobenen Nutzungsprofile, wenn es möglich ist die Daten mit persönlich identifizierenden Informationen zu verknüpfen, sodass die Nutzungsprofile konkreten Personen zuzuordnen sind. Tatsächlich können schon wenige Merkmale genügen, um eine Person eindeutig zu identifizieren. So konnte gezeigt werden, dass $87\,\%$ der US-Bevölkerung eindeutig über die Merkmalskombination Geschlecht, Geburtsdatum und Postleitzahl identifiziert werden können. [3]

Die Möglichkeit zur Identifizierung bietet sich vor allem Unternehmen, welche sowohl als First-Party agieren und als solche direkt persönliche Daten von Kunden erheben, als auch ihr Verhalten auf Webseiten Dritter verfolgen und über eine gewisse Reichweite verfügen [3]. Google bietet seinen Nutzern eine Vielzahl von Diensten wie E-Mail-Konten, Online-Office oder ein Zahlungssystem. Gleichzeitig finden sich Tracker von Google auf über $70\,\%$ der eine Millionen Seiten umfassenden Topliste des Unternehmens Alexa [8]. Damit ist Google in der Lage einen großen Teil des Surfverhaltens seiner Nutzer nachzuverfolgen.

Auch Tracking-Unternehmen, welche ausschließlich als Third-Party agieren, können in den Besitz potentiell identifizierender Informationen kommen. Zum einen ist es möglich, dass

First-Partys, welche über Daten eines Nutzers verfügen, diese an eingebettete Third-Party-Tracker verkaufen. Nutzerdaten werden dann beim Laden des Trackers übergeben und können so direkt mit dem Identifikator verknüpft werden. In anderen Fällen scheint die Weitergabe von Informationen unbeabsichtigt zu sein. Beim Laden des Trackers setzt der Browser in der Regel die HTTP-Kopfzeile `Referrer`, welche die Seite angibt von welcher die Ressource geladen wurde. Ist Nutzername oder E-Mail-Adresse Teil der URL der First-Party, so erhält auch die Third-Party diese Informationen über die Kopfzeile. [3]

Im Zweifelsfall genügt es, dass eine einzige den Tracker einbettende First-Party identifizierende Informationen teilt, um das bisherige und zukünftige mit einem Identifikator verknüpfte Nutzungsverhalten einer konkreten Person zuzuordnen. Wozu die aus den Daten gewonnenen Informationen jetzt oder in der Zukunft, eventuell nach erfolgtem Weiterverkauf, verwendet werden, ist für den Nutzer nicht transparent. Die folgenden Beispiele geben einen Eindruck, wie Tracking die Privatsphäre von Nutzern bedrohen oder zu wirtschaftlichen Nachteilen führen kann. Auch wenn die dafür verwendeten Daten wohl selten allein über das User-Tracking auf Webseiten gesammelt werden, ist davon auszugehen, dass dieses einen gewichtigen Anteil der Daten generiert [29].

- Auskunftsdienste wie Rapleaf bieten die Möglichkeit gegen Bezahlung Informationen über beliebige Personen zu erhalten. Abfragbare Merkmale sind unter anderem Geschlecht, Haushaltseinkommen, Familienstand, Kinder, Ausbildung, Hobbys, Interessen und Immobilienwert. Der Dienst ist für jedermann zugänglich, die Kosten für Abfragen liegen im Cent-Bereich. [29]
- Tracking-Daten werden verwendet, um die finanzielle Situation potentieller Kunden einzuschätzen. Abhängig von der angenommenen Zahlungskraft und -bereitschaft werden automatisch verschiedene Preise oder Konditionen vorgeschlagen. So wurde festgestellt, dass Kreditkartenanbieter sich auf die Berechnungen von Tracking-Unternehmen stützen, um den anzubietenden Zinssatz zu berechnen. Ein Tracking-Unternehmen gibt an, tausende von Informationen über einen einzelnen Nutzer zur Bewertung der Kreditwürdigkeit heranziehen zu können. [3] Ein gemeinsames Projekt der Schufa und des Hasso-Plattner-Instituts, in welchem untersucht werden sollte, wie Daten aus sozialen Netzwerken und anderen Internetquellen zur Beurteilung der Kreditwürdigkeit von Personen verwendet werden können, wurde nach einer Welle der Kritik nicht durchgeführt [29].

2.2.2 Anforderungen der DSGVO

Der Einsatz von User-Tracking unterliegt rechtlichen Beschränkungen. Ohne sich explizit auf das User-Tracking zu beziehen, werden diese Beschränkungen in der Europäischen Union seit dem 25.05.2018 durch die Datenschutz-Grundverordnung (DSGVO) vorgegeben.

Die DSGVO regelt die Verarbeitung personenbezogener Daten. Der Begriff der Verarbeitung umfasst praktisch jeden systematischen Umgang mit Daten. Daten werden als personenbezogen bezeichnet, wenn sie sich auf eine identifizierte oder identifizierbare Personen beziehen. Identifizierbarkeit ist dann gegeben, wenn die Kombination verschiedener Informationen, die für sich allein keinen Rückschluss auf eine Person zulassen, eine Identifizierung ermöglichen. Dies kann für den Bereich des User-Tracking relevante Merkmale wie IP-Adressen und Cookies einschließen. [35]

Anwendung findet die DSGVO bei der Verarbeitung personenbezogener Daten im Rahmen der Tätigkeiten einer EU-Niederlassung des Verantwortlichen oder Auftragsverarbeiters. Überraschender und mit Relevanz für das User-Tracking ist die Regelung, dass die DSGVO auch stets anwendbar ist, wenn eine Datenverarbeitung damit im Zusammenhang steht, das in der EU stattfindende Verhalten von Personen zu beobachten, unabhängig vom Standort des Verantwortlichen. Damit ist die DSGVO grundsätzlich auch anwendbar auf außereuropäische Betreiber von Webseiten, die sich nicht explizit an ein europäisches Publikum wenden, jedoch Techniken des User-Tracking verwenden, um über diese Kenntnisse zu erhalten. [35]

In der Folge waren nach dem 25.05.2018 die Webseiten diverser US-Zeitungen aus der Europäischen Union nicht mehr abrufbar [2].

Jede Verarbeitung persönlicher Daten setzt die vorherige Einwilligung der betroffenen Person oder das Vorliegen eines anderen Erlaubnistatbestands voraus. Einwilligung heißt, dass eine eindeutige bestätigende Handlung der betroffenen Person erfolgen muss. Im Kontext der User-Tracking hieße dies, dass beispielsweise eine nicht bereits aktivierte Checkbox angeklickt werden müsste. Voigt und von dem Bussche [35] gehen davon aus, dass das Opt-out-Modell, welches die Einwilligung des Nutzers standardmäßig voraussetzt und ihm die Möglichkeit zum Widerspruch gibt, unzulässig ist.

In einem Positionspapier vertritt auch die Datenschutzkonferenz die Auffassung, dass die Anwendung von Tracking-Mechanismen wie das Setzen entsprechender Cookies ohne die eindeutige vorherige Einwilligung des Betroffenen unzulässig sei. Dem widerspricht die

Gesellschaft für Datenschutz und Datensicherheit und verweist darauf, dass Werbung nach der DSGVO ein berechtigtes Interesse darstelle und daher nicht von einer Einwilligung abhängig sei. Pseudonymisiertes Tracking greife weniger stark in das Persönlichkeitsrecht ein als direkte Kundenansprache und müsse daher ebenfalls prinzipiell zulässig sein. [23]

2.3 Verwandte Arbeiten

Während keine Arbeiten gefunden werden konnten, die sich explizit mit dem User-Tracking auf deutschsprachigen Webseiten beschäftigen, gibt es einige Arbeiten, welche ähnliche Aspekte untersuchen.

Englehardt und Narayanan [8] konnten in einer Untersuchung, in welcher sie eine Million Webseiten besuchten, feststellen, dass eine Vielzahl von Tracking-Unternehmen das Nutzerverhalten als Third-Party verfolgt, allerdings nur wenige auf vielen Webseiten vertreten sind. So waren lediglich sechs Unternehmen auf mindestens $10\,\%$ der besuchten Webseiten vertreten, wobei Google mit über $70\,\%$ mit weitem Abstand die weiteste Verbreitung hatte. Ferner konnten sie feststellen, dass die Anzahl eingebundener Tracker mit der Webseitenkategorie variiert. Auf Nachrichtenseiten fand mit durchschnittlich 35 Trackern deutlich mehr Nutzerverfolgung statt als auf Wissenschaftsseiten mit durchschnittlich 9 Trackern. Der Durchschnitt aller besuchten Seiten lag bei 19 Trackern.

Um festzustellen, ob der geographische Standort des Nutzers einen Einfluss auf das Ausmaß des User-Tracking hat, besuchten Fruchter u. a. [13] die 500 beliebtesten Internetseiten jeweils von den Standorten Deutschland, USA, Japan und Australien aus. Dabei konnte kein signifikanter Einfluss des Nutzerstandortes festgestellt werden. Ferner besuchten sie die jeweils 250 beliebtesten Webseiten der genannten Länder. Dabei konnten sie feststellen, dass auf den beliebtesten Seiten der USA signifikant mehr Third-Party-Cookies gesetzt werden als auf den beliebtesten Seiten der anderen Länder, woraus auf ein höheres Tracking geschlossen wird. Zu kritisieren ist an dieser Stelle, dass die Anzahl der Third-Party-Cookies nicht zwangsläufig Aufschluss über das Ausmaß des Tracking gibt. Der Einsatz einer Heuristik zur Identifizierung von Tracking-Cookies kann die Aussagekraft derartiger Ergebnisse erhöhen.

Libert u. a. [24] beschäftigten sich mit den Auswirkungen der DSGVO auf die Anzahl von Third-Partys auf europäischen Nachrichtenseiten. Dazu wurden die Seiten im April und Juli

2018 besucht und entsprechende Kenngrößen verglichen. Dabei konnte festgestellt werden, dass die Anzahl von Third-Party-Cookies im Juli im Vergleich zum April um 22 % gesunken war. Eine Heuristik zur Identifizierung von Tracking-Cookies wurde allerdings auch hier nicht verwendet. Darüber hinaus sank der Anteil an Webseiten, welche Social-Media-Inhalte wie Sharing-Buttons verwendeten, von 84 % auf 77 %. Die Verbreitung spezieller Tracking-Unternehmen veränderte sich ungleichmäßig. Während sich die Verbreitung von Google nur geringfügig verringerte (April: 97 %, Juli: 96 %), war Oracle auf deutlich weniger Webseiten vertreten (April: 53 %, Juli: 32 %).

3 Methodik der Untersuchung

In diesem Kapitel wird die Methodik der Untersuchung dargelegt.

Abschnitt 3.1 beschreibt, wie die Daten erhoben werden, auf Grundlage welcher die Fragestellungen beantwortet werden sollen. Darauf wird in Abschnitt 3.2 dargestellt, wie aus den Rohdaten erste Informationen extrahiert werden und mit welchen weiteren Informationen die Datenbasis angereichert wird. Schließlich fasst Abschnitt 3.3 für die Analyse benötigte Methoden zusammen.

3.1 Datenerhebung

3.1.1 Auswahl der zu besuchenden Webseiten

```
1,google.com
2,youtube.com
3,facebook.com
4,baidu.com
5,wikipedia.org
6,reddit.com
7,yahoo.com
8,qq.com
9,taobao.com
10,amazon.com
```

Abbildung 3.1: Die ersten zehn Einträge der Alexa-Liste vom 11.06.2018

Die zu besuchenden Webseiten werden, wie in verwandten Arbeiten zuvor, der Liste der 1 000 000 global beliebtesten Webseiten des Unternehmens Alexa entnommen [vgl. 8, 13]. Der einer Webseite zugewiesene Rang wird über ein Dreimonatsfenster aus der Anzahl der Besucher (*Unique Visitors*) und der Anzahl der Besuche (*Pageviews*) bestimmt [1]. Die Stichprobe gibt so einen Einblick in das Ausmaß des Tracking in dem häufig frequentierten Teil des World Wide Web.

Verwendet wird die Liste vom 11. Juni 2018. Abbildung 3.1 stellt die ersten zehn Einträge der Liste dar.

Nicht alle in der Liste enthaltenen Domains sind zum direkten Besuch durch Menschen vorgesehen. Dies beinhaltet beispielsweise Domains die zur Auslieferung von Inhalten im Rahmen eines Content Delivery Networks (CDN) genutzt werden. Bei Besuch einer solchen Domain wird häufig eine Fehlermeldung angezeigt. Um eine potentielle Verzerrung der Ergebnisse zu reduzieren, sollen CDN-Domains möglichst nicht besucht werden. Die Filterung von CDN-Domains erfolgt in dieser Arbeit mit dem Prädikat in Algorithmus 1.

Algorithmus 1 Filtern von CDN-Domains

Eingabe: Domainname `domain`
Ausgabe: **true**, wenn CDN-Domain. **false**, sonst.
return `domain` = `t.co`
or `domain` = `googleusercontent.com`
or `domain` enthält `cdn`
or Subdomains von `domain` enthalten `static`
or Second-Level-Anteil von `domain` endet mit `static`

Die Liste von Alexa enthält Domains, für das Crawling werden jedoch URLs benötigt. Einige der Domains lösen zudem nicht direkt auf, sondern nur unter Verwendung der Subdomain www. Es ist daher notwendig zu prüfen, ob für die Domain oder für mit dem Präfix www erweiterte Domain ein A-Eintrag im DNS existiert. Für die Namensauflösung werden hier die Nameserver `1.1.1.1` von Cloudflare und `8.8.8.8` von Google verwendet.

Algorithmus 2 Ermittlung der zu besuchenden URL

Eingabe: Domainname `domain`
Ausgabe: Zu besuchende URL oder `null`
if `domain` ist CDN-Domain nach Algorithmus 1 **then**
return `null`
end if
if `domain` hat A-Eintrag **then**
return „`http://`“ + `domain`
end if
if Subdomain www von `domain` hat A-Eintrag **then**
return „`http://www.`“ + `domain`
end if
return `null`

Die Bestimmung der zu besuchenden URL aus einer Domain erfolgt nach Algorithmus 2. In dieser Arbeit werden die ersten 10 000 Webseiten der Alexa-Liste besucht, welche unter Anwendung des Algorithmus ein Ergebnis verschieden von `null` liefern.

3.1.2 Crawling der Webseiten

Für das Besuchen der Webseiten wird die Software OpenWPM verwendet. OpenWPM automatisiert das Besuchen von Webseiten durch Verwendung des Testframeworks Selenium zur Instrumentation des Browsers Firefox [17]. Die Software wurde inzwischen in über 30 Studien zur Datenerhebung verwendet [7].

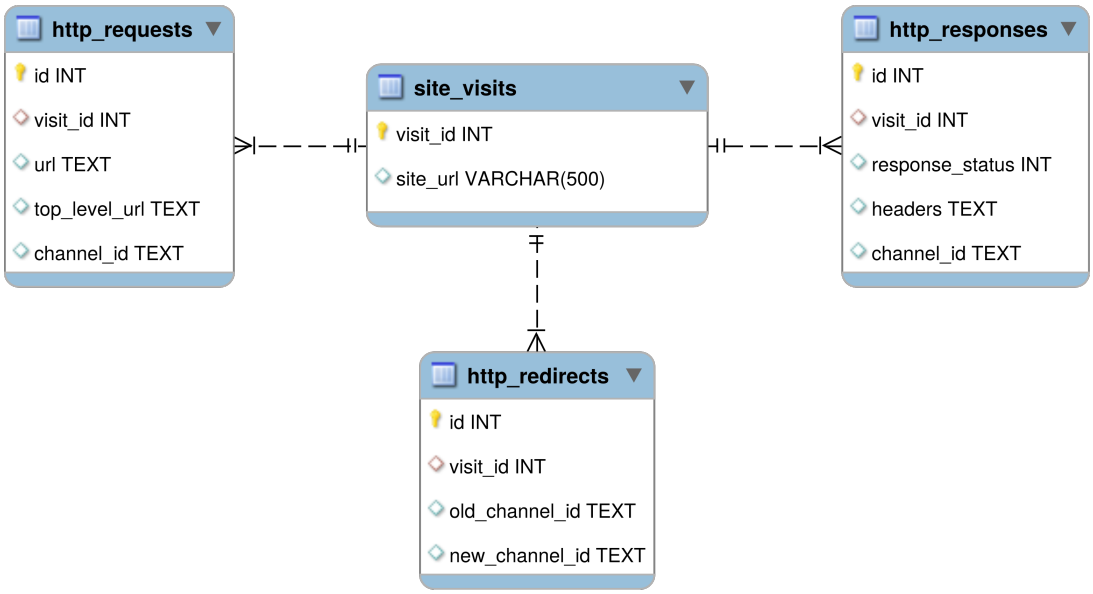

Abbildung 3.2: Datenbankschema von OpenWPM (Auszug)

OpenWPM erfasst die Daten eines jeden Besuchs in einer SQLite-Datenbank, deren Schema in Abbildung 3.2 auszugsweise dargestellt ist. Jede Anfrage, das heißt sowohl die initiale Anfrage durch das Laden der Webseite, als auch das Nachladen externer Ressourcen, wird in der Tabelle `http_requests` erfasst. Zu jeder erfolgreichen Anfrage existiert ein Eintrag in der Tabelle `http_responses`. Anfrage und Antwort können über das Attribut `channel_id` miteinander verknüpft werden. Umleitungen, d.h. Antworten mit HTTP-Statuscode `3xx`, werden in der Tabelle `http_redirects` erfasst.

Den Quelltext einer Webseite speichert OpenWPM, zusammen mit den Quelltexten eingebetteter Frames, in einem JSON-Dokument auf dem Dateisystem.

Um die Messungen voneinander unabhängig zu machen, wird in dieser Untersuchung OpenWPM so konfiguriert, dass nach jedem erfolgten Besuch das Browser-Profil zurückgesetzt wird. Darüber hinaus wird die HTTP-Anfragekopfzeile `Accept-Language` auf `de-DE,de` eingestellt, um anzuzeigen, dass die deutsche Sprachversion einer Webseite bevorzugt wird. Zudem wurden die Besuche von Hamburg aus durchgeführt, da auch der Standort des Nutzers einen Einfluss auf die Sprache einer Webseite haben kann.

3.1.3 Information über Tracking und Setzen von Cookies

In einem Teil der Untersuchung soll betrachtet werden, in welcher Beziehung das Setzen von Cookies und das Informieren durch die besuchte Webseite darüber stehen. Dazu werden stichprobenartig deutschsprachige Webseiten besucht und deren Information über den Einsatz von Tracking klassifiziert und überprüft, ob Tracking-Cookies gesetzt werden. Da die Stichprobe aus der Menge der besuchten und als deutschsprachig klassifizierten Webseiten gezogen wird, ist diese Datenerhebung dem anderen Teil der Untersuchung nachgelagert.

Zur Beurteilung, ob es sich bei einem gesetzten Cookie um einen Tracker handelt, werden die Kriterien aus Abschnitt 2.1.2 angewendet. Geprüft wird, ob die besuchte Webseite oder eine eingebettete Third-Party einen Cookie setzt, der den Kriterien entspricht.

Die auf Webseiten angezeigten Informationen werden in vier Klassen eingruppiert:

Keine Information Es wird in keiner Weise über den Einsatz von Tracking-Mechanismen informiert.

Informationstext/„Cookie-Banner“ Es wird über den Einsatz von Tracking-Mechanismen informiert. Der Nutzer stimmt durch Besuch der Seite dem Tracking zu. Opt-out-Möglichkeiten beschränken sich allenfalls auf den Hinweis, dass Cookies in der Browsereinstellungen deaktiviert werden können.

Opt-out-Modell Es wird über den Einsatz von Tracking-Mechanismen informiert. Standardmäßig wird von der Zustimmung des Nutzers ausgegangen, dem Nutzer wird aber über entsprechende Schaltflächen die Möglichkeit gegeben, zu widersprechen.

Opt-in-Modell Nutzer müssen dem Einsatz von Tracking-Mechanismen explizit zustimmen.

3.2 Transformation und Anreicherung der Rohdaten

3.2.1 Erkennung von Trackern

Die Messung des Tracking erfolgt, wie in verwandten Arbeiten zuvor, unter Verwendung einer Trackingschutzliste [vgl. 8, 13]. Die in einer Trackingschutzliste enthaltenen Filterregeln werden auf die in der Tabelle `http_requests` erfassten URLs angewandt. Auf diese Weise kann die Anzahl eingebetteter Tracker zu jedem Seitenbesuch bestimmt werden.

Der Vorteil der Verwendung von Trackingschutzlisten zum Erkennen von Trackern ist die Unabhängigkeit von der verwendeten Tracking-Technik. Wie in Abschnitt 2.1.2 erläutert, ist es möglich Tracking-Cookies zuverlässig zu erkennen. Diese stellen aber selbst unter den zustandsbehafteten Tracking-Techniken nur eine Möglichkeit der Realisierung dar, wenn auch die verbreiteste. Tracking-Schutzlisten erfassen Ressourcen, welche sicher oder mit hoher Wahrscheinlichkeit Tracking einsetzen, unabhängig von der technischen Realisierung. Problematisch für Trackingschutzlisten ist der beständige Wandel des World Wide Web und das Aufkommen neuer Tracking-Techniken. Sie werden daher nicht alle Tracker als solche erkennen und bedürfen ständiger Wartung.

Die verwendete Trackingschutzliste stammt vom *EasyList*-Projekt. Dieses verwaltet die nach dem Projekt benannte Hauptliste sowie die Liste *EasyPrivacy*. Filterregeln der EasyList müssen nach den Projektrichtlinien die Anzeige von Werbung blockieren, Tracking wird allenfalls als Nebeneffekt blockiert. Da Werbung nicht notwendigerweise Tracking-Methoden verwenden muss, würde eine Verwendung an dieser Stelle voraussichtlich zu vielen falsch positiven Klassifizierungen führen. Sie wird hier daher im Unterschied zu den Untersuchungen in [8, 13] nicht verwendet und lediglich auf EasyPrivacy zurückgegriffen, deren Filterregeln Tracking blockieren müssen [6].

Bisherige Untersuchungen haben sich auf die Messung des Tracking durch Third-Partys beschränkt. Dies ist einerseits plausibel, da Cookies von First-Partys dazu verwendet werden können, um Sitzungen zu verwalten. Dies kann insbesondere beim Online-Shopping erwünscht sein. Andererseits ist fragwürdig, ob auch der Einsatz zustandsloser Trackingmechanismen durch First-Partys als legitim zu bewerten ist. Die Richtlinie von EasyPrivacy ist,

Tracking durch First-Partys nur zu blockieren, wenn sie eine beträchtliche Menge persönlicher Daten erheben [6]. Die Verwendung von EasyPrivacy ermöglicht es daher an dieser Stelle auch als inakzeptabel eingestuftes Tracking durch First-Partys zu messen.

Für das Parsen der Filterregeln kommt die Adblock-Engine des Browsers *Brave* zum Einsatz [15].

3.2.2 Sprachklassifikation von Webseiten

Da Informationen über die Sprache der zu besuchenden Webseiten nicht von vornherein vorliegen, sollen diese aus den erhobenen Daten selbst extrahiert werden. Sowohl das Protokoll HTTP als auch die Auszeichnungssprache HTML bieten Möglichkeiten Metadaten zu erfassen und so die Sprache des Zielpublikums oder des Dokuments auszuzeichnen.

HTTP ermöglicht mit der optionalen Antwort-Kopfzeile `Content-Language` die Angabe der Sprachen des Zielpublikums. Das Feld erfasst also nicht notwendigerweise die Sprache des Textes. Bei einem Englisch-Sprachkurs, der sich an Deutschsprachige richtet, würde das Feld beispielsweise auf `de` gesetzt. Bei mehrsprachigen Seiten werden die Sprachkürzel mit Kommata voneinander getrennt. [12]

HTML bietet mit dem optionalen Attribut `lang` die Möglichkeit die inhaltliche Sprache beliebiger HTML-Elemente zu spezifizieren. Die Sprache des gesamten Dokuments wird auf dem Element `<html>` selbst definiert. Die Angabe der Textsprache ermöglicht es unter anderem Bildschirmvorlesesoftware Wörter korrekt auszusprechen. Der Wert des Attributs muss entweder ein einzelnes Sprachkürzel oder leer sein. Ein leeres Attribut zeigt an, dass die Sprache unbekannt ist. [21]

Sind beide Angaben nicht vorhanden, kann als nächstbeste Annäherung Textanalysesoftware verwendet werden. Analysebibliotheken wie Apache Tika implementieren statistische Methoden zur Erkennung der Sprache natürlichsprachlicher Texte [33]. Die oben als Beispiel gewählte Webseite mit einem Englisch-Sprachkurs für Deutschsprachige würde mit dieser Methode möglicherweise als englischsprachig erkannt. Es kann jedoch angenommen werden, dass Textsprache einer Webseite und Sprache des Zielpublikums in vielen Fällen übereinstimmen, sodass diese Methode als Rückfalllösung geeignet erscheint.

Algorithmus 3 verwendet diese Elemente zur Entscheidung der Deutschsprachigkeit einer Webseite. Das Verfahren schlägt fehl, wenn die HTTP-Kopfzeile `Content-Language` nicht gesetzt ist und das HTML-Dokument fehlt.

Algorithmus 3 Sprachklassifikation

Eingabe: HTTP-Kopfzeilen `headers`
Eingabe: HTML-Dokument `document`
Ausgabe: **true**, wenn deutschsprachig. **false**, sonst.
  if `headers` enthält Feld `Content-Language` **then**
    return **true**, wenn `Content-Language` deutsches Kürzel enthält. **false**, sonst.
  else if `document` hat nichtleeres Attribut `lang` auf `<html>`-Element **then**
    return **true**, wenn `lang` deutsches Kürzel enthält. **false**, sonst.
  else
    return **true**, wenn Text laut Analysesoftware deutsch. **false**, sonst.
  end if

Die Qualität der Ergebnisse soll an späterer Stelle stichprobenartig evaluiert werden.

3.2.3 Kategorisierung von Webseiten

Um untersuchen zu können, ob auf öffentlich-rechtlichen Seiten weniger User-Tracking eingesetzt wird als auf Seiten privater Medien, sollen die Medienseiten in der Stichprobe identifiziert werden. Dazu müssen entsprechende Informationen über die Webseiten beschafft werden. Da eine automatisierte Kategorisierung auf Grundlage der erhobenen Daten nicht trivial und eine manuelle Klassifizierung einer großen Anzahl von Webseiten nicht realistisch ist, wird eine externe Informationsquelle benötigt.

In dieser Arbeit wird der Dienst *SimilarSites* zur Kategorisierung verwendet [30]. Da die Nutzung der API kostenpflichtig ist, werden die Kategorie-Informationen aus dem HTML-Dokument extrahiert. Ist dem Dienst die Webseite nicht bekannt, kann sie nicht kategorisiert werden.

```
$ curl -I http://www.tageschau.de/
HTTP/1.1 301 Moved Permanently
Server: Apache/2.4.7 (Ubuntu) PHP/5.5.9-1ubuntu4.26
Content-Type: text/html; charset=iso-8859-1
Date: Sat, 16 Feb 2019 15:21:09 GMT
Location: https://www.tagesschau.de/
Transfer-Encoding: chunked
Connection: Keep-Alive

$ curl -I https://www.tagesschau.de/
HTTP/1.1 200 OK
Server: Apache/2.4.7 (Ubuntu)
Content-Type: text/html; charset=utf-8
Date: Sat, 16 Feb 2019 15:21:12 GMT
Connection: keep-alive
```

Abbildung 3.3: Besucher von `tageschau.de` werden auf `tagesschau.de` umgeleitet

3.2.4 Erkennung von Third-Partys

Bevor eine aufgerufene Webseite dargestellt wird, kann der Nutzer mehrfach umgeleitet werden. Ein Beispiel findet sich in Abbildung 3.3. Ruft der Nutzer versehentlich `tageschau.de` statt `tagesschau.de` auf, so wird er umgeleitet.

Die Möglichkeit von Umleitungen soll auch bei der Definition, welche Domains als First-Party zu betrachten sind, berücksichtigt werden. In diesem Beispiel bilden `tageschau.de` und `tagesschau.de` die Menge der First-Partys des Seitenbesuchs.

Um zu überprüfen, ob zwei Domains in Bezug aufeinander als First-Partys zu erachten sind, genügt es nicht, lediglich die Übereinstimmung von Top-Level-Domain und Second-Level-Domain zu überprüfen. Auf diese Weise würden alle mit `.co.uk` endenden Domains als First-Partys zueinander betrachtet. Stattdessen gelten zwei Domains als First-Partys zueinander, wenn öffentliches Suffix und die nächsttiefere Ebene übereinstimmen. Zur Erkennung öffentlicher Suffixe wird die Public Suffix List der Mozilla Foundation verwendet [25].

Zwei Domains gelten genau dann als Third-Partys zu einander, wenn sie keine First-Partys sind. Dementsprechend gilt hier eine angefragte URL als Third-Party in Bezug zur besuch-

ten Seite, wenn sie Third-Party in Bezug auf alle Domains der Menge der First-Partys des Seitenbesuchs ist.

3.2.5 Zuordnung von Domains zu Unternehmen

Um untersuchen zu können, welche Tracking-Unternehmen die weiteste Verbreitung haben, müssen Domains ihren Unternehmen zugeordnet werden. Dazu wird die Liste des Unternehmens Disconnect verwendet [16]. Die Liste ordnet über 2 300 Domains über 1 000 verschiedenen Unternehmen zu.

3.2.6 Speicherung der Sekundärdaten

Die auf Grundlage der Rohdaten gewonnenen Sekundärdaten sollen für die Analyse geeignet gespeichert werden. Dazu wird eine SQLite-Datenbank verwendet. Abbildung 3.4 stellt das verwendete Datenbankschema dar.

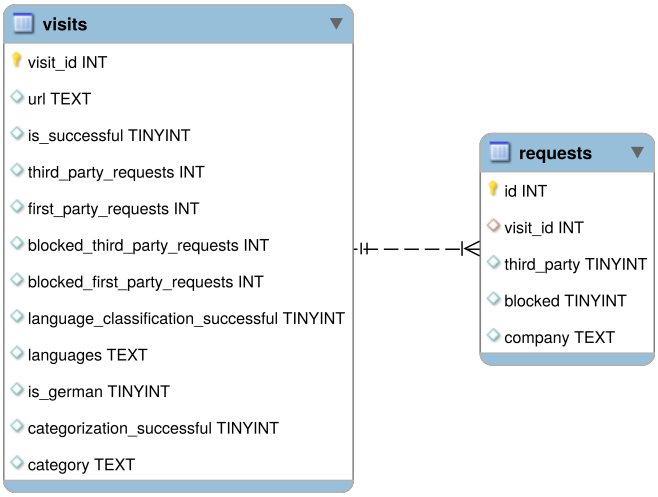

Abbildung 3.4: Datenbankschema zur Speicherung der Sekundärdaten

Ein Eintrag in der Tabelle `requests` bezieht sich auf die Anfrage mit der gleichen `id` aus der OpenWPM-Tabelle `http_requests` und enthält folgende Informationen:

- `third_party` gibt an, ob die angefragte URL eine Third-Party (siehe Abschnitt 3.2.4) in Bezug auf die besuchte Webseite ist.

- `blocked` gibt an, ob die angefragte URL als Tracker (siehe Abschnitt 3.2.1) klassifiziert wurde.
- `company` gibt an, zu welchem Unternehmen (siehe Abschnitt 3.2.5) die angefragte Domain gehört, sofern bekannt. Sonst ist das Attribut NULL.

Ein Eintrag in der Tabelle `visits` bezieht sich auf den Seitenbesuch mit der gleichen `visit_id` aus der OpenWPM-Tabelle `site_visits` und enthält folgende Informationen:

- Ein Besuch wird als erfolgreich (`is_successful`) klassifiziert, wenn es mindestens eine Antwort mit Statuscode 200 gab.
- Die Attribute `[blocked_](first|third)_party_requests` enthalten die entsprechenden Anzahlen an Anfragen und ergeben sich aus der Tabelle `requests`.
- `language_classification_successful` gibt an, ob die Sprache der Webseite nach dem Verfahren in Abschnitt 3.2.2 klassifiziert werden konnte. War die Klassifikation erfolgreich, dann enthält `languages` die Sprachkürzel und `is_german` gibt an, ob eine der Sprachen deutsch ist. Sonst sind die Attribute NULL.
- `categorization_successful` gibt an, ob die Webseite nach dem Verfahren in Abschnitt 3.2.3 kategorisiert werden konnte. War die Kategorisierung erfolgreich, dann enthält `category` die Kategorie. Sonst ist das Attribut NULL.

3.3 Aufbereitung und Analyse

3.3.1 Evaluation binärer Klassifikatoren

Um die Güte des Verfahrens zur Bestimmung der Deutschsprachigkeit zu bewerten, soll stichprobenartig die Korrektheit des Klassifikationsergebnis kontrolliert werden. Dabei gibt es vier mögliche Ausgänge, welche in Abbildung 3.5 dargestellt sind. [26]

Auf Grundlage der vorgefundenen Häufigkeiten können Kenngrößen zur Bewertung des Klassifikators berechnet werden. Folgende Kenngrößen sind üblich: [26]

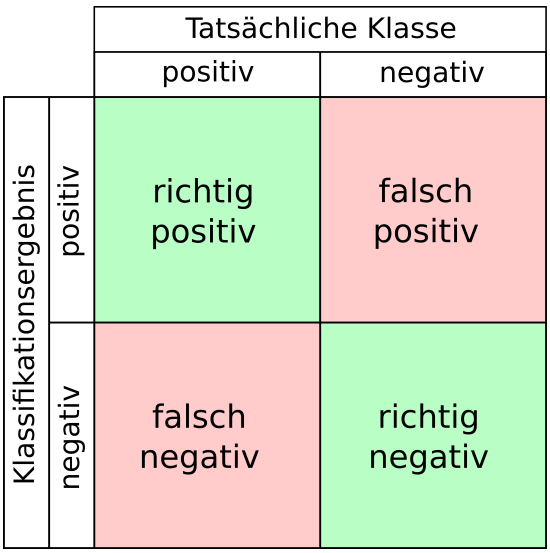

Abbildung 3.5: Konfusionsmatrix zur Evaluation binärer Klassifikatoren (nach [26])

$$recall = \frac{TP}{TP + FN}$$

Die Richtig-positiv-Rate, oder *recall*, gibt den Anteil aller positiven Objekte ($TP + FN$) an, die auch als positiv klassifiziert (TP) wurden.

$$specificity = \frac{TN}{TN + FP}$$

Die Richtig-negativ-Rate, oder *specificity*, gibt den Anteil aller negativen Objekte ($TN + FP$) an, die auch als negativ klassifiziert (TN) wurden.

$$precision = \frac{TP}{TP + FP}$$

Precision gibt den Anteil aller als positiv klassifizierten Objekte ($TP + FP$) an, die tatsächlich positiv (TP) sind.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy gibt den Anteil aller Objekte ($TP + TN + FP + FN$) an, die korrekt klassifiziert ($TP + TN$) wurden.

3.3.2 Statistische Methoden

Zur Aufbereitung und Analyse werden statistische Methoden verwendet. Bei der Beschreibung der Stichproben kommen bekannte Kenngrößen wie das arithmetische Mittel, die Standardabweichung und Quantile zum Einsatz. Veranschaulicht werden die Daten mit Histogrammen und empirischen Verteilungsfunktionen. Auf eine Definition dieser wird an dieser Stelle verzichtet und im Zweifelsfall auf [20] verwiesen.

Stattdessen soll hier in die verwendeten Methoden der schließenden Statistik eingeführt werden. Diese werden in dieser Untersuchung bei dem Vergleich des User-Tracking auf deutschsprachigen und nicht-deutschsprachigen sowie beim Vergleich des User-Tracking auf öffentlich-rechtlichen und privaten Medienseiten eingesetzt.

Formen von Inferenz

Methoden der schließenden Statistik werden verwendet, um Schlüsse aus Stichprobendaten zu ziehen. Es können zwei verschiedene Formen der Inferenz, die kausale Inferenz und die Inferenz auf Grundgesamtheiten, unterschieden werden, deren Zulässigkeit vom Studiendesign abhängen. [27]

Kausale Inferenz bedeutet, dass eine Beziehung zwischen Ursache und Wirkung hergestellt werden kann. Voraussetzung dafür ist ein randomisiertes Experiment. In einem randomisierten Experiment werden Beobachtungseinheiten per Zufall verschiedenen Behandlungen zugewiesen, deren Effektivität beurteilt werden soll. Durch die zufällige Aufteilung auf Behandlungsgruppen wird statistisch der Einfluss des Unterschieds zwischen den Beobachtungseinheiten selbst auf die Ergebnisse des Experiments minimiert. [27]

Dieses experimentelle Design ist insbesondere für medizinische Forschung interessant. Genesen beispielsweise Patienten unter Verabreichung eines neuen Medikamentes im Durchschnitt deutlich schneller als unter Verabreichung eines bisherigen, kann, unter Berücksichtigung einer Irrtumswahrscheinlichkeit, darauf geschlossen werden, dass dies durch das neue Medikament bedingt ist, und nicht in den Patienten selbst begründet liegt.

Für die Inferenz auf Grundgesamtheiten ist die Art der Auswahl der Beobachtungseinheiten maßgeblich. Die Stichprobe der Beobachtungseinheiten muss eine Zufallsstichprobe sein. Eine einfache Zufallsstichprobe der Größe n einer Grundgesamtheit ist eine n-elementige Teilmenge der Grundgesamtheit, die so gezogen wird, dass jede n-elementige Teilmenge die gleiche Wahrscheinlichkeit hat gezogen zu werden. Ist dies gegeben, so sind Schlüsse auf die Grundgesamtheiten selbst möglich. [27]

Ein Beispiel wäre die Untersuchung des Unterschieds der Körpergröße zwischen männlichen und weiblichen Individuen einer Tierpopulation. Weist eine der beiden Zufallsstichproben durchschnittlich eine deutlich höhere Körpergröße auf, so kann der Unterschied auf die Gesamtpopulation verallgemeinert werden. In der Praxis dürfte es jedoch schwierig sein, echte Zufallsstichproben zu erheben.

In dieser Arbeit soll schließende Statistik verwendet werden, um quantitative Unterschiede im User-Tracking zu beurteilen. Die Art der Datenerhebung genügt dabei weder den Kriterien eines randomisierten Experiments, noch handelt es sich um eine Zufallsstichprobe. Es ist hier daher nicht möglich, festgestellte Unterschiede auf einen weiteren Kontext zu übertragen.

Nichtsdestotrotz können Methoden der schließenden Statistik, insbesondere das Testen von Hypothesen mit Permutationstests, geeignet sein, um an dieser Stelle interessante Muster in den Daten aufzuzeigen. Auch wenn nicht auf Verursachungszusammenhänge geschlossen werden kann oder die Ergebnisse verallgemeinert werden können, können Testergebnisse Hinweise geben und zukünftige Untersuchungen anregen.

Permutationstests

Um quantitative Unterschiede im User-Tracking zwischen zwei Gruppen von Webseiten zu untersuchen, können Hypothesentests verwendet werden. Hypothesentests überprüfen zuvor formulierte Annahmen auf Basis erhobener Daten. Die zu überprüfende Hypothese wird als Arbeitshypothese oder Alternativhypothese bezeichnet und der Nullhypothese, von welcher vorläufig ausgegangen wird und welche aussagt, dass der vermutete Effekt nicht existiert, gegenübergestellt. [27]

In dieser Arbeit werden Hypothesen mit Permutationstests untersucht. Permutationstests setzen im Unterschied zu anderen üblichen Testverfahren wie dem t-Test keine Annahmen

über die Verteilung des untersuchten Merkmals in der Grundgesamtheit voraus. Das Vorgehen ist folgendermaßen: [27]

1. Es wird eine zur Untersuchung der Hypothese geeignete Teststatistik, wie beispielsweise der Unterschied der Gruppenmittelwerte, ausgewählt und berechnet.
2. Alle Neugruppierungen der $n_1 + n_2$ Stichprobenwerte in Gruppen der Größen n_1 und n_2 werden generiert und jeweils der Wert der Teststatistik berechnet.
3. Die Anzahl der Werte aus Schritt 2, welche einen mindestens so extremen Wert haben wie der in Schritt 1 berechnete, wird bestimmt.
4. Die in Schritt 3 bestimmte Anzahl wird durch die Anzahl der aller Neugruppierungen geteilt. Dieser Wert wird als p-Wert bezeichnet.

Der p-Wert gibt den Anteil aller Neugruppierungen an, welche eine mindestens so extreme Teststatistik aufweisen. Je kleiner der p-Wert, desto unwahrscheinlicher ist die Gültigkeit der Nullhypothese. Liegt der p-Wert unter einem zuvor festgelegten Signifikanzniveau α, dann wird die Nullhypothese verworfen und die Arbeitshypothese angenommen. Übliche Werte für α sind $0{,}05$, $0{,}01$ und $0{,}001$. Liegt der p-Wert über α, dann wird die Nullhypothese beibehalten. [27]

Da $\binom{n_1+n_2}{n_1}$ Neugruppierungen existieren ist es häufig nicht praktikabel für jede die Teststatistik zu berechnen. Es wird sich daher in der Regel auf eine große Zufallsstichprobe aus der Menge aller Neugruppierungen beschränkt, auf Basis welcher ein näherungsweiser p-Wert berechnet wird. [27]

Effektmaß *Cliff's delta*

Auch sehr kleine Unterschiede in Verteilungen können statistisch signifikant werden, wenn die Stichproben groß genug sind. Die statistische Signifikanz eines Effekts zeigt also nicht notwendigerweise an, dass er auch praktisch relevant ist. [22]

Eine Möglichkeit die praktische Relevanz zu beurteilen liegt in der Verwendung von Effektmaßen zur Beurteilung der Effektstärke [22]. Ein Effektmaß, das keine Annahmen über die Verteilungform der Stichproben trifft, ist *Cliff's delta*: [4]

$$d = \frac{\#(x_i > x_j) - \#(x_i < x_j)}{mn}$$

Es wird also jedes der n Elemente der ersten Gruppe mit jedem der m Elemente der zweiten Gruppe verglichen, und gezählt wie oft ein Element der ersten Gruppe größer und wie oft es kleiner ist. Anschließend wird die Differenz gebildet und durch die Anzahl der Paarungen geteilt.

Das Effektmaß gibt an, wie stark sich zwei Stichproben überlappen. Dabei gibt der Wert 0 totale Überlappung an, während -1 und 1 keinerlei Überlappung bedeuten. Zur Beurteilung der Effektstärke geben Romano u. a. [28] folgende Schwellenwerte an[1]:

$\lvert d\rvert < 0{,}147$	zu vernachlässigen
$\lvert d\rvert < 0{,}33$	klein
$\lvert d\rvert < 0{,}474$	mittel
sonst	groß

[1] Zitiert nach [34]

4 Untersuchungsergebnisse

4.1 Übersicht über die erhobenen Daten

Zeitraum der Besuche

10.07.2018 bis 11.07.2018

Anzahl der besuchten Webseiten

10 000

Anzahl der erfolgreichen Besuche

9 655 (96,6 %)

Anzahl der Besuche mit erfolgreicher Sprachklassifikation

7 930 (82,1 % der erfolgreichen Besuche)

Anzahl der als deutschsprachig klassifizierten Webseiten

1 089 (13,7 % der Besuche mit erfolgreicher Sprachklassifikation)

Anzahl der Besuche mit erfolgreicher Kategorisierung

8 437 (87,4 % der erfolgreichen Besuche)

Abbildung 4.1 stellt die Häufigkeiten der Kategorien dar.

Verteilung des Merkmals *Anzahl zu blockierender Anfragen*

Tabelle 4.1 stellt einige Kenngrößen der Merkmalsverteilung dar. Es werden durchschnittlich etwa 25 Anfragen als zu blockierende Tracker klassifiziert. Gleichzeitig liegt der Median nur bei 11, das Maximum jedoch bei 307. Die Standardabweichung liegt mit etwa 34,8 höher als der Mittelwert, die Werte streuen also sehr weit um den Mittelwert herum. Der Mittelwert charakterisiert die Verteilung nur sehr schlecht. Es handelt sich um eine sehr schiefe Verteilung mit extremen Ausreißern nach oben.

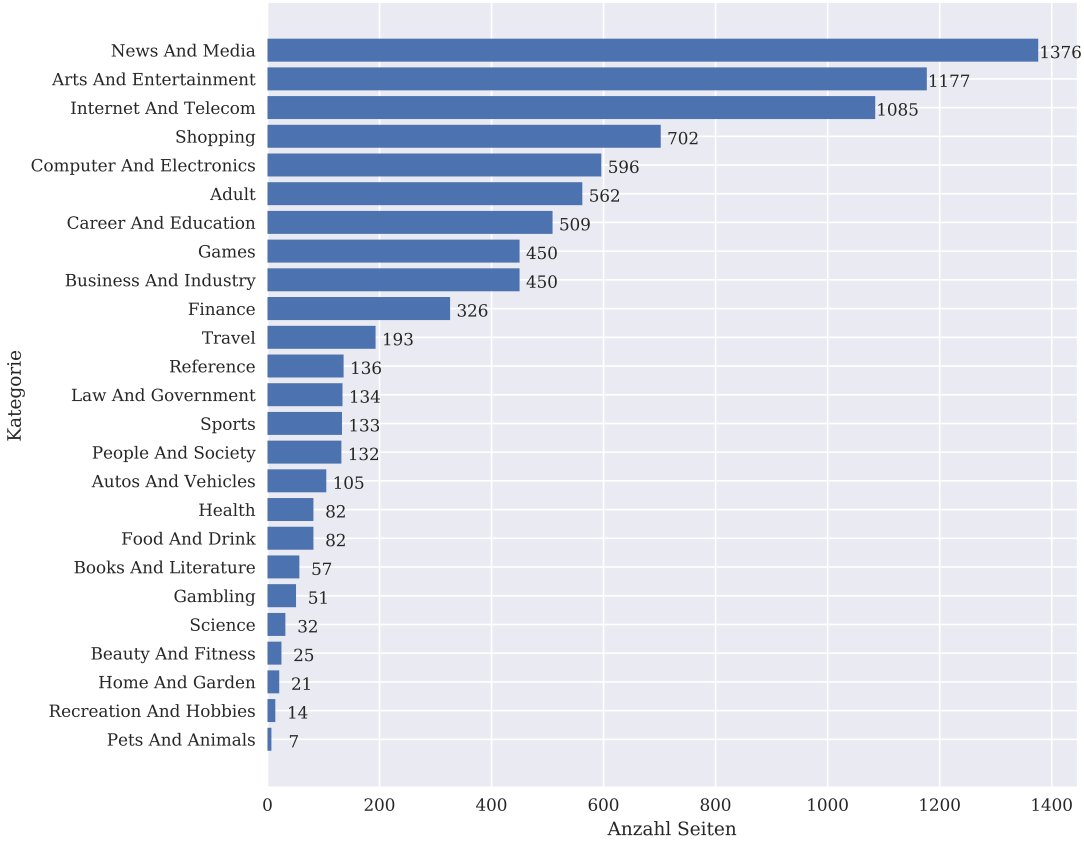

Abbildung 4.1: Häufigkeiten der Kategorien

Tabelle 4.1: Kenngrößen zur Verteilung des Merkmals *Anzahl zu blockierender Anfragen*, untergliedert nach Anfragen zum Erstanbieter (*first-party*), zu Dritten (*third-party*) und insgesamt

	first-party	*third-party*	total
Mittelwert	0,718	24,408	25,126
Standardabweichung	4,072	34,239	34,797
Minimum	0	0	0
25 %-Perzentil	0	3	3
Median	0	10	11
75 %-Perzentil	0	32	33
Maximum	118	307	307

Die Histogramme in Abbildung 4.2 veranschaulichen die ungleichmäßige Häufigkeitsverteilung. Die Form der Verteilung entspricht etwa der einer geometrischen Verteilung.

Zu blockierende Anfragen sind fast immer Anfragen zu Third-Partys. Die durchschnittliche Zahl zu blockierender Anfragen an First-Partys liegt bei etwa $0{,}7$. Auf etwa $81{,}3\,\%$ der Webseiten müssen überhaupt keine Anfragen zu eigenen Ressourcen blockiert werden.

4.2 Evaluation des Verfahrens zur Sprachklassifikation

Zur Evaluation wurden 30 als deutschsprachig und 30 als nicht-deutschsprachig eingestufte Seiten händisch besucht und die Korrektheit der Einstufung kontrolliert. Eine Webseite wurde dabei als deutschsprachig eingestuft, wenn der überwiegende Text deutsch war oder Deutsch als Sprache über ein entsprechendes Symbol oder Menü auswählbar war. Die Auswahl der Webseiten erfolgte zufällig ohne Zurücklegen. Die Seiten wurden innerhalb einer virtuellen Maschine (Linux Kubuntu 18.04, Firefox 61) besucht. Nach jedem erfolgten Besuch wurden der Browser auf den Auslieferungszustand zurückgesetzt, was insbesondere das Löschen aller Cookies und zwischengespeicherter Webseitendaten beinhaltet. So wie bei der Datenerhebung mit OpenWPM zuvor wurde die Anfragekopfzeile `Accept-Languages` auf `de-DE,de` gesetzt.

Tabelle 4.2: Konfusionsmatrix zur Evaluation des Verfahrens zur Sprachklassifikation

		Tatsächliche Sprache		
		deutsch	nicht-deutsch	total
Klassifikationsergebnis	deutsch	28 (TP)	2 (FP)	30
	nicht-deutsch	1 (FN)	27 (TN)	28
	total	29	29	58

Tabelle 4.2 stellt das Evaluationsergebnis in einer Konfusionsmatrix dar. Zwei Webseiten wurden fälschlicherweise als deutsch klassifiziert, eine Webseite fälschlicherweise als nicht-deutsch. In den übrigen Fällen war die Klassifikation korrekt. Die Gesamtzahl der Webseiten addiert sich zu 58 statt 60, da zum Zeitpunkt der Evaluation zwei Webseiten nicht aufrufbar waren.

Es ergeben sich folgende Kenngrößen:

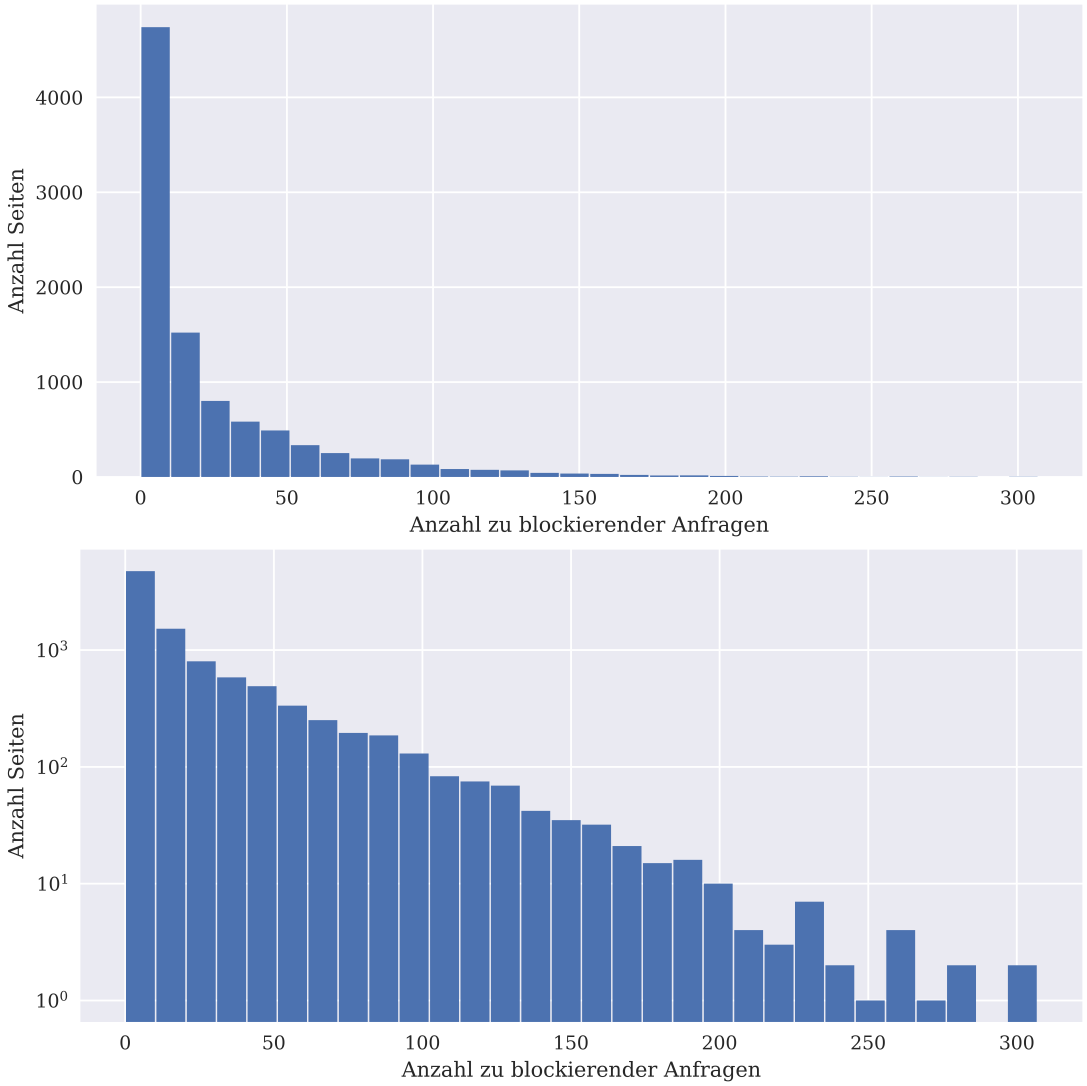

Abbildung 4.2: Häufigkeit der Anzahl zu blockierender Anfragen (Darstellung mit linearer und logarithmischer y-Achsenskalierung, verwendet wurden jeweils 30 Klassen gleicher Breite)

$$
\begin{aligned}
recall &= \frac{TP}{TP+FN} &&= \frac{28}{29} \approx 0{,}966 \\
specificity &= \frac{TN}{TN+FP} &&= \frac{27}{29} \approx 0{,}931 \\
precision &= \frac{TP}{TP+FP} &&= \frac{28}{30} \approx 0{,}933 \\
accuracy &= \frac{TP+TN}{TP+TN+FP+FN} &&= \frac{55}{58} \approx 0{,}948
\end{aligned}
$$

4.3 Tracking in Abhängigkeit von Deutschsprachigkeit

Tabelle 4.3: Kenngrößen zur Verteilung des Merkmals *Anzahl zu blockierender Anfragen* in Abhängigkeit von Deutschsprachigkeit

	deutschsprachig	*nicht-deutschsprachig*
Mittelwert	19,317	19,802
Standardabweichung	27,934	28,309
Minimum	0	0
25 %-Perzentil	1	3
Median	7	9
75 %-Perzentil	26	25
Maximum	202	258

Es soll die Hypothese überprüft werden, die Anzahl der Tracker auf deutschsprachigen Webseiten sei niedriger als auf nicht-deutschsprachigen Webseiten. Es wird zunächst die Verteilung des Merkmals *Anzahl zu blockierender Anfragen* in Abhängigkeit von Deutschsprachigkeit betrachtet (Tabelle 4.3). Standardabweichung und Mittelwert beider Gruppen sind sehr ähnlich, kleinere Unterschiede gibt es bei den Quantilen.

Abbildung 4.3 veranschaulicht die Ähnlichkeit beider Verteilungen. Die empirischen Verteilungsfunktionen liegen sehr nahe bei einander und schneiden einander mehrfach. Zunächst scheinen deutschsprachige Webseiten häufiger kein Tracking einzusetzen als nicht-deutschsprachige Webseiten. So müssen auf etwa 22,8 % der deutschsprachigen Webseiten keinerlei Tracker blockiert werden, während dies nur bei 12,4 % der nicht-deutschsprachigen Seiten der Fall ist. Bei genauerer Betrachtung relativiert sich dieses Ergebnis jedoch schnell,

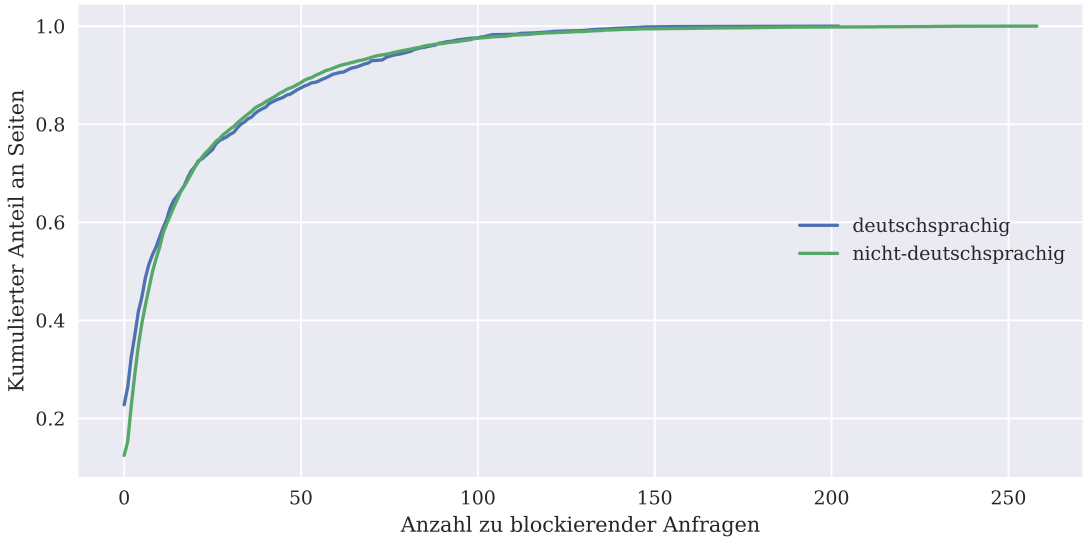

Abbildung 4.3: Empirische Verteilungsfunktion des Merkmals *Anzahl zu blockierender Anfragen* in Abhängigkeit von Deutschsprachigkeit

da die Menge der deutschsprachigen Seiten, die kein Tracking einsetzen, auch 114 länderspezifische Google-Domains enthält. Google scheint den Nutzerstandort für die Auswahl der anzuzeigenden Sprache zu nutzen, sodass alle Domains zur gleichen Seite führen. Berücksichtigt man bei der Bestimmung des Verhältnisses Google nur einmal, so reduziert sich der Anteil deutschsprachiger Seiten, die kein Tracking einsetzen, auf etwa $13{,}7\,\%$.

Das Effektmaß *Cliff's delta* liegt bei $|d| = 0{,}072$, was eine starke Überlappung anzeigt. Es wäre daher von einem zu vernachlässigendem Effekt auszugehen. Auf einen Hypothesentest kann daher an dieser Stelle verzichtet werden. Die Hypothese, die Anzahl der Tracker sei auf deutschsprachigen Seiten niedriger als auf nicht-deutschsprachigen Seiten, kann auf Grundlage der erhobenen Daten nicht bestätigt werden.

4.4 Tracking auf deutschsprachigen Seiten: öffentlich-rechtliche und private Medien

Es soll die Hypothese überprüft werden, die Anzahl der Tracker auf Webseiten öffentlich-rechtlicher Medien sei geringer als auf Webseiten privater Medien.

Tabelle 4.4: Kenngrößen zur Verteilung des Merkmals *Anzahl zu blockierender Anfragen* bei öffentlich-rechtlichen und privaten Medienseiten

	öffentlich-rechtlich	*privat*
Mittelwert	8,667	33,341
Standardabweichung	7,738	37,778
Minimum	3	0
25 %-Perzentil	5,75	4
Median	6	14,5
75 %-Perzentil	7,25	58
Maximum	31	147

Nur eine Seite in der Kategorie *News And Media* konnte als öffentlich-rechtlich identifiziert werden. Weitere elf wurden in der Kategorie *Arts And Entertainment* ausfindig gemacht. Die Trackerzahlen der 12 öffentlich-rechtlichen Seiten werden mit den übrigen 131 deutschsprachigen Webseiten in den beiden genannten Kategorien verglichen.

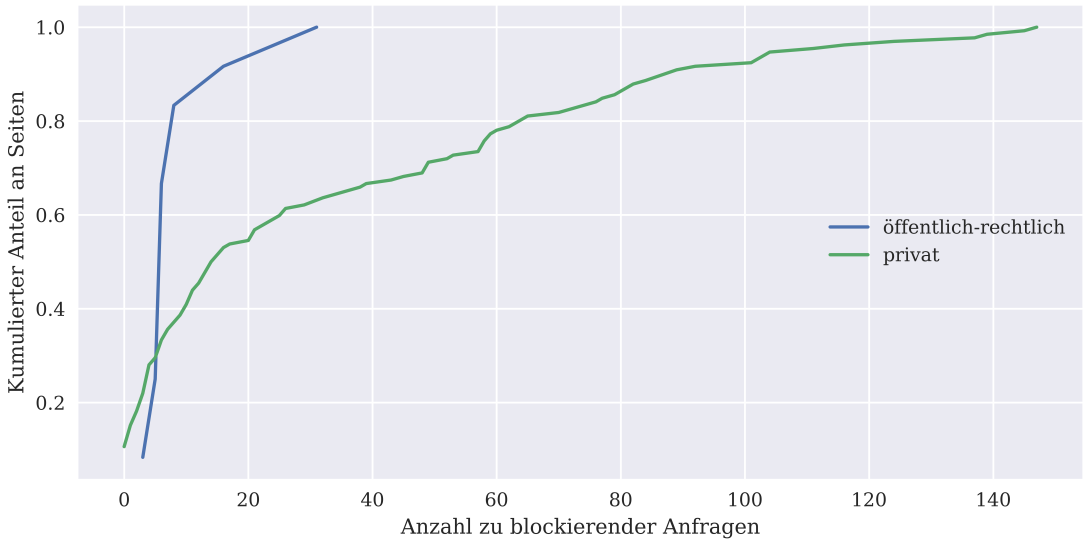

Abbildung 4.4: Empirische Verteilungsfunktion des Merkmals *Anzahl zu blockierender Anfragen* auf Medienseiten

Tabelle 4.4 zeigt, dass die durchschnittliche Trackerzahl bei privaten Medienseiten mit etwa 33,3 deutlich höher ist als bei öffentlich-rechtlichen. Gleichzeitig ist bei privaten Medienseiten jedoch auch die Standardabweichung größer als der Mittelwert, Minimalwert und

25 %-Perzentil kleiner als bei öffentlich-rechtlichen Webseiten. Auch hier charakterisiert der Mittelwert die Verteilung nur schlecht. Abbildung 4.4 veranschaulicht die Verteilung.

Es wird auch hier mit dem Effektmaß *Cliff's delta* die Überlappung beider Gruppen hinsichtlich des Merkmals geprüft. Das Ergebnis von $d \approx -0{,}306$ zeigt eine noch relativ starke Überlappung an, was bei Vorhandensein statistischer Signifikanz für einen noch schwachen Effekt spräche. Mit einem linksseitigen Permutationstest mit 1 000 000 Permutationen wird geprüft, ob der Unterschied der Verteilungen signifikant ist. Da der Mittelwert an dieser Stelle aufgrund der schiefen Verteilungsform nur schlecht beschreibt, welche Anzahl an Trackern in den Gruppen jeweils zu erwarten ist, wird auf den robusteren Median zurückgegriffen. Als Teststatistik wird daher die Medianabweichung zwischen beiden Gruppen gewählt.

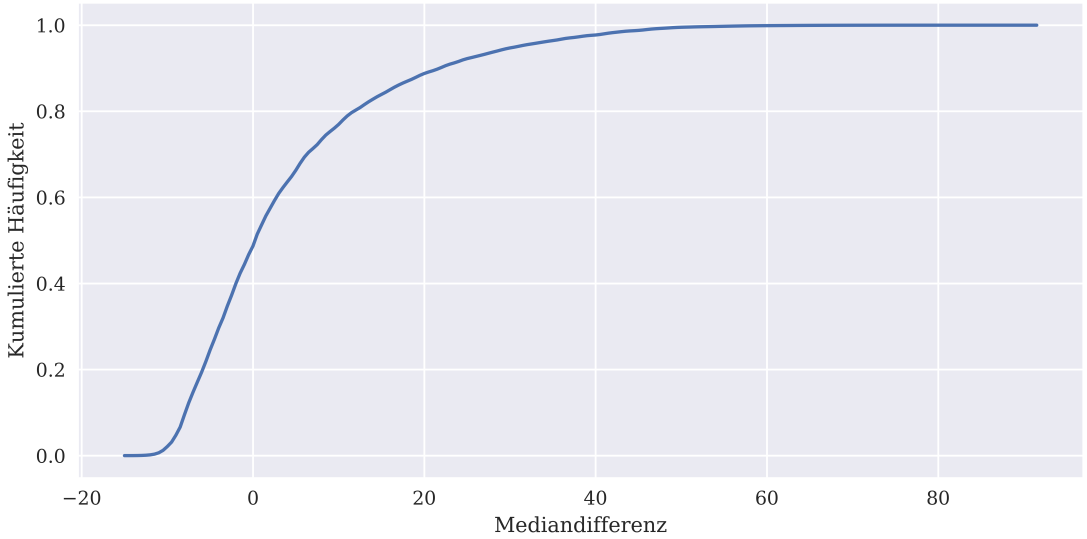

Abbildung 4.5: Empirische Verteilungsfunktion der Permutationsverteilung

Abbildung 4.5 zeigt die Verteilungsfunktion der Permutationsverteilung. Die Mediandifferenz zwischen der öffentlich-rechtlichen Gruppe und der privaten Gruppe liegt bei $-8{,}5$. Der p-Wert liegt bei etwa $0{,}065$, d.h. etwa $6{,}5\,\%$ der Permutationen weisen einen ebenso kleinen oder kleineren Wert auf. Das Ergebnis ist bei einem Signifikanzniveau von $\alpha = 0{,}05$ nicht signifikant, die Nullhypothese wird beibehalten. Die Hypothese, auf Webseiten öffentlich-rechtlicher Medien finde weniger Tracking statt als auf Webseiten privater Medien, kann an dieser Stelle daher nicht bestätigt werden.

4.5 Information über den Einsatz von Tracking-Cookies

Tabelle 4.5: Information über das Setzen von Tracking-Cookies und tatsächliches Setzen bei Besuch der Webseite

		Tracking-Cookies		
		ja	nein	total
Cookie-Information	keine	13 (28,9 %)	0 (0 %)	13 (28,9 %)
	Info	26 (57,8 %)	2 (4,4 %)	28 (62,2 %)
	Opt-out	1 (2,2 %)	0 (0 %)	1 (2,2 %)
	Opt-in	1 (2,2 %)	2 (4,4 %)	3 (6,7 %)
	total	41 (91,1 %)	4 (8,9 %)	45 (100 %)

Zur Untersuchung der Tracking-Informationen und des tatsächlich stattfindenden Setzen von Cookies wurden 50 als deutschsprachig klassifizierte Webseiten besucht und entsprechend der Methodik in Abschnitt 3.1.3 bewertet. Die Besuche erfolgten am 15. August 2018. Die Ergebnisse sind in Tabelle 4.5 dargestellt.

Von den 50 besuchten Webseiten waren vier nicht deutschsprachig (fehlklassifiziert) und eine war nicht aufrufbar, sodass sich die Gesamtzahl auf 45 reduziert. Mit weitem Abstand wurden auf den besuchten Webseiten Cookie-Banner eingesetzt, gefolgt von keiner Information zum Tracking. Trotz des Hinweises auf Tracking setzten zwei Webseiten mit Cookie-Bannern keine Tracking-Cookies. Lediglich drei Webseiten setzten auf die Opt-in-Lösung, nur eine auf die Opt-out-Variante. Trotz des Verwendens der Opt-in-Lösung setzte eine Webseite Tracking-Cookies, bevor diesem zugestimmt wurde.

4.6 Tracking-Firmen auf deutschsprachigen und nicht-deutschsprachigen Webseiten

Abbildung 2.1 stellt die Verbreitung von Tracking-Unternehmen auf deutschsprachigen und nicht-deutschsprachigen Webseiten dar. Abgebildet sind alle Unternehmen, die auf mindestens 5 % aller deutschsprachigen oder auf mindestens 5 % aller nicht-deutschsprachigen Webseiten vertreten waren.

Grundsätzlich sind die Tracking-Unternehmen auf deutschsprachigen und nicht-deutschsprachigen Seiten ähnlich verteilt. Bei beiden Gruppen ist Google mit Abstand der häufigste Tracker,

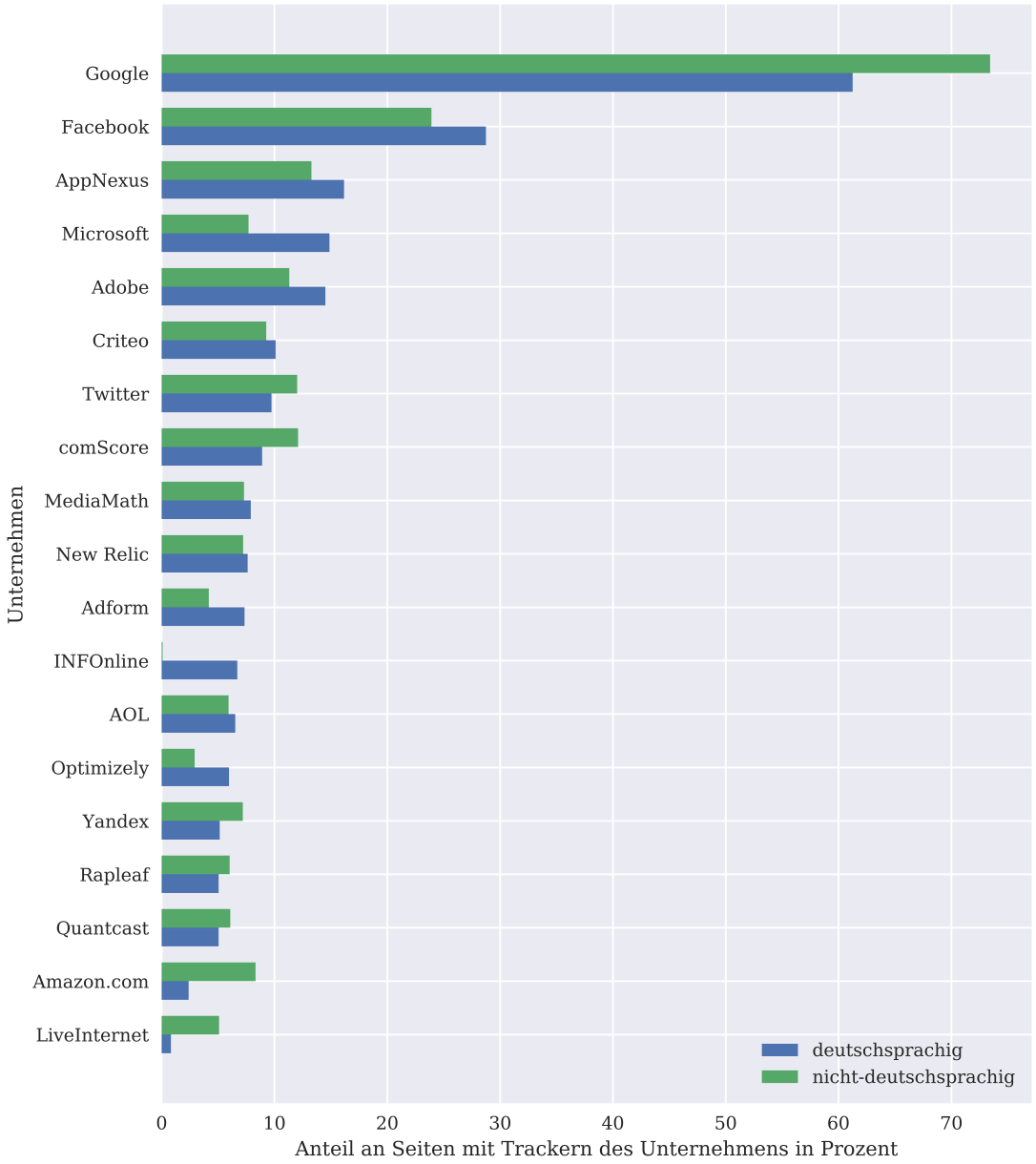

Abbildung 4.6: Tracking-Unternehmen auf deutschsprachigen und nicht-deutschsprachigen Webseiten

wenn auch die Häufigkeit auf deutschsprachigen Seiten merklich geringer ist. Hier gibt es eine Differenz von 12,2 %. Auch Amazon (Differenz 5,9 %) und LiveInternet (Differenz 4,3 %) sind auf deutschsprachigen Seiten im Vergleich weniger stark vertreten. Letztgenanntes Unternehmen scheint in Deutschland fast unbekannt (Anteil 0,8 %).

Auf deutschsprachigen Seiten stärker vertreten als auf nicht-deutschsprachigen Seiten sind Microsoft (Differenz 7,2 %), INFOnline (Differenz 6,6 %) und Facebook (Differenz 4,8 %). INFOnline scheint fast exklusiv im deutschsprachigen Raum aktiv zu sein, da es nur auf 0,1 % der nicht-deutschsprachigen Seiten vertreten ist.

5 Diskussion

In diesem Kapitel werden die Ergebnisse der Untersuchung diskutiert. Dabei wird in Abschnitt 5.1 zunächst darauf eingegangen, welche methodischen Unzulänglichkeiten festzustellen sind. Anschließend diskutiert Abschnitt 5.2 die Ergebnisse in Hinblick auf die zu Beginn der Arbeit formulierten Fragestellungen.

5.1 Methodische Einschränkungen

5.1.1 Probleme beim Crawling

Zum Besuch der 10 000 Webseiten wurde die Software OpenWPM verwendet. Im Einsatz erwies sich diese leider nicht als sonderlich robust. So stürzte das Programm häufig während des Crawlingvorgangs ab. Ein Fortsetzen an der Stelle des Abbruchs ist leider nicht möglich, sodass auch der gesamte Crawlingvorgang bis zum Punkt der Absturzes wiederholt werden muss. Dies ist angesichts des Zeitbedarfs von etwa 28 Stunden für einen vollständigen Crawl ärgerlich. Der Fehler ist seit Dezember 2017 bekannt, jedoch bisher nicht behoben [14].

Ein weiteres Problem stellte das unzuverlässige Speichern der Seitenquelltexte dar. Diese wurden bei etwa 18 % der erfolgreichen Besuche nicht gespeichert. Der Grund dafür ist unklar. Dies hatte Auswirkungen auf die Anzahl der Webseiten mit erfolgreicher Sprachklassifikation.

5.1.2 Verfahren der Sprachklassifikation

Zur Evaluation des Sprachklassifikationsverfahrens wurde stichpunktartig die Korrektheit der Klassifikation überprüft.

Von 29 tatsächlich deutschsprachigen Webseiten wurden 28 richtig klassifiziert ($recall \approx 0{,}966$). Bei der falsch klassifizierten Seite war deutsch als Sprache auswählbar, standardmäßig war der Text jedoch englischsprachig. Dieses Problem war zu befürchten, scheint angesichts des Verhältnisses von 28 zu 1 aber selten aufzutreten.

Von 29 tatsächlich nicht-deutschsprachigen Webseiten wurden 27 richtig klassifiziert ($specificity \approx 0{,}931$). Bei einer Seite war die Fehlklassifikation nicht nachvollziehbar, bei der anderen gab der Server in der Antwortkopfzeile `Content-Language de-DE` an, obwohl die Webseite weder deutschsprachig war, noch sich irgendwie erkennbar an ein spezifisch deutschsprachiges Publikum wendete. Dieses Zurückgeben der in der Anfragekopfzeile `Accept-Language` angegebenen bevorzugten Sprache konnte auch bei fehlklassifizierten Webseiten im Rahmen der Untersuchung der Nutzerinformation zum Tracking beobachtet werden. Dieses Verhalten überrascht zunächst, kann aber möglicherweise als Versuch der Optimierung der Suchmaschinenposition gewertet werden. In jedem Fall büßt damit `Content-Language` als Entscheidungskriterium an Zuverlässigkeit ein.

Insgesamt erscheint das Verfahren verbesserungswürdig. Nur etwa $93{,}3\,\%$ der als deutschsprachig klassifizierten Webseiten waren es tatsächlich (*precision*). Richtig klassifiziert wurden insgesamt etwa $94{,}8\,\%$ der Webseiten (*accuracy*). Die Möglichkeit von Klassifikationsfehlern muss daher bei der Bewertung der Ergebnisse berücksichtigt werden.

5.1.3 Kategorisierung von Webseiten

Zur Kategorisierung von Webseiten wurde der Dienst SimilarSites verwendet. Über ihn konnten $84{,}7\,\%$ der erfolgreich besuchten Webseiten kategorisiert werden. Allerdings wurden nur eine Webseite öffentlich-rechtlicher Medien in der Kategorie *News And Media* ausfindig gemacht, während weitere in der Kategorie *Arts And Entertainment* gefunden wurden. Gleichzeitig enthält *Arts And Entertainment* auch den Streamingdienst Spotify, während der Streamingdienst Netflix in der Kategorie *Internet And Telecom* gelistet ist. Dies lässt an der Kohärenz der Kategorisierung zweifeln, oder zeigt zumindest auf, dass den Webseiten nicht immer diejenige Kategorie zugeordnet ist, die man erwarten würde.

Da die Nutzung der API kostenpflichtig ist, wurden die Informationen direkt aus dem jeweiligen HTML-Dokument extrahiert. Dies ist inzwischen nicht mehr praktikabel, da der Dienst nun auf clientseitiges JavaScript zum Seitenaufbau setzt, sodass dieses erst wie in einem

Browser interpretiert werden müsste. Zusätzlich kommen Captchas zum Einsatz, um gerade das automatisierte Auslesen zu verhindern. Die Wiederverwendung der Implementation zur Kategorisierung scheint daher ausgeschlossen.

5.2 Bewertung der Ergebnisse

5.2.1 Tracking in Abhängigkeit von Deutschsprachigkeit

Es sollte untersucht werden, ob das Tracking auf deutschsprachigen Seiten geringer ausfällt als auf nicht-deutschsprachigen Seiten. Dabei wurde festgestellt, dass die Anzahl der Tracker in beiden Gruppen sehr ähnlich verteilt ist. Die festgestellten Mittelwerte liegen nah bei dem, der von Englehardt und Narayanan [8] festgestellt worden war. Tatsächlich überlappten sich beide Verteilungen so stark ($|d| = 0{,}072$), dass auf einen Hypothesentest verzichtet werden konnte. Beide Gruppen sind von einer sehr hohen Variabilität innerhalb der Gruppen gekennzeichnet, eine Variabilität zwischen den Gruppen ist dagegen praktisch nicht erkennbar. Für den Nutzer bedeutet dies, dass er nicht erwarten kann, auf deutschsprachigen Webseiten grundsätzlich weniger Tracking ausgesetzt zu sein.

Die Erhebung der Daten erfolgte im Juli 2018. Unklar ist, wie stark sich zu dem Zeitpunkt, etwa sechs Wochen nach dem Anwendbarkeitsdatum der DSGVO, diese schon auf das Tracking auf Webseiten ausgewirkt hatte. Es ist davon auszugehen, dass die DSGVO auf praktisch alle deutschsprachigen Webseiten anwendbar ist, während die Gruppe der nicht-deutschsprachigen Webseiten auch Webseiten mit nicht-europäischer Zielgruppe umfassen, welche also die DSGVO nicht berücksichtigen müssen. Ein Unterschied im Tracking erschien daher im Vorfeld plausibel, konnte jedoch nicht bestätigt werden.

Es sollte berücksichtigt werden, dass die Sprachklassifikation als verbesserungswürdig zu betrachten ist. Fehlklassifikationen können zu kleineren Verzerrungen in der Verteilung geführt haben.

Es erscheint angebracht, die Erhebung zu wiederholen, um untersuchen zu können, ob es seit Juli 2018 Veränderungen gegeben hat. Dabei wäre von Interesse, ob sich das Tracking in absoluten Zahlen verringert hat, ob der Anteil an Seiten ohne Tracking gestiegen ist, und ob inzwischen eine stärkere Variabilität zwischen den Gruppen existiert.

5.2.2 Tracking auf deutschsprachigen Seiten: öffentlich-rechtliche und private Medien

Es sollte untersucht werden, ob das Tracking auf Webseiten öffentlich-rechtlicher Medien geringer ausfällt als auf Webseiten privater Medien. Dazu wurden während der Erhebung besuchte öffentlich-rechtliche Webseiten identifiziert. Diese fanden sich in den Kategorien *News And Media* und *Arts And Entertainment*, sodass die anderen Webseiten der beiden Kategorien die Vergleichsgruppe bildeten.

Mittelwerte und Quantile beider Gruppen unterscheiden sich deutlich voneinander. Trotzdem zeigt $d \approx -0{,}306$ noch eine relativ starke Überlappung der Gruppen an. Dies lässt sich in Teilen dadurch erklären, dass die öffentlich-rechtliche Gruppe lediglich 12 Seiten umfasst, von denen zwei Messwerte als Ausreißer einzustufen sind (16 und 31), und die schon größer sind als der Median der privaten Seiten. In der Folge zeigt ein Hypothesentest mit der Medianabweichung als Teststatistik keinen signifikanten Unterschied.

An dieser Stelle sei jedoch an die zuvor festgestellten methodischen Einschränkungen erinnert. Die Vergleichsgruppe beinhaltet zwar Webseiten, die zu erwarten waren wie `bild.de`, `welt.de` oder `focus.de`, allerdings auch Streaminganbieter oder Wetterseiten. Damit beinhaltet die Vergleichsgruppe nicht unbedingt nur die Webseiten, die man vielleicht für einen Vergleich von Webseiten öffentlich-rechtlicher und privater Anbieter heranziehen würde. Durch das verhältnismäßig häufige Scheitern der Sprachklassifikation ist auch anzunehmen, dass in beiden Gruppen relevante Webseiten fehlen.

Der Versuch, durch eine automatisierte Kategorisierung der Webseiten eine große Stichprobe für die Untersuchung zu erhalten, scheint nur mäßig erfolgreich. Das Ergebnis ist daher nur bedingt aussagekräftig. Künftige Untersuchungen, welche den Unterschied zwischen beiden Gruppen beleuchten wollen, sollten die zu besuchenden Seiten auf eine andere Art auswählen.

5.2.3 Information über Tracking und Setzen von Tracking-Cookies

Es sollte untersucht werden, wie Webseiten über den Einsatz von Tracking informieren und ob unmittelbar bei Besuch der Seite Tracking-Cookies gesetzt werden. Dazu wurden stichprobenartig Informationen von 45 Webseiten erhoben.

93,3 % der besuchten Webseiten setzten weiterhin die Zustimmung des Nutzers voraus, obwohl die Zulässigkeit dieser Praxis seit der DSGVO fraglich ist. Überwiegend wurde ohne Widerspruchsmöglichkeit informiert (62,2 %), aber auch überraschend viele zeigten keinerlei Information an (28,9 %), obwohl diese Seiten alle Tracking-Cookies setzten.

Lediglich drei Webseiten setzten auf die Opt-in-Lösung. Von diesen setzten zwei tatsächlich keine Tracking-Cookies vor der Zustimmung des Nutzers, eine tat es dennoch.

Die Erhebung wurde etwa 12 Wochen nach der Anwendbarkeit der DSGVO durchgeführt. Es ist möglich, dass seitdem weitere Webseiten ihren Hinweis angepasst haben, zumal noch im März und April 2018 über die Hälfte von 300 befragten KMU angaben, bisher keinerlei Maßnahmen bezüglich der DSGVO getroffen zu haben [19].

Andererseits existieren, wie in Abschnitt 2.2.2 erwähnt, andere Interpretationen der DSGVO, die davon ausgehen, dass die bisherige Praxis weiterhin zulässig sei. Es ist anzunehmen, dass hier eine Konkretisierung der Rechtslage durch den Gesetzgeber oder in Form einer gerichtlichen Grundsatzentscheidung abgewartet wird.

Es scheint daher lohnenswert, in regelmäßigen Abständen, insbesondere aber nach einer Konkretisierung der Rechtslage, die Untersuchung zu wiederholen.

5.2.4 Tracking-Firmen auf deutschsprachigen und nicht-deutschsprachigen Webseiten

Es sollte untersucht werden, wie stark Tracking-Unternehmen auf deutschsprachigen und nicht-deutschsprachigen Webseiten vertreten sind. Es wurde sich auf Unternehmen beschränkt, welche auf mindestens 5 % aller deutschsprachigen oder nicht-deutschsprachigen Seiten vertreten waren.

Dabei wurde festgestellt, dass die Unternehmen grundsätzlich ähnlich verteilt sind. Tracking-Unternehmen, welche auf deutschsprachigen Webseiten wahrnehmbar vertreten sind, sind dies in der Regel auch auf nicht-deutschsprachigen Seiten und umgekehrt. Ausnahmen sind die Firmen LiveInternet, vermutlich ein russisches Informationsportal, welches auf deutschsprachigen Seiten nur wenig vertreten ist, und INFOnline, ein deutsches Analytics-Unternehmen, welches außerhalb von Deutschland nahezu unbekannt zu sein scheint.

6 Zusammenfassung

Diese Arbeit hat versucht das Ausmaß des User-Tracking auf deutschsprachigen Webseiten zu erfassen. Dazu wurde eine quantitative Studie durchgeführt, in deren Rahmen 10 000 Webseiten besucht wurden. Die besuchten Webseiten wurden hinsichtlich ihrer Sprache und Kategorie klassifiziert und jeweils die Anzahl der als Tracker zu bewertenden Anfragen bestimmt. Zusätzlich wurde stichprobenartig untersucht, wie Webseiten ihre Nutzer über stattfindendes Tracking informieren und ob unmittelbar nach Aufruf der Seite Tracking-Cookies gesetzt werden.

In Bezug auf die formulierten Fragestellungen kann folgendes festgehalten werden:

1. Es wurde festgestellt, dass deutschsprachige Webseiten nicht weniger Tracking einsetzen als nicht-deutschsprachige Webseiten.
2. Die mittlere öffentlich-rechtliche Medienseite weist in dieser Untersuchung nicht signifikant weniger Tracker auf als die mittlere private Medienseite, obwohl sich die Stichprobenstatistiken augenscheinlich stark unterscheiden.
3. Bei der Untersuchung der Nutzerinformation über stattfindendes Tracking konnte festgestellt werden, dass die Opt-in-Lösung bisher keine weite Verbreitung erfährt. Stattdessen setzen Webseiten weiterhin mehrheitlich auf Informationsbanner. Eine von insgesamt drei Webseiten der Stichprobe, welche bereits die Opt-in-Lösung nutzten, setzte dennoch Tracking-Cookies, ohne die Zustimmung des Nutzers abzuwarten.
4. Größere Tracking-Unternehmen sind auf deutschsprachigen und nicht-deutschsprachigen Webseiten ähnlich stark vertreten. Es konntde lediglich ein Unternehmen identifiziert werden, welches auf deutschsprachigen Webseiten relativ stark vertreten und gleichzeitig auf nicht-deutschsprachigen Webseiten nahezu unbekannt ist.

Während der Arbeit konnten einige methodische Unzulänglichkeiten identifiziert werden. Die Genauigkeit des Verfahrens zur Bestimmung der Deutschsprachigkeit scheint verbesserungswürdig, ein erster Ansatzpunkte konnte jedoch aufgezeigt werden. Des Weiteren schienen die Ergebnisse der Kategorisierung für die Untersuchung des Tracking auf öffentlich-rechtlich und privaten Medienseiten nur mäßig geeignet. Die Ergebnisse diesbezüglich können daher nur als bedingt aussagekräftig bezeichnet werden.

Nachfolgende Arbeiten können daher an einer Verbesserung der Methodik ansetzen. Darüber hinaus stellen Untersuchungen wie diese stets nur eine Momentaufnahme dar. Die Durchführung weiterer Erhebungen würde es erlauben, den Wandel des Ausmaßes des User-Tracking nachzuvollziehen. Die verschiedenen Interpretationen der Rechtslage lassen vermuten, dass mit der Klärung dieser weitere Verschiebungen zu erwarten sind.

Literaturverzeichnis

[1] Alexa Internet: *How are Alexa's traffic rankings determined? – Alexa Support.* – URL https://support.alexa.com/hc/en-us/articles/200449744. – Zugriffsdatum: 04.02.2019

[2] BBC: *GDPR: US news sites unavailable to EU users under new rules - BBC News.* – URL https://www.bbc.com/news/world-europe-44248448. – Zugriffsdatum: 21.02.2019

[3] Bujlow, Tomasz ; Carela-Espanol, Valentin ; Solé-Pareta, Josep ; Barlet-Ros, Pere: A Survey on Web Tracking: Mechanisms, Implications, and Defenses. In: *Proceedings of the IEEE* 105 (2017), 03, S. 1–35. – URL https://doi.org/10.1109/JPROC.2016.2637878

[4] Cliff, Norman: *Ordinal Methods for Behavioral Data Analysis.* Psychology Press, März 2014. – URL https://doi.org/10.4324/9781315806730

[5] Statistisches Bundesamt: *Informationsgesellschaft in Deutschland 2009.* URL https://www.destatis.de/DE/Publikationen/Thematisch/EinkommenKonsumLebensbedingungen/Querschnitt/Informationsgesellschaft.html, November 2009. – ISBN 978-3-8246-0868-3

[6] EasyList: *EasyList - Policy.* – URL https://easylist.to/pages/policy.html. – Zugriffsdatum: 14.02.2019

[7] Englehardt, Steven ; Narayanan, Arvind: *Online tracking: A 1-million-site measurement and analysis.* – URL https://webtransparency.cs.princeton.edu/webcensus/. – Zugriffsdatum: 5.02.2019

[8] Englehardt, Steven ; Narayanan, Arvind: Online Tracking: A 1-million-site Measurement and Analysis. In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. New York, NY, USA : ACM, 2016 (CCS '16), S. 1388–1401. – URL http://doi.acm.org/10.1145/2976749.2978313. – ISBN 978-1-4503-4139-4

[9] Eurostat: *Internet advertising of businesses - statistics on usage of ads*. 2016. – URL https://ec.europa.eu/eurostat/statistics-explained/index.php/Internet_advertising_of_businesses_-_statistics_on_usage_of_ads. – Zugriffsdatum: 28.01.2019

[10] Facebook: *Facebook-Pixel - Dokumentation - Facebook for Developers*. – URL https://developers.facebook.com/docs/facebook-pixel. – Zugriffsdatum: 29.01.2019

[11] Facebook: *Targeting-Optionen für Facebook-Werbung | Facebook Business*. – URL https://www.facebook.com/business/ads/ad-targeting. – Zugriffsdatum: 28.01.2019

[12] Fielding, Roy T. ; Gettys, James ; Mogul, Jeffrey C. ; Nielsen, Henrik F. ; Masinter, Larry ; Leach, Paul J. ; Berners-Lee, Tim: Hypertext Transfer Protocol – HTTP/1.1 / RFC Editor. RFC Editor, June 1999 (2616). – Forschungsbericht. – URL http://www.rfc-editor.org/rfc/rfc2616.txt. – ISSN 2070-1721

[13] Fruchter, Nathaniel ; Miao, Hsin ; Stevenson, Scott ; Balebako, Rebecca: Variations in Tracking in Relation to Geographic Location. In: *CoRR* abs/1506.04103 (2015). – URL http://arxiv.org/abs/1506.04103

[14] GitHub: *Crawl crashes after AssertionError in the 'tab_restart_browser' · Issue #166 · mozilla/OpenWPM · GitHub*. – URL https://github.com/mozilla/OpenWPM/issues/166. – Zugriffsdatum: 19.02.2019

[15] GitHub: *GitHub - brave/ad-block: Ad block engine used in the Brave browser for ABP filter syntax based lists like EasyList.*. – URL https://github.com/brave/ad-block. – Zugriffsdatum: 21.02.2019

[16] GitHub: *GitHub - disconnectme/disconnect-tracking-protection: Canonical repository for the Disconnect services file.* – URL https://github.com/disconnectme/disconnect-tracking-protection. – Zugriffsdatum: 14.02.2019

[17] Github: *GitHub - mozilla/OpenWPM: A web privacy measurement framework.* – URL https://github.com/mozilla/OpenWPM. – Zugriffsdatum: 05.02.2019

[18] Google: *Analysetools und -lösungen für Ihr Unternehmen – Google Analytics.* – URL https://marketingplatform.google.com/about/analytics/. – Zugriffsdatum: 29.01.2019

[19] Haufe: *DSGVO laut Umfragen bei KMU weiterhin erschreckend unbekannt | Compliance | Haufe.* – URL https://www.haufe.de/compliance/recht-politik/dsgvo-laut-umfragen-bei-kmu-weiterhin-erschreckend-unbekannt_230132_449834.html. – Zugriffsdatum: 20.02.2019

[20] Hedderich, Jürgen ; Sachs, Lothar: *Angewandte Statistik.* Springer Berlin Heidelberg, 2018. – URL https://doi.org/10.1007/978-3-662-56657-2

[21] Faulkner, Steve (Hrsg.) ; Eicholz, Arron (Hrsg.) ; Leithead, Travis (Hrsg.) ; Danilo, Alex (Hrsg.) ; Moon, Sangwhan (Hrsg.): *HTML 5.2.* 2017. – URL https://www.w3.org/TR/2017/REC-html52-20171214/. – Zugriffsdatum: 22.01.2019

[22] Janczyk, Markus ; Pfister, Roland: *Inferenzstatistik verstehen.* Springer Berlin Heidelberg, 2015. – URL https://doi.org/10.1007/978-3-662-47106-7

[23] Krempl, Stefan: *DSGVO und Telemedien: Cookies und Tracking nur noch mit expliziter Einwilligung?* Mai 2018. – URL https://www.heise.de/newsticker/meldung/DSGVO-und-Telemedien-Cookies-und-Tracking-nur-noch-mit-expliziter-Einwilligung-4045908.html. – Zugriffsdatum: 12.02.2019

[24] Libert, T ; Graves, L ; Nielsen, RK: Changes in third-party content on European news websites after GDPR. Reuters Institute for the Study of Journalism, 2018. – Forschungsbericht. – URL https://ora.ox.ac.uk/objects/uuid:5a5d4eea-6e74-49b4-8c77-71ec6760f127

[25] Mozilla Foundation: *Public Suffix List.* – URL https://publicsuffix.org/. – Zugriffsdatum: 16.02.2019

[26] Olson, David L. ; Delen, Dursun: *Advanced Data Mining Techniques.* First Edition. Springer Publishing Company, Incorporated, 2008. – URL https://doi.org/10.1007/978-3-540-76917-0. – ISBN 3540769161, 9783540769163

[27] Ramsey, Fred ; Schafer, Daniel: *The Statistical Sleuth: A Course in Methods of Data Analysis.* Third Edition. Cengage Learning, 2013. – ISBN 978-1-133-49067-8

[28] Romano, J. ; Kromrey, J. D. ; Coraggio, J. ; Skowronek, J.: Appropriate statistics for ordinal level data: Should we really be using t-test and cohen's d for evaluating group differences on the NSSE and other surveys?, 2006

[29] Schneider, Markus ; Enzmann, Matthias ; Stopczynski, Martin ; Waidner, Michael (Hrsg.): Web-Tracking-Report 2014. Fraunhofer-Verlag, Stuttgart, 2014. – Forschungsbericht. – 113 S. – URL https://www.sit.fraunhofer.de/de/wtr/. – ISBN 978-3-8396-0700-8

[30] SimilarSites: *Similarsites.com - Easily Find Similar Websites.* – URL https://www.similarsites.com/. – Zugriffsdatum: 14.02.2019

[31] Sinus Institut: *Studie zu Datenschutz: Mehrheit der Deutschen zweifelt an Datensicherheit.* – URL https://www.sinus-institut.de/veroeffentlichungen/meldungen/detail/news/studie-zu-datenschutz-mehrheit-der-deutschen-zweifelt-an-datensicherheit/news-a/show/news-c/NewsItem/. – Zugriffsdatum: 20.02.2019

[32] Statistisches Bundesamt: *Private Haushalte in der Informationsgesellschaft (IKT).* 2018. – URL https://www.destatis.de/DE/Publikationen/Thematisch/EinkommenKonsumLebensbedingungen/PrivateHaushalte/PrivateHaushalteIKT.html. – Zugriffsdatum: 28.01.2019

[33] The Apache Software Foundation: *Apache Tika - a content analysis toolkit.* – URL https://tika.apache.org/. – Zugriffsdatum: 23.01.2019

[34] Torchiano, Marco: *cliff.delta function | R Documentation.* – URL https://www.rdocumentation.org/packages/effsize/versions/0.7.4/topics/cliff.delta. – Zugriffsdatum: 18.02.2019

[35] Voigt, Paul ; Bussche, Axel von dem: *EU-Datenschutz-Grundverordnung (DSGVO).* Springer Berlin Heidelberg, 2018. – URL https://doi.org/10.1007/978-3-662-56187-4

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

Hamburg, 25. Februar 2019 Niklas Hagemann