Den Einfluss der Suchmaschinenoptimierung messbar machen

Ein halb-automatisierter Ansatz zur Bestimmung von optimierten Ergebnissen auf Googles Suchergebnisseiten

Sebastian Sünkler, Dirk Lewandowski

Hamburg University of Applied Sciences, Department of Information,
Finkenau 35, 22081 Hamburg (Germany)
{sebastian.suenkler, dirk.lewandowski}@haw-hamburg.de

Abstract

Suchmaschinenoptimierung (SEO) kann einen erheblichen Einfluss auf die Suchergebnisseiten in kommerziellen Suchmaschinen haben. Bisher ist jedoch unklar, welcher Anteil der (Top-)Ergebnisse tatsächlich optimiert wurde und wie dies die Auswahl und das Ranking der angezeigten Treffer bestimmt. Für die Untersuchung über den Einfluss von SEO auf die Suchergebnisse haben wir halb-automatische Prozesse und ein Software-Framework entwickelt, das mit einem regelbasierten Klassifikator die Wahrscheinlichkeit von Optimierungsmaßnahmen auf Suchergebnissen zu ermittelt. Der Ansatz basiert auf 20 Merkmalen, zu denen die Analyse über genutzte SEO-Plug-ins und Analytics Tools, die Auswertung technischer Indikatoren wie der Ladegeschwindigkeit und der Verwendung bestimmter Tags wie zur Description sowie eine manuelle Klassifikation auf Basis von vorab zusammengestellten Listen mit optimierten und nicht optimierten Webseiten gehören. Dieser Ansatz wurde auf drei Datensätzen mit insgesamt 2.043 Suchanfragen und 263.790 Ergebnissen angewendet. Die Ergebnisse zeigen, dass ein großer Teil der in Google gefundenen Seiten zumindest wahrscheinlich optimiert ist, was im Einklang mit Aussagen von SEO-Experten steht, die sagen, dass es sehr schwierig ist, ohne die Anwendung von SEO-Techniken in Suchmaschinen sichtbar zu werden.

Keywords: Suchmaschinen; Suchmaschinenoptimierung; SEO; Datenanalyse

1 Einleitung

Mit Suchmaschinenoptimierung (SEO) werden Maßnahmen bezeichnet, die das Ziel haben, Ranking-Algorithmen von Suchmaschinen zu entsprechen, um dadurch eine verbesserte Platzierung in den organischen Suchergebnissen zu erreichen (Griesbaum, 2013). Sie kann dadurch zwischen der Unterstützung von Suchmaschinen beim Auffinden und Indexieren relevanter Inhalte und der Manipulation ihrer Ergebnisse eingeordnet werden. Inhaltsanbieter wie Online-Shops und Medienangebote sind auf gute Platzierungen in den Suchergebnissen in kommerziellen Suchmaschinen angewiesen, um Traffic für ihre Angebote zu generieren, da Suchmaschinen den wichtigsten Zugang zu Inhalten im Web darstellen (Van Couvering, 2009). Dadurch entsteht eine oft stark ausgeprägte Abhängigkeit insbesondere von Google, dem Marktführer in der Websuche. Google hat einen Marktanteil von 87% in den Vereinigten Staaten (StatCounter, 2020a) und 93% in Europa (StatCounter, 2020b) über alle Plattformen hinweg (Stand: August 2020). Durch die hohe Relevanz von Suchmaschinenoptimierung für Webseiten hat sich eine starke Branche für SEO ausgebildet, die beispielsweise in den USA im Jahr 2020 einen Umsatz von 80 Milliarden Dollar erreichen wird (McCue, 2018). Eine zentrale Frage, die sich daraus ergibt, ist, inwieweit Ergebnislisten in Suchmaschinen durch Suchmaschinenoptimierung (SEO) extern beeinflusst werden. Nach unserem Wissen gibt es keine Untersuchungen darüber, wie man technisch messen kann, ob eine Webseite (oder Website) durch SEO-Maßnahmen „manipuliert“ wurde.

Die Untersuchung dieses Themas ist für die Informationswissenschaft auf verschiedenen Ebenen relevant. Für den Bereich des Information Retrieval (IR) ist sie von Bedeutung, da es externe Einflüsse auf die Ergebnisse von Information-Retrieval-Systemen berücksichtigt und somit unsere Sichtweise auf IR-Systeme erweitert wird. Abgesehen von der Konzentration auf Maßnahmen, die Anbieter von IR-Systemen ergreifen können, um ihre Ranking-Funktionen zu verbessern, hat sich die bisherige Forschung darauf konzentriert, die Benutzer durch die Analyse von Klicks und weiteren Interaktionen mit den Ergebnissen in die Ranking-Modelle einzubeziehen (z.B. Wang et al., 2016; Zamani et al., 2017). Darüber hinaus gibt es einige Forschungsarbeiten über die Eigeninteressen von Suchmaschinenunternehmen und darüber, wie diese das, was auf den Ergebnisseiten angezeigt wird, beeinflussen können (z.B. Lewandowski & Sünkler, 2013). Einige Forschungs-

arbeiten haben sich auch auf die Mischung aus bezahlten und organischen Ergebnissen auf den Ergebnisseiten von Suchmaschinen (SERPs) konzentriert und darauf, wie Nutzer, die zwischen den beiden Ergebnistypen unterscheiden können bzw. nicht unterscheiden können, ein unterschiedliches Auswahlverhalten zeigen (z. B. Lewandowski, 2017).

Neben den genannten Aspekten sind auch die Fragen über den Einfluss von suchmaschinenoptimierten Inhalten auf den Wissenserwerb von Suchmaschinennutzenden und auch im Bereich der Infomationskompetenz relevant. Suchmaschinenoptimierung kann als Teil von Suchmaschinenwerbung gesehen werden, da diese dazu führen soll, dass Inhalte prominenter in Suchergebnissen auftauchen. Dazu werden Maßnahmen durchgeführt, die zum Teil auch als Manipulation des Rankings eingesetzt werden. Zur Suchmaschinenwerbung gehört auch das Suchmaschinenmarketing (SEM) durch die Schaltung von Textanzeigen. Diese unterscheiden sich von den organischen Ergebnissen vor allem dadurch, dass Suchmaschinen sie als Werbung kennzeichnen; für suchmaschinenoptimierte Inhalte gibt es allerdings keine entsprechende Kennzeichnung. Während bereits Studien zu der Wahrnehmung von Werbeanzeigen in Suchmaschinen durchgeführt wurden und zeigen, dass Benutzer Schwierigkeiten bei der Unterscheidung zwischen Werbung und organischen Suchergebnissen haben (Lewandowski et al., 2018), gibt es keine Forschung zur Wahrnehmung von SEO – auch nicht dazu, wie sich diese Inhalte auf den Wissenserwerb auswirken können.

Für die Durchführung von Studien zu den genannten Bereichen ist es wichtig, dass Möglichkeiten geschaffen werden, um Aussagen darüber zu treffen, ob Dokumente suchmaschinenoptimiert sind oder nicht. In diesem Beitrag stellen wir einen Ansatz zur Identifizierung von suchmaschinenoptimierten Inhalten vor. Dabei wird aus einer Kombination von Analyseverfahren eine Aussage über die Wahrscheinlichkeit von SEO auf einer Webseite getroffen. In diesem Beitrag beschreiben wir zunächst die Vorgehensweise zur Identifizierung der SEO-Merkmale als Basis für unseren regelbasierten Klassifikator zur Einordnung der URLs. Darauf folgend stellen wir erste Ergebnisse über den Einsatz unseres Ansatzes auf drei Datensätzen mit insgesamt 263.790 Suchergebnissen vor. Anschließend folgen ein Fazit und eine Darstellung des weiteren Vorgehens zur Weiterentwicklung unserer Methoden.

2 Identifizierung von SEO-Faktoren

Unsere Diskussion der SEO-Faktoren und ihrer Umsetzung in unserem System basiert auf einer ausführlichen Durchsicht der Fachliteratur (z.B. Enge, 2015; Erlhofer, 2019) und Interviews mit SEO-Experten (Schultheiß/Lewandowski, 2020). Insgesamt besteht unser Modell aus 47 Faktoren, die wir für die Umsetzung in unserem System priorisiert haben. Es sei darauf hingewiesen, dass das aktuelle Modell nur 20 Faktoren berücksichtigt.

Da dies jedoch diejenigen sind, die von den Experten und Forschern als die fruchtbarsten angesehen wurden, sind wir davon überzeugt, dass wir bereits in diesem Stadium zuverlässig optimierte Inhalte identifizieren können. Bei der Auswahl der Kriterien haben wir uns ferner auch nicht daran orientiert, ob es sich um Merkmale handelt, die auf gute oder erfolgreiche SEO-Praktiken hinweisen. Auf den ersten Blick könnte man annehmen, dass Suchmaschinenoptimierer versuchen, mit ihren Optimierungsmaßnahmen die Ranking-Faktoren von Google zu bedienen und so zu beeinflussen, dass ihre Inhalte sichtbarer werden. Dadurch wären SEO-Faktoren und Ranking-Faktoren identisch. Allerdings sind zum einen die genauen Ranking-Faktoren unbekannt und zum anderen sind auch nicht alle SEO-Bemühungen erfolgreich. Einige Methoden werden auch von den Suchmaschinen bestraft oder einfach ignoriert. Zum Beispiel könnte jemand, der versucht, Sichtbarkeit in Google zu erlangen, Keyword-Stuffing verwenden, d.h. ein Keyword auf einer Webseite sehr oft wiederholen, um der Suchmaschine zu suggerieren, dass die Seite für dieses Keyword relevant ist. Offensichtlich wird dieser Ansatz nicht funktionieren, da Suchmaschinen solche einfachen Manipulationsversuche erkennen können. Da wir jedoch feststellen wollen, ob Inhaltsanbieter versuchen, ihre Seiten zu optimieren, kann das Keyword-Stuffing immer noch ein Faktor zur Erkennung optimierter Seiten sein. Ob Inhaltsanbieter ihre Seiten erfolgreich optimiert haben, ist kein Kriterium, das für unsere Klassifizierung relevant ist. Wir nutzen verschiedene halb-automatisierte Methoden, um die Wahrscheinlichkeit von Suchmaschinenoptimierung zu bewerten.

2.1 Halbautomatische Analysemethoden zur Bewertung der SEO-Wahrscheinlichkeit auf einer Webseite

Für die abschließende Analyse der Suchmaschinenoptimierung sind drei Prozesse notwendig, die zum einen zur Berechnung der Input-Variablen für die regelbasierte Klassifikation notwendig sind und zum anderen die eigentliche Klassifikation durchführen. Abbildung 1 zeigt die Prozesse, die in folgende Schritte eingeteilt sind:

1. Abrufen der URL, um den HTML-Code und die Metadaten zu extrahieren
2. Generierung des Inputs für den regelbasierten Klassifikator auf drei Stufen und
3. Bestimmen der Wahrscheinlichkeit von SEO.

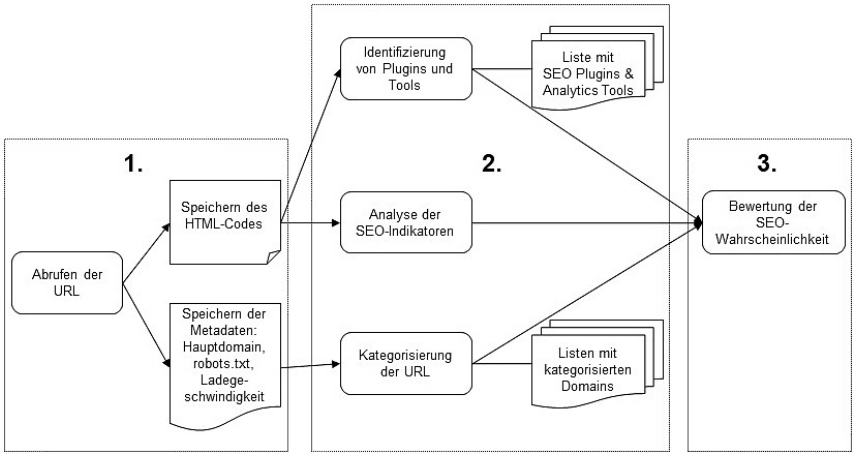

Abb. 1 Halb-automatisierte Methoden zur Bewertung der SEO-Wahrscheinlichkeit auf einer Webseite

Im ersten Schritt findet ein Scraping der Ergebnisseiten von Suchmaschinen statt. Scraping bezeichnet allgemein den Vorgang, durch technische Hilfsmittel automatisiert auf Texte und Webseiten zuzugreifen, um Informationen daraus zu extrahieren und für weitere Zwecke, z. B. Analysen, verfügbar zu machen. In diesem Fall werden die URLs von den Suchergebnisseiten von Google extrahiert und in einer Datenbank gespeichert. Dabei werden sowohl der HTML-Quelltext als auch einige zusätzliche Daten wie die Hauptdomain, der Inhalt der robots.txt und die Ladegeschwindigkeit der

URL erfasst. Anschließend werden SEO-Indikatoren sowie SEO-Plug-ins and Analytics Tools aus dem Quelltext extrahiert und die URL wird mit Listen abgeglichen, die optimierte und nicht optimierte Dokumente enthalten. Die Ergebnisse aus diesem Schritt sind die Grundlage für die abschließende Bewertung. Im Folgenden werden die einzelnen Prozesse aus dem zweiten und dritten Schritt näher erläutert.

2.2 Identifizierung von Plug-ins und Tools

Für die Identifizierung von SEO-Plug-ins und Tools wurden manuelle Listen erstellt, die sowohl die Bezeichnung als auch ein Suchmuster enthalten, um Hinweise auf die Nutzung im Quelltext der URL zu finden. Eine Nutzung von SEO-Plug-ins ist ein eindeutiger Hinweis darauf, dass Optimierungsmaßnahmen eingesetzt werden, aber auch Analytics Tools lassen zumindest darauf schließen, dass Webanalysen durchgeführt werden. Webanalysen sind ein sehr guter Indikator für ein kommerzielles Interesse des Anbieters und da insbesondere kommerzielle Anbieter Maßnahmen für gute Positionen im Suchmaschinenranking durchführen, sind Analytics Tools ebenfalls ein guter Indikator.

Bei den SEO-Plug-ins umfasst die Liste momentan 58 Plug-ins und bei den Analytics Tools wurden bisher 54 Tools manuell in einer Datenmenge von 30.000 gespeicherten Quelltexten identifiziert. Die folgen Ausschnitte aus dem HTML-Quelltext zeigen jeweils ein Beispiel für ein SEO-Plug-in und ein Analytics Tool:

SEO-Plug-in: Yoast SEO Plugin[1]

HTML-Code: <!--This site is optimized with the Yoast SEO plugin v12.4 - https://yoast.com/wordpress/plugins/seo/-->

Suchmuster: "*yoast seo*"

Analytics Tool: Google Analytics[2]

HTML-Code: <!-- Google Analytics --> <script>(function(i,s,o,g,r,a,m){i['GoogleAnalyticsObject']=r;…

Suchmuster: "*google analytics*"

1 https://yoast.com/wordpress/plugins/seo/

2 https://analytics.google.com/analytics/web/

2.3 Kategorisierung der Seite anhand bekannter Domains

Für unseren Ansatz kombinieren wir die automatische Identifizierung von SEO-Indikatoren aus Webseiten und Websites mit der manuellen Generierung von Listen mit optimierten und nicht optimierten Websites. Wir verfolgen diesen Ansatz, da die Ergebnisse in den Suchmaschinen nicht gleichmäßig verteilt sind, d. h. es gibt nur eine relativ kleine Anzahl von Webseiten, die einen großen Teil der gezeigten URLs ausmachen (Petrescu, 2014) und in den Top-Ergebnissen angeklickt werden (Goel et al., 2010). Dies bedeutet, dass wir durch die manuelle Klassifizierung einer begrenzten Anzahl von Websites bereits einen relativ großen Teil der optimierten Seiten erkennen können (Petrescu, 2014). Für die Klassifizierung haben wir Listen mit Websites nach Kategorien zusammengestellt, die wir anhand der Wahrscheinlichkeit von Suchmaschinenoptimierung in den Kategorien unterteilt haben. Insgesamt wurden sechs Kategorien definiert. Die erste Kategorie ist eine Liste mit Kunden von SEO-Agenturen, da davon ausgegangen werden kann, dass Kunden von solchen Agenturen Suchmaschinenoptimierung betreiben.

Die Agenturen wurden anhand von Recherchen in Fachportalen identifiziert und die Websites für die Liste anhand der Kundenreferenzen auf den Seiten der Agenturen ermittelt. Durch diese Vorgehensweise konnten insgesamt 1.004 Websites identifiziert werden. Eine weitere Kategorie bezieht sich auf Websites, die ganz sicher keine Suchmaschinenoptimierung betreiben. Da unser Ansatz noch in der Entwicklung ist, sind Aussagen zu solchen Seiten schwer zu treffen. Daher befindet sich momentan nur Wikipedia in dieser Liste, da von Wikipedia bekannt ist, dass dort keine aktiven SEO-Maßnahmen genutzt werden. Für die weiteren Kategorien wurde ein Datensatz mit 13.403 Dokumenten manuell ausgewertet. Die URLs wurden dabei klassifiziert und in die restlichen Kategorien eingeteilt. Wir sind dabei von den möglichen Interessen der Seitenbetreiber ausgegangen. So ist definiert, dass klare kommerzielle Absichten die Wahrscheinlichkeit von SEO erhöhen. Die Kategorie, die dabei die meisten Dokumente enthält, sind Nachrichten-Websites wie beispielsweise spiegel.de. Insgesamt konnten 1.203 Nachrichtenangebote im Datensatz gefunden werden. Die weiteren Kategorien sind Online-Shops (178 Seiten), Websites, die Werbeanzeigen schalten (325 Seiten), und Unternehmens-Webseiten (72 Seiten). Alle Dokumente in den Listen sind dabei allerdings nicht exklusiv eingeordnet. Es sind Überschneidungen möglich. Die Listen werden fortlaufend erweitert und dabei ist auch insbesondere die Erweiterung der nicht optimierten Webseiten hochrelevant, da bisher nur eine

Domain in diese Liste aufgenommen wurde. Zum jetzigen Zeitpunkt ist eine ganz konkrete Einordnung solcher Seiten allerdings schwierig, da unser Klassifikator noch nicht abschließend und genau genug die SEO-Wahrscheinlichkeit bestimmten kann, um konkret auszusagen, dass eine Seite nicht optimiert ist.

2.4 Analyse der SEO-Indikatoren

Bei den SEO-Indikatoren extrahieren wir Informationen direkt aus dem HTML-Quelltext der Seiten und speichern dazu auch weitere Merkmale wie die Hauptdomain, die Inhalte der robots.txt und die Geschwindigkeit, mit der die Seite vollständig geladen wird. Bei den aus dem Quelltext gewonnen Daten handelt es sich beispielsweise um den Seitentitel (Title Tag), die Seitenbeschreibung (Description Tag) und Nofollow-Links. Ferner wird noch geprüft, ob eine Sitemap vorhanden ist. In Bezug auf die robots.txt werden typische Inhalte gesucht, die auf SEO hinweisen. Eine robots.txt wird genutzt, um Anweisungen an Suchmaschinen-Crawler zu geben. Wir messen, ob diese vorhanden ist und SEO-relevante Anweisungen enthält (z. B. explizite Ausschlüsse von Inhalten auf der Domain für die Crawler). Schließlich messen wir die Geschwindigkeit der Seiten, da einer der grundlegenden technischen Faktoren bei der Suchmaschinenoptimierung die Optimierung der Seiten für ein schnelles Laden ist. Der Vorteil bei dieser Vorgehensweise liegt darin, dass wir unabhängig von externen Indikatoren agieren können, da wir nur auf direkt verfügbare technische Informationen zugreifen.

Durch die Kombination mit der Auswertung der Plug-ins und Tools sowie der Kategorisierung der URL durch die Abgleiche mit bekannten Domains wird eine Basis für die Einschätzung der Suchmaschinenoptimierung geschaffen. Eine vollständige Übersicht der genutzten Merkmale findet sich in Tabelle 1.

Die halb automatische Erfassung und Analyse dieser Merkmale bildet die Grundlage für die regelbasierte Klassifikation, um die Wahrscheinlichkeit von Suchmaschinenoptimierung auf der Seite zu bewerten.

Tab. 1: SEO-Merkmale

Merkmal	Beschreibung	Beispiel
SEO Plugins und Analytics Tools		
SEO Plug-ins	Tools, die direkt zur Suchmaschinenoptimierung eingesetzt werden	Yoast SEO Plugin
Analytics Tools	Tools, die für das Tracking und für die Webseitenanalyse verwendet werden	Google Analytics
URL-Listen		
Kunden von SEO-Agenturen	In dieser Liste sind Domains hinterlegt, die zu Kunden von SEO-Agenturen gehören (1.004 Seiten).	faz.net
nicht optimierte Seiten	Liste mit Domains, die garantiert keine Suchmaschinenoptimierung betreiben (1 Seite)	wikipedia.de
Nachrichtendienste	Domains von Nachrichtendiensten (1.203 Seiten).	spiegel.de
Online-Shops	Liste mit Domains von Online-Shops (178 Seiten)	amazon.de
Unternehmens-Webseiten	Liste mit Domains von Unternehmen (72 Seiten)	dhl.de
Webseiten mit Werbung	Liste mit Domains, die Werbeanzeigen schalten (325 Seiten)	facebook.com
Technische Indikatoren		
Microdata-Formate	Nutzung von Microdata-Formaten auf einer Webseite strukturiert	JSON-LD
Textanzeigen auf der Seite	Nutzung von Diensten zur automatischen Schaltung von kontextualisierten Werbeanzeigen	Google Ads
HTTPS	Verwendung des Hypertext Transfer Protocol Secure	–
SEO in der robots.txt	Hinweise zur Konfiguration von Suchmaschinen-Crawler in der robots.txt, die zu der Hauptdomain der Seite gehört	crawl-delay
Sitemap	Verwendung einer Sitemap auf der Seite für eine bessere Navigation	–
Viewport	Definition eines Viewports für eine responsive Darstellung der Webseite.	<meta name="viewport" content="width=device-width, initial-scale=1.0">

Merkmal	Beschreibung	Beispiel
Nofollow-Links	Nofollow-Links werden gesetzt, um Suchmaschinen anzuweisen, diese Links nicht für das Ranking zu berücksichtigen. Die Anweisung wird über das rel-Tag gesetzt.	<a href ="example.com" rel="nofollow">
Canonical-Links	Canonical-Links werden gesetzt, um bei mehrfach verwendeten Inhalten auf die Originalressource zu verweisen. Die Anweisung wird über das rel-Tag gesetzt.	<a href ="example.com" rel="canonical">
Ladegeschwindigkeit	Messen der Ladegeschwindigkeit. Wenn der Wert unter drei Sekunden liegt, ist das ein Indikator für optimierte Inhalte.	driver.execute_script(""" var loadTime = ((window.performance.timing.domComplete - window.performance.timing.navigationStart)/1000); return loadTime;) """
Description	Überprüfung, ob eine Description auf der Seite vorhanden ist. Das Tag für die Beschreibung der Seite; die Beschreibung kann dabei sowohl in den meta-Tags einer Seite und in Open Graph Tags gesetzt werden.	<meta name="description" content="example" />
Title	Überprüfung, ob das Tag für den Titel der Seite gesetzt ist. Der Titel wird normalerweise in dem dafür vorgesehenen Tag gesetzt. Es ist aber auch möglich, dafür Open Graph Tags zu verwenden.	<title>example title</title>
Open Graph Tags	Überprüfung, ob Open Graph Tags auf der Seite verwendet werden.	<meta property="og:title" content="example" />

2.5 Regelbasierte Klassifikation der SEO-Wahrscheinlichkeit

Durch die regelbasierte Klassifikation wird eine Aussage darüber getroffen, ob eine Dokument höchstwahrscheinlich optimiert, wahrscheinlich optimiert, wahrscheinlich nicht optimiert oder höchstwahrscheinlich nicht optimiert ist. Die Regeln werden im Folgenden näher erläutert.

2.5.1 Höchstwahrscheinlich optimiert

Ein Dokument ist höchstwahrscheinlich optimiert, wenn entweder mindestens ein SEO-Plug-in im Quelltext identifiziert wurde oder die Seite Kunde einer SEO-Agentur ist oder die Seite ein Nachrichtenangebot ist oder Werbeanzeigen auf der Seite geschaltet sind oder Microdata-Formate gefunden wurden.

Bei dieser Regel wurden die eindeutigsten Merkmale für SEO genutzt. So sind die Verwendung von SEO-Plug-ins und die Zusammenarbeit mit einer SEO-Agentur ganz klare Hinweise für Suchmaschinenoptimierung. Zusätzlich sind Nachrichtenangebote als Informationsangebote stark an guten Positionen in den Suchergebnissen interessiert und daher ist auch hier von einer ausgeprägten SEO-Intention auszugehen. Diese Annahme wurde durch unseren Beirat von Suchmaschinenexperten bestätigt. Zusätzlich sind eindeutige kommerzielle Absichten mit der Schaltung von Werbeanzeigen ein klarer Hinweis auf SEO und auch Microdata-Formate werden häufig dafür genutzt, sogenannte Rich Snippets, also angereicherte Trefferbeschreibungen für Suchmaschinen, zu generieren (Ronallo, 2012).

2.5.2 Wahrscheinlich optimiert

Die Seite ist wahrscheinlich optimiert, wenn sie nicht höchstwahrscheinlich optimiert ist und mindestens eines der folgenden Kriterien erfüllt: (1) Die Seite ist ein Online-Shop oder eine Unternehmensseite, (2) auf der Seite wurde mindestens ein Analytics Tools identifiziert, (3) als Übertragungsprotokoll wird HTTPS eingesetzt, (4) es wurden Hinweise zu SEO in der robots.txt gefunden, (5) die Seite hat eine Sitemap, (6) ein Viewport ist definiert, (7) es wurde mindestens ein Nofollow- oder ein Canonical-Link gefunden, (8) die Ladegeschwindigkeit der Seite liegt unter drei Sekunden.

Bei dieser Regel wird eine Vielzahl von Merkmalen einbezogen, die Hinweise auf eine wahrscheinliche Suchmaschinenoptimierung geben. Hier wird davon ausgegangen, dass kommerzielle Anbieter wie Online-Shops und Unternehmensseiten sowie Seiten, die Programme für kontextualisierte Werbeanzeigen wie Google Ads verwenden, ein großes Interesse an guten Positionen im Ranking von Suchmaschinen haben. Die weiteren Indikatoren, die geprüft werden, ergeben sich aus der Analyse des Quelltexts, der robots.txt und der gemessenen Ladegeschwindigkeit. Hier werden die Nutzung von HTTPS als Übertragungsprotokoll, das Angebot einer Sitemap für eine klare Navigation und als expliziter Pfad für Suchmaschinen-Crawler (Erlhofer,

2019), die Verwendung von Viewports zur Sicherstellung einer responsiven Darstellung der Inhalte sowie die Ladegeschwindigkeit berücksichtigt. Noch eindeutiger sind allerdings explizite Anweisungen für Suchmaschinencrawler in der robots.txt sowie die Definition von Nofollow- und Canonical-Links, da diese auch als direkte Anweisungen für Suchmaschinen und Crawler definiert werden (ebd.).

2.5.3 Wahrscheinlich nicht optimiert

Die Seite ist wahrscheinlich nicht optimiert, wenn sie nicht höchstwahrscheinlich nicht optimiert und nicht höchstwahrscheinlich optimiert ist und wenn mindesten eines der folgenden Kriterien erfüllt ist: (1) kein Description Tag, (2) kein Title Tag oder, wenn identische Title Tags auf Unterseiten sind, (3) keine Open Graph Tags definiert.

Wie bereits bei den anderen Regeln gezeigt, sind viele Merkmale als Hinweise für SEO zu deuten, auch wenn diese zum Teil als gute Praktiken für die Erstellung von Websites gelten können. Daher wird bei der Einschätzung, ob ein Dokument nicht optimiert ist, auf eine Prüfung der Mindestanforderungen für Suchmaschinenoptimierung zurückgegriffen, d.h., wenn diese nicht erfüllt sind, kann davon ausgegangen werden, dass wahrscheinlich keine Optimierungsmaßnahmen unternommen wurden. Zu diesen Mindestanforderungen zählt, dass eine Seitenbeschreibung in Form eines Description Tags gesetzt wurde. Dies gilt ebenfalls für den Seitentitel, der als Title Tag definiert wird. Bei dem Seitentitel wird allerdings ebenfalls geprüft, ob sich die Seitentitel auf Unterseiten einer Webseite unterscheiden, denn die Nutzung eines beispielsweise automatisch generierten identischen Titels zeigt, dass keine Anstrengungen unternommen wurden, diese entsprechend den Seiteninhalten anzupassen. Da Description und Title Tags auch in den sogenannten Open Graph Tags definiert werden können, wird bei dieser Regel auch geprüft, ob solche Tags gefunden wurden. Open Graph Tags werden von Social-Media-Angeboten wie Facebook und Twitter unterstützt und dienen einer strukturierten Vorschaudarstellung der Seiteninhalte in diesen sozialen Medien (Krrabaj et al., 2017).

2.5.4 Höchstwahrscheinlich nicht optimiert

Die Domain der Seite ist auf der Liste mit nicht optimierten URLs.

Bei dieser Regel wird bisher nur geprüft, ob das Suchergebnis ein Wikipedia-Artikel ist. Wenn unsere weiteren Auswertungen eindeutigere Hinweise auf nicht optimierte Dokumente liefern, wird diese Liste um die Domains dieser Dokumente erweitert.

2.6 Technische Umsetzung der Methoden in einem Software-Tool

Für die Anwendung der Ansätze und Methoden wurde ein Software-Tool in Python programmiert, das sich aus verschiedenen Programmbibliotheken zusammensetzt, die für die Datenerfassung und Datenanalyse gut geeignet sind.

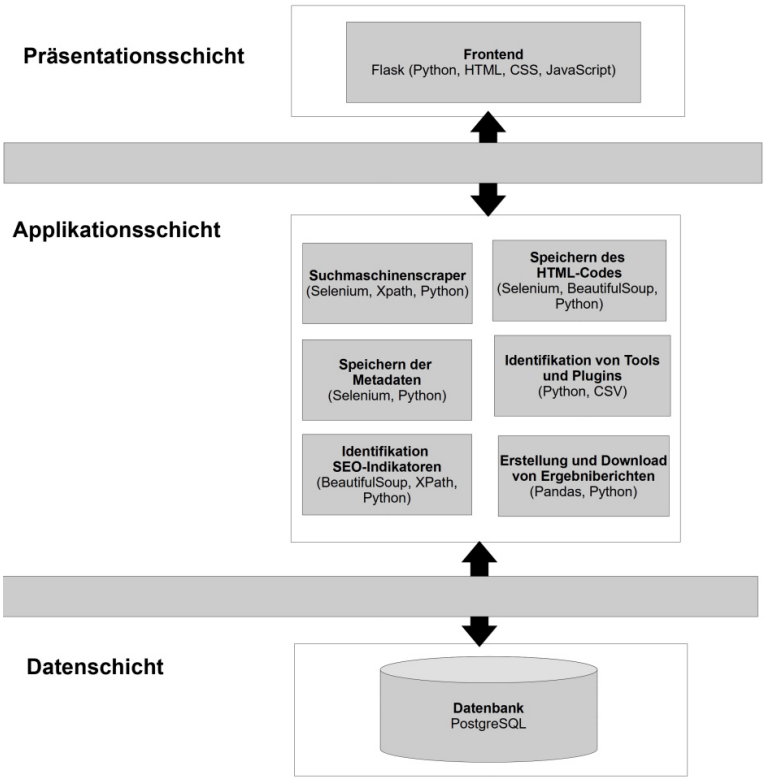

Abb. 2 MVC-Architektur der Software

Sämtliche Ergebnisse des Tools werden in einer PostgreSQL-Datenbank[3] gespeichert. Für eine Strukturierung der Software wurde auf die MVC-Architektur zurückgegriffen, die durch die Aufteilung der Präsentations-, Applikations- und Datenschicht eine bessere Wartbarkeit der Software ermöglicht. Die Datenbank ist dabei als zentraler Speicher auf einem Datenbankserver installiert, währen die Applikationen verteilt arbeiten. Damit ist es möglich, sämtliche Module für die Speicherung und Analyse verteilt auf mehreren Servern zu installieren, um damit die Performance zu erhöhen. Abbildung 2 zeigt die Software-Architektur des Tools mit den Schichten und den jeweiligen Modulen. Im Folgenden werden die Module näher erläutert.

2.6.1 Präsentationsschicht

Das Frontend der Software basiert auf dem Web-Framework Flask[4], da dieses sämtliche Funktionen anbietet, die für die Webentwicklung notwendig sind. So lassen sich mit geringem Aufwand Formulare erstellen und über HTML-Templates abbilden. Das Frontend in der Anwendung dient dazu, neue Datensätze zu generieren und die Datenanalysen zu steuern. Der Nutzende kann in dem Frontend Namen für Studien anlegen und dazu Suchanfragen definieren. Sämtliche weitere Prozesse werden dann anschließend automatisiert auf der Applikationsschicht durchgeführt.

2.6.2 Applikationsschicht

Auf dieser Schicht werden sämtliche Prozesse durchgeführt, die zu Beginn dieses Abschnitts näher erläutert wurden und in Abbildung 1 dargestellt sind. Eine weitere Hauptanwendung in dieser Schicht ist der Suchmaschinenscraper, der auf Selenium[5] (portables Framework zum Testen von Webanwendungen) und Beautiful Soup[6] (Paket zum Parsen von HTML und XML) basiert. Mit dem Scraper werden vorab definierte Suchanfragen automatisiert an Google gesendet und alle Suchergebnisse von allen Suchergebnisseiten (von der ersten Seite bis zur letzten Seite) gespeichert. Die URLs der Such-

3 https://www.postgresql.org/

4 https://flask.palletsprojects.com/

5 https://www.selenium.dev/

6 https://www.crummy.com/software/BeautifulSoup/

ergebnisse werden mithilfe von XPath extrahiert und anschließend werden der Quelltext, die Trefferposition, die Domain, die robots.txt und die Ladegeschwindigkeit jedes einzelnen Suchtreffers erfasst und in Relation mit der Suchanfrage in der Datenbank gespeichert. Alternativ zum Suchmaschinenscraper lassen sich auch Listen mit URLs für eine Analyse importieren. Die URLs werden über eine eindeutige ID in der Datenbank gespeichert.

Die Software prüft automatisch, ob neue URLs in der Datenbank erfasst wurden, und falls noch keine Auswertung für die SEO-Indikatoren zu der URL durchgeführt wurde, werden diese ausgeführt. Dabei werden in dem Modul für die Analyse der SEO-Indikatoren die technischen Merkmale über verschiedene Prüfalgorithmen aus dem Quelltext extrahiert und gespeichert. Für die Erfassung der SEO-Plug-ins und Analytics Tools werden die jeweiligen Listen mit den Suchmustern für jedes Plug-in geladen und abgeglichen. Für die Kategorisierung des Dokuments wird die Hauptdomain mit den hinterlegen URL-Listen verglichen und, falls ein Treffer vorliegt, in der Datenbank gespeichert. Sind alle Informationen zu einer URL gespeichert, wird die Wahrscheinlichkeit von SEO in diesem Dokument über den regelbasierten Klassifikator ermittelt und ebenfalls gespeichert. Mit dem Berichtsmodul können zu jeder Zeit Ergebnisberichte zu den gespeicherten Datensätzen erzeugt und als CSV-Datei heruntergeladen werden. Dabei werden sowohl die Rohdaten zur Verfügung gestellt als auch Zusammenfassungen zu den jeweiligen Datensätzen mit deskriptiven Kennzahlen zur SEO-Wahrscheinlichkeit.

Alle Module in dieser Schicht kommunizieren sowohl mit der Datenschicht, die bei dieser Anwendung aus einer großen PostgreSQL-Datenbank besteht, als auch mit der Applikationsschicht, auf der die Anwendung gestartet und gesteuert wird.

2.6.3 Datenschicht

Die Datenschicht der Anwendung ist eine PostgreSQL-Datenbank, die insgesamt aus sieben Tabellen besteht. Dort werden die Studien, die Suchanfragen, eine Jobverwaltung für die Scraper, die Suchergebnisse, die Quelltexte und die SEO-Merkmale gespeichert.

2.6.4 Demo-Tool

Ein Teil der Software ist ein Demo-Tool, das genutzt werden kann, um eine Analyse gezielt für eine gewünschte URL durchzuführen. Dadurch lassen sich schnelle Tests für den regelbasierten Klassifikator durchführen, um die Ansätze weiterzuentwickeln. Das Tool kann direkt im Web aufgerufen und für eigene Tests genutzt werden. Abbildung 3 zeigt einen Ergebnisbericht, der direkt in der Webanwendung generiert wird.

Abb. 3 Demo-Tool

In dem Bericht werden die Ergebnisse aus allen Analysen in dazugehörigen Kategorien ausgewiesen. So wird am Anfang die Einschätzung zur SEO-Wahrscheinlichkeit angezeigt. Darauf folgt die Auswertung zu den SEO-Plug-ins und Analytics Tools, die Kategorisierung der URL und darauffolgend werden die technischen SEO-Indikatoren erläutert. Der Bericht kann ebenfalls als CSV-Datei für weitere Verarbeitungen heruntergeladen werden.

3 Studien zur Überprüfung von Suchmaschinenoptimierung in kommerziellen Suchergebnissen

Für die praktische Anwendung unseres Ansatzes haben wir bisher drei Studien zu verschiedenen Themenbereichen durchgeführt, um zum einen unsere Vorgehensweise zu überprüfen, aber auch um erste Ergebnisse zu der Anzahl von Dokumenten in Suchergebnissen zu erhalten, die durch Suchmaschinenoptimierung beeinflusst sind. Insgesamt wurden 263.790 Suchergebnisse aus drei Datensätzen analysiert. Bei der Erhebung wurden dafür jeweils alle Suchtreffer berücksichtigt, die von Google zu einer Suchanfrage zurückgegeben wurden. Dabei sind dies in der Regel nicht mehr als 300 Ergebnisse je Anfrage. Die Datensätze werden im Folgenden näher erläutert und anschließend die Ergebnisse der Studien präsentiert.

3.1 Datensätze

Für die Anwendung unserer Methoden wurden drei sehr verschiedene Datensätze definiert, die sich thematisch stark voneinander unterscheiden. So wurden beispielsweise die populärsten Suchanfragen bei Google durch die Ausweisung in Google Trends über einen definierten Zeitraum gesammelt, um eine vielfältige Sammlung von Suchergebnissen zu generieren, die thematisch nicht festgelegt sind. Die weiteren Datensätze sind dagegen themenspezifisch und weisen eine stärkere Heterogenität auf. So konnten durch die Zusammenarbeit mit der Medienanstalt Hamburg Schleswig-Holstein Suchanfragen generiert werden, die potenziell strafrechtlich relevante Dokumente in Bezug auf rechtsradikale Inhalte auffinden. Der letzte Datensatz wurde

mithilfe von deutschen Suchanfragen generiert, die sich auf die Corona-Pandemie beziehen.

3.1.1 Google Trends

Bei diesem Datensatz wurden 244.985 Suchergebnisse durch 1.478 Suchanfragen erfasst. Die Suchanfragen wurden dabei in dem Zeitraum von März bis Juni 2020 direkt von der Google-Trends-Seite[7] entnommen.

Dabei wurde deutlich, dass sich die Suchanfragen in der Regel immer wieder auf ähnlich populäre Themen über alle Monate hinweg beziehen. Dazu zählen durch die Corona-Pandemie Suchanfragen zu Corona, zu Prominenten, zu Fernsehsendungen und Anfragen zu Sport, z. B. zur Fußball-Bundesliga. Dieser Datensatz wurde erstellt, um ein möglichst großes und vielfältiges Set von Suchergebnissen zu generieren.

3.1.2 Rechtsradikale Inhalte

Dieser Datensatz ist durch eine Kooperation mit den Landesmedienanstalten Hamburg Schleswig Holstein entstanden. Eine Aufgabe einer Medienanstalt ist die Prüfung strafrechtlich relevanter Inhalte im Web. Daher werden dort spezialisierte und ständige angepasste Suchanfragen genutzt, um solche Inhalte zu finden. Durch den Zugriff auf diese Anfragen wurden insgesamt 82 Suchanfragen im März 2020 automatisiert an Google geschickt und insgesamt 13.403 Suchergebnisse mit potenziell rechtsradikalen Inhalten gespeichert und durch unsere Software analysiert. Die Auswahl dieses Themas für einen Datensatz folgte dabei der Frage, ob Betreiber solcher Angebote Suchmaschinenoptimierung betreiben. Die Annahme war, dass zu solchen Themen häufiger Privatangebote zu finden sind, die weniger professionalisiert im Online-Marketing agieren und daher auch kein SEO einsetzen.

3.1.3 Corona

Microsoft Bing stellt Suchanfragen zur Verfügung, die als Suchintention zur Corona-Pandemie passen.[8] Dafür haben wir 483 Top-Anfragen für Deutschland aus dem Datensatz von Microsoft Bing im September 2020 extrahiert

7 https://trends.google.de/trends/?geo=DE

8 https://github.com/microsoft/BingCoronavirusQuerySet

und 5.402 Suchergebnisse gespeichert. Dieser Datensatz wurde generiert, da die Corona-Pandemie ein hochaktuelles Thema ist und wir auf tatsächlich genutzten Suchanfragen von Nutzer*innen zugreifen konnten.

3.2 Auswertung der Datensätze

Im Folgenden werden die Ergebnisse der Analyse der SEO-Wahrscheinlichkeit in den Dokumenten aus den Datensätzen näher erläutert und anschließend diskutiert.

3.2.1 Ergebnisse

Eine Analyse der Datensätze zeigt, dass die überwiegende Mehrheit der Ergebnisse entweder höchstwahrscheinlich oder zumindest wahrscheinlich optimiert ist. In Abbildung 3 sind die jeweiligen Ergebnisse zu sehen. Je nach Datensatz können wir sehen, dass zwischen 41,9% und 62% der gefundenen Ergebnisse als höchstwahrscheinlich optimiert eingestuft werden.

Tab. 2: Verteilung der Domains (Top-10)

Domain	Anzahl der Dokumente mit dieser Domain in allen Datensätzen	Anteil der Dokumente in allen Datensätzen in %
books.google.de	5.724	12,13
de.wikipedia.org	2.442	5,18
youtube.com	2.122	4,50
t-online.de	2.075	4,40
focus.de	1.915	4,06
welt.de	1.814	3,85
spiegel.de	1.713	3,63
sueddeutsche.de	1.699	3,60
stern.de	1.622	3,44
rtl.de	1.535	3,25

Die Unterschiede zwischen den Datensätzen lassen sich auf den höheren Anteil von Nachrichteninhalten im Trends- und Corona-Datensatz im Vergleich zum rechtsradikalen Datensatz zurückführen. So waren im Google-Trends-Datensatz 56% der Ergebnisse Nachrichtenangebote (136.806 Dokumente), im Corona-Datensatz 45% (2.454 Dokumente) und im Datensatz mit

den potenziell rechtsradikalen Inhalten nur 34%. Insgesamt konnten 47.177 eindeutige Domains über alle Datensätzen hinweg bei 263.790 Dokumenten festgestellt werden. Eine Auswertung der Top 10 zeigt, dass insbesondere Angebote von Google sehr häufig in den Dokumenten zu finden sind (books.google.de und youtube.com). Wikipedia ist auf dem zweiten Platz. Alle weiteren Angebote sind Nachrichtenangebote.

Eine weitere Analyse in Bezug auf genutzte SEO Plug-ins und Analytics Tools ergibt eine ähnliche Verteilung in allen Datensätzen. Tabelle 3 zeigt die Verteilung.

Tab. 3: Verteilung der SEO-Plug-ins und Analytics Tools

Datensatz	Anzahl an Dokumenten mit SEO-Plug-ins	Anzahl an Dokumenten mit Analytics Tools
Google Trends (244.985 Dokumente)	8% (19.515 Dokumente)	36% (87.175 Dokumente)
Potenziell rechtsradikale Inhalte (13.403 Dokumente)	8% (1.206 Dokumente)	36% (4.765 Dokumente)
Corona (5.402 Dokumente)	6% (332 Dokumente)	33% (1.793 Dokumente)

Es ist deutlich, dass die Anzahl an SEO-Plug-ins über alle Datensätze hinweg nicht sehr stark ausgeprägt ist. Das ist darauf zurückzuführen, dass diese in der Regel bei Webangeboten eingesetzt werden, die auf dem CMS Wordpress basieren und dies in der Regel nicht bei professionellen Nachrichtenangeboten verwendet wird. Eine Auswertung der Analytics Tools zeigt allerdings, dass die meisten Angebote Webanalysen für ihre Angebote durchführen, was auf kommerzielle Absichten hindeutet.

Die Auswertung der technischen Merkmale zeigt, dass 96% aller Dokumente HTTPS einsetzt (252.519 Dokumente) und Open Graph Tags sogar bei 98% (257.396 Dokumente) gesetzt werden. Das zeigt beispielsweise, wie hoch die Relevanz von Social Media für die Webseitenbetreiber ist, da fast alle Anbieter diese speziellen Tags für eine optimierte Darstellung einer Vorschau ihrer Inhalte definieren.

Abbildung 4 zeigt die Auswertung der SEO-Wahrscheinlichkeit in den Datensätzen. Dabei zeigt sich, dass nur ein kleiner Teil der Ergebnisse als höchstwahrscheinlich nicht optimiert eingestuft wurde (1,6% über alle Datensätze). Dies sind alles Ergebnisse von Wikipedia, da dies die einzige

Website auf unserer Liste der definitiv nicht optimierten Seiten ist. Ein relativ kleiner Teil der Seiten (zwischen 3,5 und 7%) ist wahrscheinlich nicht optimiert. Für eine bessere Darstellung wurden die Ergebnisse aus höchstwahrscheinlich nicht optimiert und wahrscheinlich nicht optimiert zusammengefasst. Wir fanden auch eine leichte Überlappung von bis zu 3% über alle Datensätze hinweg. Die Überlappung lässt sich dadurch erklären, dass die Ergebnisse von unserem Klassifikator sowohl als wahrscheinlich optimiert als auch als wahrscheinlich nicht optimiert eingestuft wurden. Zusammenfassend haben wir festgestellt, dass ein großer Teil der in Google gefundenen Ergebnisse entweder definitiv oder wahrscheinlich optimiert ist. Über 90% der gefundenen Ergebnisse gehören zu diesen Kategorien.

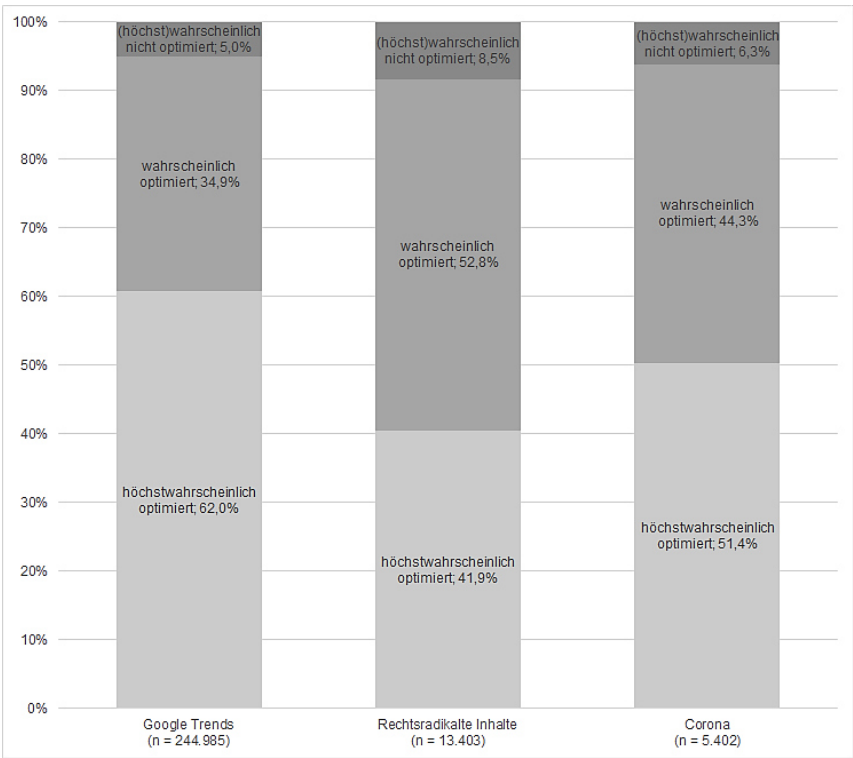

Abb. 4 Verteilung der SEO-Wahrscheinlichkeit

3.2.2 Diskussion

Die ersten Studien zur Ermittlung der Wahrscheinlichkeit der Suchmaschinenoptmierung zeigen, dass davon auszugehen ist, dass optimierte Inhalte die Suchergebnisse stark dominieren. So sind nur bis zu 10% nicht oder wahrscheinlich nicht optimiert. Dies ist keine Überraschung, da wir wissen, dass SEO ein Multimilliarden-Dollar-Geschäft ist und Unternehmen und andere Akteure oft von der Sichtbarkeit abhängig sind, die ihre Websites durch Traffic von Websuchmaschinen erhalten. Damit kann die Annahme bestätigt werden, dass Suchmaschinenoptimierung einen starken Einfluss auf die Suchergebnisse hat – auch unabhängig von den Themenbereichen, in denen recherchiert wird.

So unterscheiden sich die Datensätze inhaltlich stark voneinander. Während in dem Ergebnis-Set zu den Google Trends ein breites Themenspektrum nach der Popularität von Suchanfragen abgedeckt wird, sind die Ergebnisse zu den potenziell rechtsradikalen Inhalten und zu dem Coronavirus sehr spezifisch und weniger kommerziell ausgerichtet. Allerdings lässt sich eine Aussage darüber, ob der Einfluss der Suchmaschinenoptimierung sich positiv oder negativ auf die Ergebnisqualität auswirkt, auch in Bezug auf einzelne Suchanfragen oder die Gesamtheit der Suchanfragen mit den durchgeführten Analysen nicht beantworten. Für solche Analysen müsste die Methodik beispielsweise durch Retrievalstudien, in denen Juroren die Qualität der Suchergebnisse bewerten, ergänzt werden.

Ferner ist es auch notwendig, den bisherigen regelbasierten Ansatz weiter auszudifferenzieren, da bisher nur eine beschränkte Menge von Merkmalen in die Analyse einfließt. Dabei sind insbesondere die Prüfung auf den Einsatz von SEO-Plug-ins, was ein sehr starker Indikator für Suchmaschinenoptimierung ist, und die Überprüfung auf fehlende Trefferbeschreibungen (Description-Tag), was als starker Indikator für fehlende Suchmaschinenoptimierung ist, besonders aussagekräftig Indikatoren. Die weiteren Merkmale, die wir bisher betrachten, sind zum Teil ebenfalls gut für die Einschätzung von SEO geeignet, doch fehlt es an einer klaren Gewichtung – auch, um die Aussagekraft besser einschätzen zu können. Weitere Merkmale, die sich beispielsweise auf die Keywords in den Suchanfragen beziehen, oder auch sogenannte Offsite-Faktoren wie die Anzahl von Verlinkungen auf eine Webseite, fließen bisher nicht in das Modell ein.

In Bezug auf die Kategorisierung der Webseiten nach eher optimierten und nicht optimierten Sites kann die Identifikation eines Seitenbetreibers als

Kunde einer SEO-Agentur als sicherer Indikator angenommen werden. Zusätzlich ist eine hohe Suchmaschinenoptimierung bei einem Nachrichtenangebot anzunehmen. Dies zeigt auch die nähere Auswertung der Häufigkeiten in Bezug auf die Domains in den Datensätzen. Es ist davon auszugehen, dass besonders häufige aufgetretene Domains auch optimiert sind, und da insbesondere Nachrichtenangebote in der Quellenverteilung auftauchen, kann angenommen werden, dass diese Art von Websites auch SEO-Maßnahmen nutzt.

In der Auswertung der Verteilung der häufigsten Quellen in den Datensätzen waren in sieben von zehn Fällen prominente Nachrichtenanbieter mit einem Anteil von ca. 3% an der Gesamtheit aller Suchergebnisse zu finden. Dies lässt sich zum einen auf die überwiegend informationsorientierten Suchanfragen für alle Datensätze zurückführen, zum anderen erklären wir uns diese Dominanz von Nachrichtendiensten aber auch damit, dass die Suchmaschinen verlässliche Quellen in den Suchergebnissen bevorzugen, um beispielsweise die Gefahr von Falschinformationen und Fake News zu minimieren (Wingfield et al., 2016). Durch diese Vorgehensweise können allerdings kleinere und von großen Medienunternehmen unabhängige Informationsangebote benachteiligt werden. Eine weitere Benachteiligung für Inhaltsanbieter besteht auch dadurch, dass Eigenangebote von Suchmaschinentreibern sehr prominent in den Suchergebnissen aufgeführt werden. So zeigt unsere Analyse bei der Verteilung der Domains, dass eigene Angebote von Google (Google Books und YouTube) stark vertreten sind. Diese Art der Benachteiligung führt bereits seit mehreren Jahren zu rechtlichen Auseinandersetzungen (Edelman, 2015).

Die Auswertung aller genannten Aspekte zeigt, wie komplex der Einfluss verschiedener Faktoren auf die Anzeige von Dokumenten in den Suchergebnissen ist. So zeigen unsere Auswertungen bereits, dass neben dem Einfluss der Suchmaschinenoptimierung auf die Ergebnisse auch Aspekte wie die Diversität von Suchergebnissen und die Eigenangebote der Suchmaschinenbetreiber relevant sind.

4 Fazit und weiteres Vorgehen

Der Einfluss von Suchmaschinenoptimierung auf die Suchergebnisse in kommerziellen Suchmaschinen ist bisher kaum erforscht. Mit unserem halb-

automatisierten Verfahren und einer Umsetzung in einem Software-Tool konnten wir erste Auswertungen dazu vornehmen, wie groß der Anteil von optimierten Inhalten in Suchergebnissen unabhängig von Themenschwerpunkten ist. Die Vorgehensweise enthält dabei bewährte Ansätze, um Suchergebnisse automatisiert zu speichern, relevante Merkmale zu extrahieren und eine Kategorisierung der URLs vorzunehmen, um den Seitentyp nach Kategorien, aufgeteilt nach wahrscheinlich und nicht wahrscheinlich optimierten Seiten, aufzuteilen. Für diese Analysen haben wir Merkmale identifiziert, die verlässliche Aussagen über die Wahrscheinlichkeit von SEO auf Webseiten ermöglichen. Allerdings ist dieses Modell noch nicht vollständig und wird durch weitere Merkmale, Faktorenanalysen, Gewichtungen von Merkmalen und Merkmalsgruppen sowie Analysen mit Machine-Learning-Methoden aus dem unüberwachten Lernen (z. B. durch Clustering) erweitert. Dabei wird beispielsweise auch mit semi-überwachten Methoden aus der automatischen Klassifikation gearbeitet und mit Regressionsanalysen soll ein konkreter Wert für die SEO-Wahrscheinlichkeit berechnet werden. Ferner wird auch die Trefferpositionen im Ranking der Suchmaschinen näher betrachtet, um die Zusammenhänge von SEO und den Positionen zu untersuchen.

Mit der Durchführung weiterer Studien werden wir unsere Methodik weiterentwickeln und den Effekt der Suchmaschinenoptimierung weiter untersuchen. Dabei werden wir auch noch stärker auf die Diversität von Quellen in den Suchergebnissen und auf die Verteilung von Eigenangeboten der Suchmaschinenbetreiber eingehen, da sich bereits jetzt gezeigt hat, dass diese Faktoren ebenfalls einen Einfluss auf die Dokumentenauswahl für die Suchergebnisse und auf die Darstellung der Suchergebnisse haben.

Forschungsdaten

Der Quelltext des Tools kann über die OSF-Plattform (dx.doi.org/10.17605/OSF.IO/ETZHD) abgerufen werden. Die Suchergebnisse für die Auswertung sind ebenfalls auf der OSF-Plattform unter dx.doi.org/10.17605/OSF.IO/RVX54 abrufbar.

Förderung

Das Projekt „SEO-Effekt“, aus dem dieses Tool hervorgeht, wird von der Deutschen Forschungsgemeinschaft (DFG) unter der Projektnummer 417552432 gefördert.

Literaturverzeichnis

Edelman, B. (2015): Does Google Leverage Market Power Through Tying and Bundling? *Competition Law & Economics* 11 (2), 365–400. https://doi.org/10.1093/joclec/nhv016

Enge, E.; Spencer, S.; Stricchiola, J. (2015): *The Art of SEO: Mastering Search Engine Optimization*. Sebastopol, CA: O'Reilly.

Erlhofer, S. (2019): *Suchmaschinen-Optimierung: Das umfassende Handbuch*. Bonn: Rheinwerk Verlag.

Goel, S.; Broder, A.; Gabrilovich, E.; Pang, B.(2010): Anatomy of the long tail. In: Davison, B. D.; Suel, T.; Craswell, N.; Liu, B. (Hrsg.): *Proceedings of the Third ACM International Conference on Web Search and Data Mining - WSDM '10*. New York, NY: ACM Press.

Griesbaum, J. (2013): Online-Marketing. In: R. Kuhlen, W. Semar, D. Strauch (Hrsg.): *Grundlagen der praktischen Information und Dokumentation*. Berlin, Boston, MA: de Gruyter Saur. https://doi.org/10.1515/9783110258264.411

Krrabaj, S.; Baxhaku, F.; Sadrijaj, D. (2017): Investigating search engine optimization techniques for effective ranking. A case study of an educational site. In: *6th Mediterranean Conference on Embedded Computing (MECO), Bar, 2017*. IEEE. DOI: https://doi.org/10.1109/meco.2017.7977137

Lewandowski, D. (2017): Users' Understanding of Search Engine Advertisements. *Journal of Information Science Theory and Practice* 5 (4), 6–25. https://doi.org/10.1633/JISTaP.2017.5.4.1

Lewandowski, D.; Kerkmann, F.; Rümmele, S.; Sünkler, S. (2018): An empirical investigation on search engine ad disclosure. *Journal of the Association for Information Science and Technology* 69 (3), 420–437. DOI: https://doi.org/10.1002/asi.23963

Lewandowski, D.; Sünkler, S (2013): Representative online study to evaluate the revised commitments proposed by Google on 21 October 2013 as part of EU competition investigation AT.39740-Google: Country comparison report. http://searchstudies.org/wp-content/uploads/2015/10/Google_Country_Comparison_Report.pdf

Li, K.; Lin, M.; Lin, Z.; Xing, B. (2014): Running and Chasing – The Competition between Paid Search Marketing and Search Engine Optimization. In: *47th Hawaii International Conference on System Sciences, Waikoloa, HI*. IEEE, S. 3110–3119. https://doi.org/10.1109/HICSS.2014.640

McCue, T. (2018): SEO Industry Approaching $80 Billion But All You Want Is More Web Traffic. *Forbes*. https://www.forbes.com/sites/tjmccue/2018/07/30/seo-industry-approaching-80-billion-but-all-you-want-is-more-web-traffic/

Petrescu, P. (2014): Google Organic Click-Through Rates in 2014. https://moz.com/blog/google-organic-click-through-rates-in-2014

Ronallo, J. (2012): HTML5 Microdata and Schema.org. *Code4Lib Journal* (16). https://journal.code4lib.org/articles/6400

Schultheiß, S.; Lewandowski, D. (2020): "Outside the industry, nobody knows what we do". SEO as seen by search engine optimizers and content providers. *Journal of Documentation* 77 (2), 542–557. https://doi.org/10.1108/JD-07-2020-0127

StatCounter (2020a): StatCounter: Search Engine Market Share United States Of America | StatCounter Global Stats.

StatCounter (2020b): StatCounter: Search Engine Market Share Europe | StatCounter Global Stats.

Van Couvering, E. (2009). *Search Engine Bias. The Structuration of Traffic on the World-Wide Web.* Diss., London School of Economics and Political Science. http://etheses.lse.ac.uk/41/

Wang, X.; Bendersky, M.; Metzler, D.; Najork, M. (2016): Learning to rank with selection bias in personal search. In: *SIGIR '16: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval.* New York, NY: ACM, S. 115–124. https://doi.org/10.1145/2911451.2911537

Wingfiedl, N.; Isaas, M.; Benner, K. (2016): Google and Facebook take aim at fake news sites. *The New York Times*. https://www.nytimes.com/2016/11/15/technology/google-will-ban-websites-that-host-fake-news-from-using-its-ad-service.html

Zamani, H.; Bendersky, M.; Wang, X.; Zhang, M. (2017): Situational context for ranking in personal search. In: *WWW '17: Proceedings of the 26th International Conference on World Wide Web*. Geneva: International World Wide Web Conferences Steering Committee, S. 1531–1540. https://doi.org/10.1145/3038912.3052648

In: T. Schmidt, C. Wolff (Eds.): Information between Data and Knowledge. Information Science and its Neighbors from Data Science to Digital Humanities. Proceedings of the 16th International Symposium of Information Science (ISI 2021), Regensburg, Germany, 8th–10th March 2021. Glückstadt: Verlag Werner Hülsbusch, pp. 273–298. DOI: doi.org/10.5283/epub.44948.