HOCHSCHULE OSNABRÜCK
UNIVERSITY OF APPLIED SCIENCES

Fakultät Management, Kultur und Technik

Masterprogramm

Management und Technik (M. Sc.)

Hausarbeit

im Modul Performance Measurement und Big Data

Thema: Analyse der Berichterstattung von Tageszeitungen zum Thema Nachhaltigkeit

Am Beispiel: *die tageszeitung* und *Focus (Online)*

Dozent*In:	Prof. Dr. A. Schierenbeck
	Prof. Dr. R. Buschermöhle
Semester:	WS 2020/21
vorgelegt von:	Frank Knese, 801221
	Mario Gierke, 804405
	Niklas Fischer, 806114
	Massimiliano Kuck, 804823
	Alexander Lutterbeck, 827714
Abgabedatum:	Lingen, den 22.01.2021

Kurzfassung

Die Nutzung und Möglichkeiten informationstechnischer Systeme gewinnen zunehmend an Bedeutung. Die Verarbeitung großer Datenmengen zu entscheidungsrelevanten Informationen ist ein zentraler Anwendungsbereich dieser Systeme. In der vorliegenden Arbeit werden durch Data Mining und unter Anwendung von Big Data Techniken die Onlinezeitungen die tageszeitung und der Focus (Online) analysiert. Die Analyse bezieht sich auf das Thema Nachhaltigkeit und die 17 Nachhaltigkeitsziele der Vereinten Nationen. Die zugrundeliegende Forschungsfrage der Arbeit lautet: Welche Präsenz hat das Thema Nachhaltigkeit in der Berichterstattung von taz und Focus und wie ausgeglichen ist diese bezüglich der einzelnen Nachhaltigkeitsziele? Die Untersuchungsergebnisse zeigen, dass die Berichterstattung zu den einzelnen Nachhaltigkeitszielen unausgeglichen ist. Artikel zum Thema Nachhaltigkeit sind im Focus durchschnittlich länger als andere Artikel. Beide Zeitungen berichten über Nachhaltigkeit neutral und objektiv; der Focus hat dabei einen Hang zu Subjektivität. Das redaktionelle Konzept und die politische und inhaltliche Ausrichtung beider Zeitungen werden durch die Analysen deutlich.

Abstract

The utilization and possibilities of information technology systems are becoming increasingly important. The processing of large amounts of data into decision-relevant information is an essential area of these systems. In this paper, the online newspapers die tageszeitung and Focus (Online) are analysed through data mining and big data methods. The analysis relates to the topic of sustainability and the 17 sustainability goals of the United Nations. The fundamental research question of this paper is: What presence does the topic of sustainability have in the reporting of taz and Focus and how balanced is this presence regarding the individual sustainability goals? The results of the study show that the reporting on the individual sustainability goals is unbalanced. Sustainability articles in Focus are longer on average than other articles. Both newspapers report on sustainability neutrally and objectively; Focus has a tendency towards subjectivity. The editorial concept as well as the political and content orientation of both newspapers become clear through the analyses.

Inhalt

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

BICE	Basel Insitut of Commons and Economics
BoW	Bag-of-Words
ID	Identifikationsnummer
KI	Künstliche Intelligenz
ML	Machine Learning
NLP	Natural Language Processing
NLTK	Natural Language Tool Kit
taz	die tageszeitung

Symbolverzeichnis

n	Stichprobengröße
$\overline{x}$	arithmetisches Mittel in der Stichprobe
μ	Erwartungswert in der Grundgesamtheit
s	Standardabweichung in der Stichprobe
σ	Standardabweichung in der Grundgesamtheit
i	Zählindizes der Urliste
j	Zählindizes der Merkmalsausprägungen
p	Anteil an der Grundgesamtheit
k	Zeilenanzahl
m	Spaltenanzahl
h	absolute Häufigkeit
a	Ausprägungen des Merkmals X
b	Ausprägungen des Merkmals Y
χ^2	Chi-Quadrat
K	Kontingenzkoeffizient
K^*	korrigierter Kontingenzkoeffizient
R^2	Bestimmtheitsmaß
α	Achsenabschnitt (lineare Regression)
β	Steigungsparameter (lineare Regression)
ϵ	Fehler (lineare Regression)

1 Einleitung

Die Nutzung und Möglichkeiten informationstechnischer Systeme gewinnen übergreifend an Bedeutung. Führende Unternehmen handeln am Weltmarkt nicht mehr mit Gütern wie Öl oder Mineralien, sondern mit Daten. Daten sind der Rohstoff des 21. Jahrhunderts. Der Begriff *Web Scraping* beschreibt das automatisierte Auslesen und Speichern von Inhalten einer Webseite, um so Daten zu gewinnen. Mit verschiedenen Methoden und Techniken können die gewonnenen und meist sehr großen Datenmengen ausgewertet und analysiert werden. *Big Data* verfolgt den Ansatz diese große und oftmals unstrukturierte Datenmenge zu verarbeiten, um so sinnvolle und entscheidungsrelevante Erkenntnisse zu gewinnen. *Big Data* nutzt neue und effiziente Analysemethoden und kann bisher unbekannte Zusammenhänge sichtbar machen, die einer enormen Datenmenge zugrunde liegen. Das *Data Mining* deckt durch die systematische Anwendung von statistischen und mathematischen Tools Muster und Strukturen in den Datenbeständen auf. Als Teilgebiet des *Data Minings* extrahiert das *Text Mining* Wissen aus Texten, um die Informationen der Texte zu verarbeiten und zu analysieren.

Eine mögliche Datengrundlage für das *Text Mining* bieten Tageszeitungen. Diese können inhaltlich sowie in Bezug auf die Metadaten der Artikel analysiert und ausgewertet werden. Speziell das Onlineangebot der Tageszeitungen macht das *Text Mining* sowie das *Data Mining* besonders interessant. Mittels *Web Scraping* können die veröffentlichten Artikel einer Zeitung ausgelesen und gespeichert werden, um sie im Nachgang zu analysieren. Durch die große Text- und Datenvielzahl können hier *Big Data Methoden* angewendet werden, um Strukturen und Zusammenhänge zu erkennen.

Im Rahmen der vorliegenden Arbeit werden die beschriebenen Methoden eingesetzt, um die Berichterstattung einzelner Zeitungen im Themenkomplex *Nachhaltigkeit* zu analysieren. Jeden Freitag machen Schüler[1] und Studenten durch die *Fridays for Future Bewegung* darauf aufmerksam, dass Umweltschutz und nachhaltiges Handeln heute darüber entscheiden, wie die Erde morgen aussieht.

[1] Auf Grund der besseren Lesbarkeit wird in dieser Arbeit ausschließlich die männliche Form verwendet. Sie bezieht sich auf Personen aller Geschlechter.

Doch was bedeutet Nachhaltigkeit? Dazu haben die *Vereinte Nationen* mit der *Agenda 2030* Nachhaltigkeit in Form von 17 konkreten Ziele beschrieben. Diese Ziele haben eine weltweite Gültigkeit und fördern die nachhaltige Entwicklung auf verschiedenen Ebenen.

Die Kombination aus *Web Scraping, Text Mining* und die gesellschaftliche Relevanz der Nachhaltigkeit eröffnet vielseitige Möglichkeiten zur Analyse. Das Ziel der vorliegenden Arbeit ist die Analyse von Zeitungsartikeln. Dazu werden zwei Zeitungen repräsentativ verglichen. Der Themenkomplex Nachhaltigkeit steht dabei im Vordergrund. Durch die Programmierung, Anwendung und Nutzung verschiedener Systeme sowie die statistische Auswertung der Daten soll folgende Forschungsfrage im Zuge der Arbeit beantwortet werden:

> *Welche Präsenz hat das Thema Nachhaltigkeit in der Berichterstattung von taz und Focus und wie ausgeglichen ist diese bezüglich der einzelnen Nachhaltigkeitsziele?*

Mit der Beantwortung dieser Forschungsfrage sollen mögliche Strukturen, Zusammenhänge und Verbindungen innerhalb der Berichterstattung aufgedeckt werden. Daraus können Rückschlüsse über die Zeitungen selbst aber auch über die Präsenz der Nachhaltigkeit innerhalb des Mediums Zeitung geschlossen werden.

2 Theoretische Grundlagen

Im folgenden Absatz werden die in der Arbeit verwendeten theoretischen Grundlagen beschrieben. Dazu gehört zum einen die inhaltliche Einordnung der Tageszeitungen und der Nachhaltigkeit, zum anderen die technischen Grundlagen des Web Scrapings und der Textanalyse durch das Natural Language Processing.

2.1 Einordnung der Tageszeitungen

Die in dieser Arbeit analysierten Zeitungen werden im Folgenden hinsichtlich verschiedener Kriterien verglichen. Eine generelle Einordnung des redaktionellen Konzepts wird vorgenommen. Die wichtigsten Zahlen in Bezug auf Printmedien und das Online-Angebot werden genannt, um die Unterschiede darzustellen.

2.1.1 Die tageszeitung

Die tageszeitung (Eigenschreibweise, taz) ist eine unabhängige und überregionale deutsche Tageszeitung. Seit dem 17. April 1979 erscheint werktags eine gedruckte Auflage. Im dritten Quartal 2020 lag die Anzahl der verkauften Auflagen bei 41.170 Stück. Die taz hat eine Reichweite von 0,32 Mio. Lesern pro gedruckte Ausgabe. Die Webseite der taz (taz.de) verzeichnet ca. 2 Mio. Unique User[2] pro Monat. Derzeit beschäftigt die taz ca. 250 Mitarbeiter. Die taz erscheint mit Lokalteilen für Berlin und Nord (Hamburg und Bremen). Als eine der ersten Tageszeitungen war die taz bereits 1995 im Internet verfügbar. Seitdem sind die Artikel kostenfrei zugänglich. Leser können jedoch einen freiwilligen Beitrag zur Unterstützung zahlen. [1], [2]

Das redaktionelle Konzept der taz wird als grün-links, linksalternativ sowie system- und konsumkritisch beschrieben. Laut eigener Aussage engagiert sie sich für eine kritische Öffentlichkeit, tritt für die Verteidigung und Entwicklung der Menschenrechte ein und gibt denjenigen eine Stimme, die in der Gesellschaft selten ein Gehör finden. Sowohl inländische als auch ausländische Themen haben bei der Berichterstattung den gleichen Rang. Die Redaktion der taz weist jegliche Einflussnahme auf die Berichterstattung durch einzelne Personen, Gruppen, Parteien, Unternehmen oder Ähnlichem zurück. [2], [3]

[2] Dt. unterschiedliche Nutzer, Kennzahl zur Ermittlung der Reichweite eines Online-Angebots. Es werden nur die unterschiedlichen Nutzer innerhalb eines Zeitraumes gezählt, unabhängig davon, wie oft die Seite pro Monat aufgerufen wird.

Im Zuge des *Web Scrapings* werden die Artikel aller Hauptressorts und Ressorts (s. Anhang A) der taz gespeichert. Für die inhaltliche Analyse der taz sind besonders die Artikel der Hauptressorts „Politik“, „Öko“, „Gesellschaft“ und „Kultur“ relevant. In Bezug auf das Thema der Arbeit wird dem Hauptressort „Öko“ eine besondere inhaltliche Bedeutung zugesprochen.

2.1.2 Focus (Online)

Das Nachrichtenmagazin Focus ist eines der reichweitestärksten Magazine Deutschlands; es gehört zum *Hubert Burda Media* Medienkonzern. Seit 1993 erscheint wöchentlich am Samstag eine Ausgabe; seit 2000 wird ebenfalls wöchentlich das Wirtschaftsmagazin *Focus Money* herausgegeben. Die Anzahl der verkaufen Auflagen lag im dritten Quartal 2020 bei knapp 262.000 Stück mit einer Reichweite von ca. 3,5 Mio. Lesern pro gedruckte Ausgabe. Die Webseite des Focus Magazins (Focus Online) verzeichnete im Jahr 2020 ca. 29 Mio. Unique User pro Monat und ist damit eine der am höchsten frequentierten Nachrichten-Webseite Deutschlands. [4], [5]

Das redaktionelle Konzept des Focus bedient das Konzept eines eher konservativen Nachrichtenmagazins. Neben der politischen Berichterstattung beinhaltet das Konzept gesellschaftliche Themen wie Familie und Gesundheit, aber auch Finanzen und Karriere. Der Focus kann zum Teil als Ratgeber verstanden werden. Durch kurze Texte, eine Vielzahl an Illustrationen und vielen Grafiken sind Informationen für den Leser einfach zu verarbeiten. [4], [6], [7]

2.2 Nachhaltigkeit

> *Der Begriff Nachhaltigkeit beschreibt in seinem ursprünglichen Sinn die Nutzung eines regenerierbaren natürlichen Systems in einer Weise, dass dieses System in seinen wesentlichen Eigenschaften erhalten bleibt und sein Bestand auf natürliche Weise nachwachsen kann.* [8]

Der rasante Anstieg der Weltbevölkerung in den letzten Jahrzehnten, die Industrie oder die Zunahme des Auto-, Luftfahrt-, Schiff- oder Zugverkehrs haben unter anderem einen hohen Ressourcenverbrauch, hohe CO_2-Emissionen und den anthropogenen Treibhauseffekt zur Folge. Dies stellt folglich eine nicht nachhaltige Entwicklung dar. [8] Um die Auswirkungen einzudämmen und der negativen Entwicklung entgegenzuwirken, wurde in den 1990er-Jahren das Drei-Säulen-Modell (s. Abbildung 1) der Nachhaltigkeit entwickelt.

Gemäß des Drei-Säulen-Modells setzt sich der Begriff Nachhaltigkeit aus den drei gleichgewichteten Komponenten Wirtschaft, Umwelt und Soziales zusammen. Jede dieser Stützen ist für eine zukunftsfähige Entwicklung gleichermaßen zu berücksichtigen und dient der Sicherstellung und Verbesserung der ökonomischen, ökologischen und sozialen Leistungsfähigkeit. [8]

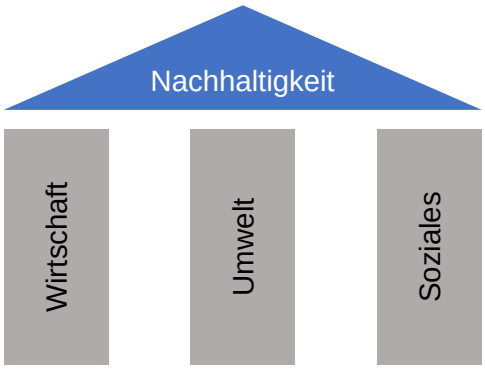


Abbildung 1: Drei-Säulen-Modell der Nachhaltigkeit [8]

Hinsichtlich der ökologischen Nachhaltigkeit gilt es im Wesentlichen die Regenerationsfähigkeit des Ökosystems zu beachten. Der Abbau von natürlichen Ressourcen darf die Regenerationsrate nicht übersteigen. Die ökonomische Nachhaltigkeit betrifft die effiziente und gerechte Verteilung von Gütern und Dienstleistungen. Damit die Wirtschaft langfristig bestehen kann, sind ein funktionsfähiges Preissystem, ein Wettbewerb am Markt, Privateigentum sowie ein stabiler Geldwert anzustreben. Die soziale Nachhaltigkeit umfasst die Sicherstellung der Existenz und Lebensqualität für weitere Generationen. Um dies zu erreichen, gilt es die Menschenwürde zu achten, die Grundbedürfnisse zu erfüllen, Gerechtigkeitsempfinden sicherzustellen und Chancengleichheit herzustellen. [9]

Nach dem entwickelten Drei-Säulen-Modell und weiteren Klimaabkommen in den Folgejahren, verhandelten die Vereinten Nationen die *Agenda 2030* mit Gültigkeit ab 1. Januar 2016 und mit einer Laufzeit von 15 Jahren. Diese hat Gültigkeit für alle Staaten der Welt und dient der nachhaltigen Entwicklung auf ökologischer, ökonomischer und sozialer Ebene unter Festlegung von 17 konkreten Nachhaltigkeitszielen. Zu den Kernaspekten der *Agenda 2030* zählt die Umgestaltung von Volkswirtschaften zur nachhaltigen Entwicklung durch eine emissionsarme Lebensweise, der Erzeugung sauberer Energie, der verantwortungsvollen Produktion und des Konsumverhaltens sowie der Klimapolitik. Zudem werden die

fünf Kernbotschaften – Mensch, Planet, Wohlstand, Frieden, Partnerschaft – benannt, die als handlungsleitende Prinzipien den Nachhaltigkeitszielen vorangestellt sind. [8] Die 17 Nachhaltigkeitsziele der Vereinten Nationen sind in Abbildung 2 visualisiert und lauten wie folgt:

1. Armut in jeder Form und überall beenden;
2. Den Hunger beenden, Ernährungssicherheit und eine bessere Ernährung erreichen und eine nachhaltige Landwirtschaft fördern;
3. Ein gesundes Leben für alle Menschen jeden Alters gewährleisten und ihr Wohlergehen fördern;
4. Inklusive, gerechte und hochwertige Bildung gewährleisten und Möglichkeiten des lebenslangen Lernens für alle fördern;
5. Geschlechtergerechtigkeit und Selbstbestimmung für alle Frauen und Mädchen erreichen;
6. Verfügbarkeit und nachhaltige Bewirtschaftung von Wasser und Sanitärversorgung für alle gewährleisten;
7. Zugang zu bezahlbarer, verlässlicher, nachhaltiger und zeitgemäßer Energie für alle sichern;
8. Dauerhaftes, inklusives und nachhaltiges Wirtschaftswachstum, produktive Vollbeschäftigung und menschenwürdige Arbeit für alle fördern;
9. Eine belastbare Infrastruktur aufbauen, um inklusive und nachhaltige Industrialisierung zu fördern und Innovationen zu unterstützen;
10. Ungleichheit innerhalb von und zwischen Staaten verringern;
11. Städte und Siedlungen inklusiv, sicher, widerstandsfähig und nachhaltig machen;
12. Für nachhaltige Konsum- und Produktionsmuster sorgen;
13. Umgehend Maßnahmen zur Bekämpfung des Klimawandels und seiner Auswirkungen ergreifen;
14. Ozeane, Meere und Meeresressourcen im Sinne einer nachhaltigen Entwicklung erhalten und nachhaltig nutzen;
15. Landökosysteme schützen, wiederherstellen und ihre nachhaltige Nutzung fördern, Wälder nachhaltig bewirtschaften, Wüstenbildung bekämpfen, Bodenverschlechterung stoppen und umkehren sowie den Biodiversitätsverlust stoppen;

16. Friedliche und inklusive Gesellschaften im Sinne einer nachhaltigen Entwicklung fördern, allen Menschen Zugang zur Justiz ermöglichen und effektive, rechenschaftspflichtige und inklusive Institutionen auf allen Ebenen aufbauen;
17. Umsetzungsmittel stärken und die globale Partnerschaft für nachhaltige Entwicklung wiederbeleben. [10]

Abbildung 2: 17 Nachhaltigkeitsziele der Vereinten Nationen [11]

2.3 Web Scraping

Unter Web Scraping wird der Prozess des Extrahierens von Daten aus Webseiten verstanden. Die extrahierten Daten werden gespeichert, um anschließend analysiert oder anderweitig verwertet zu werden. Der Prozess des Web Scrapings kann manuell oder automatisiert erfolgen. Aufgrund des hohen Aufwands wird dies häufig automatisiert durchgeführt. Grundsätzlich dient das Web Scraping dem Extrahieren von explizit gewünschten Informationen von Webseiten. Folglich werden für den Zweck unerwünschte Informationen herausgefiltert. [12]

Der sogenannte Scraper ruft zunächst die Webseite auf. Anschließend erfolgt die automatisierte und strukturierte Extraktion der Daten. Darauffolgend werden die extrahierten Daten beispielsweise in einer lokalen Datenbank gespeichert. Der ganze Prozess kann mit Hilfe der multiparadigmatischen Programmiersprache Python umgesetzt werden. [12] *BeautifulSoup* ist eine Python-Bibliothek für das

Auslesen von Daten aus HTML- und XML-Dateien. Als Eingabe wird ein HTML-Dokument an den BeautifulSoup-Konstruktor übergeben. Der Konstruktor transformiert ein komplexes HTML-Dokument in ein BeautifulSoup-Objekt, welches als eine Struktur aus einzelnen Python-Objekten verstanden werden kann. Auf dieses Objekt sind verschiedene Methoden anwendbar. Beispielsweise kann durch die Struktur navigiert, darin gesucht oder Teile davon modifiziert werden. Wichtige Methoden für das vorliegende Projekt sind dabei explizit die *find()* und *find_all()* Methode. Diese Methoden ermöglichen das Auffinden und Extrahieren von gewünschten Informationen aus dem persistierten HTML-Quelltext. Es kann beispielweise nach Strings, Tags oder IDs im Dokument gesucht werden. [13]

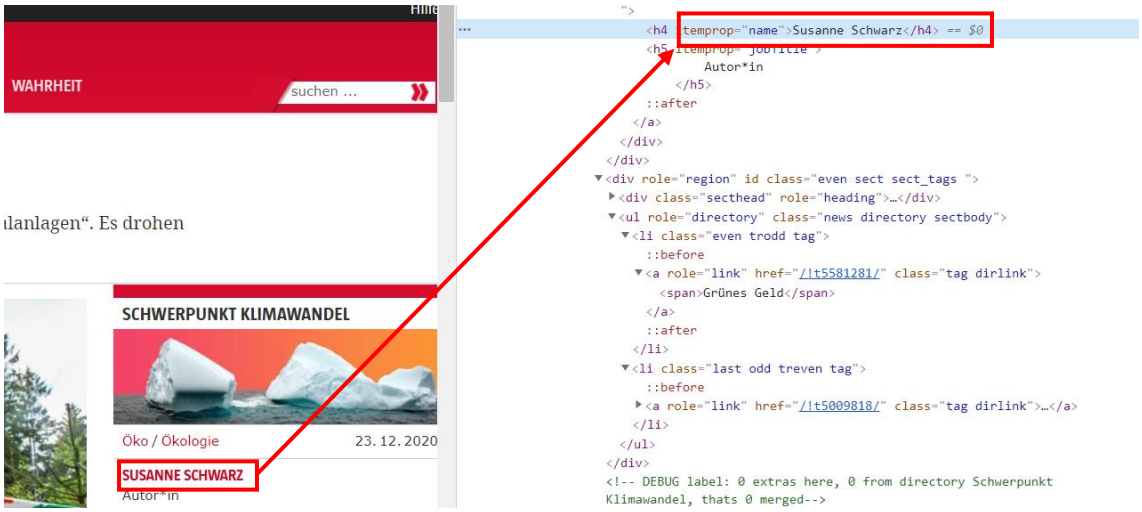


Abbildung 3: Webseite und Quellcode – Extraktion des Autors / taz

In Abbildung 3 ist beispielhaft ein Auszug aus einem Artikel der taz und der dazugehörige HTML-Quelltext abgebildet. Der Name der Autorin des Artikels ist durch die rote Markierung sowohl auf der Webseite als auch in der Quelltextzeile hervorgehoben. Das beispielhafte Merkmal „Autor“ kann wie folgt extrahiert werden:

```
def getAuthor(rawHtml):
    try:
        soup = BeautifulSoup(rawHtml, 'html.parser')
        author = (soup.find('h4').get_text())
        return author
    except:
                print("no author extractable")
```

Abbildung 4: Beispielhafte Extraktion des Merkmals "Autor"

Bei dieser Methode wird zunächst der HTML-Quelltext übergeben und in ein BeautifulSoup-Objekt transformiert. Daraufhin wird nach dem Element „h4“ gesucht und mit der Methode *get.text()* der lesbare Text als Unicode-String ausgegeben. [12] Nach diesem Muster werden alle Merkmale für das Forschungsprojekt aus den gescrapten HTML-Quelltexten erhoben.

2.4 Natural Language Processing

Natural Language Processing (kurz: NLP, dt. Verarbeitung natürlicher Sprache) bezeichnet die maschinelle Verarbeitung menschlicher Sprache. Dazu gehört sowohl das Verstehen (Natural Language Understanding) als auch das Erzeugen (Natural Language Generation) von natürlicher Sprache. Dabei können Bücher, Mails, Kurznachrichten, informelle Gespräche, Reden oder Telefonate verarbeitet werden. NLP ist eine Unterkategorie der sog. künstlichen Intelligenz (KI) und ein Anwendungsfall für das *Deep Learning*. Über bestimmte Algorithmen kann die KI die Absicht hinter Wörtern verstehen und so beispielsweise einen Chat-Bot[3] dazu befähigen, die gewünschten Aktionen durchzuführen und in der Sprache des jeweiligen Nutzers zu antworten. Ursprünglich diente NLP zur Sicherstellung der Lesefähigkeit von Computern. Heute umfasst es grundlegende Aspekte der Linguistik. Dabei ist die Komplexität der Sprache problematisch. Sowohl Satzbau als auch der Inhalt variieren hinsichtlich der Komplexität. Je nach Kontext und Situation haben Sätze eine grundlegende andere Bedeutung, was das Verstehen anspruchsvoll gestaltet. Die Einsatzgebiete für NLP sind vielseitig. Die bereits oben erwähnten Chat-Bots sind nur ein Anwendungsgebiet. NLP wird ebenfalls in Suchmaschinen, Question-Answering-Systemen, Sprachassistenten, Übersetzungen, Textklassifikationen und vielem mehr verwendet. [14], [15]

Ausgangspunkt des NLP sind die Rohdaten – der Rohtext. Dieser kann aus diversen Quellen, z. B. Spracherkennung, Mail oder Web Scraping, stammen. Dabei existieren für jede Quelle eigene Techniken und Bibliotheken, die zur Extraktion des Rohtextes dienen. [16]

Bei dem NLP wird zwischen der syntaktischen und der semantischen Methode unterschieden. Die syntaktische Methode fokussiert sich auf die Strukturen des

[3] Ein textbasiertes Dialogsystem, welches das schriftliche Kommunizieren mit technischen Systemen ermöglicht.

Satzes und den Funktionen der einzelnen Satzelemente. Die semantische Methode beschäftigt sich hingegen mit der Bedeutung einzelner Elemente. Welche Methode des NLP gewählt wird, hängt von der Aufgabe, der vorliegenden Datenqualität, der Vollständigkeit und der Zielsetzung der Sprachverarbeitung ab. Die gewählte Methode ist während der Anwendung kontinuierlich zu überprüfen und ggf. zu optimieren. Die natürliche Sprache wird durch das NLP in eine Zahlenfolge umgewandelt, welche durch den Rechner verarbeitet werden kann. [16]

Abbildung 5: Sechs Schritte des NLP nach [17], [18]

Abbildung 5 zeigt, dass jede NLP-Anwendung in mehreren aufeinanderfolgenden Schritten erfolgt, die als Kreislauf verstanden werden können. Je nach Anwendungsfall sind diese unterschiedlich, jedoch ist die Grundform aller NLP-Anwendungen ähnlich. Im Folgenden werden mögliche Schritte des NLP dargestellt: [17], [18]

1. Datenbereinigung

Bei der Datenbereinigung werden die rohen Text- oder Sprachdaten entnommen und so transformiert, dass nur noch die reinen Textdaten ohne Format zur Verfügung stehen. Das nachfolgende Beispiel zeigt wie Textdaten in Form von HTML-Quellcode um sogenannte HTML-Tags bereinigt werden. [17]

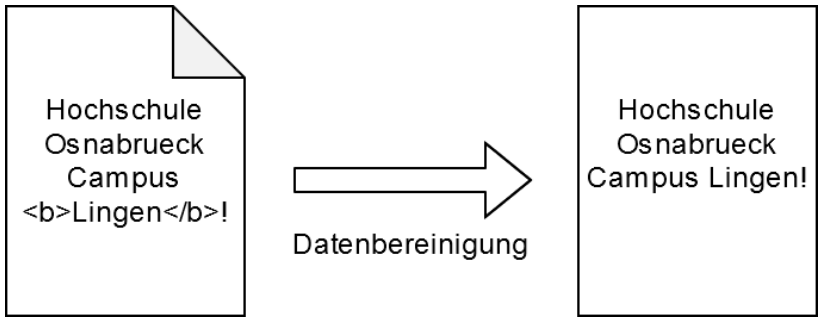


Abbildung 6: NLP Datenbereinigung

2. Tokenisierung und Normalisierung (Tokenizing and Normalizing)

Nach der Datenbereinigung liegen die Daten in reiner Textform ohne Format vor. Allerdings beinhaltet der Text Sprachelemente (z. B. Groß- und Kleinschreibung) sowie Satzzeichen. Daher werden in diesem Schritt die Sprachelemente und Satzzeichen entfernt, sowie der Text auf seine Wortbestandteile (sog. Tokens) reduziert. [17]

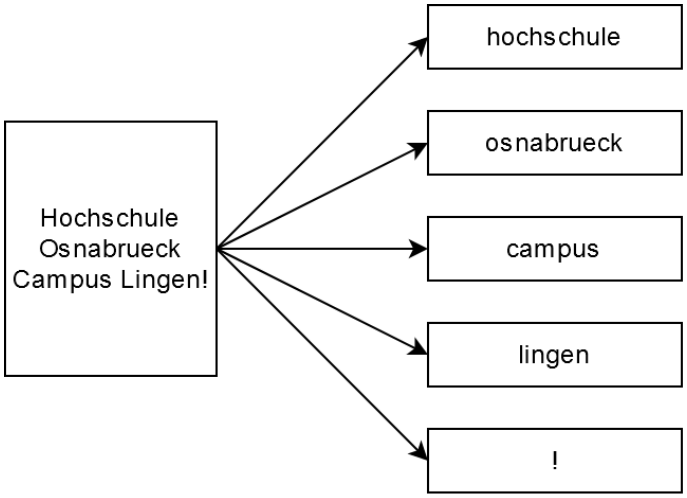


Abbildung 7: NLP Tokenisierung und Normalisierung

3. Stoppwort-Entfernung (Stop Word Removal)

Im Anschluss an die Tokenisierung und Normalisierung werden Stoppwörter entfernt. Dies sind Wörter, welche keine Relevanz für den Dokumenteninhalt besitzen. Übliche Stoppwörter sind bestimmte Artikel („der“, „die“ und „das“), unbestimmte Artikel („ein“, „eine“ und „einer“), Konjunktionen („und“, „oder“ und „weil“), häufig verwendete Präpositionen („an“, „von“ und „in“) sowie die Negation („nicht“). Die Entfernung von Stoppwörtern kann jedoch den jeweiligen Kontext einer Aussage verändern. Daher ist dieser Schritt optional und die Stoppwörter sollten in Bezug auf den Kontext gewählt werden. [15]

4. Wortartzuweisung (Pats of Speech)

Die Wortartzuweisung identifiziert und markiert alle übrigen Wörter mit der entsprechenden Wortart. Dabei werden Tags, wie z. B. „NNP“ (noun, proper, singular) oder „CC“ (conjuction, coordination), genutzt. Dieser Schritt wird für den darauffolgenden Schritt zum Bilden von Wortstämmen benötigt. [17]

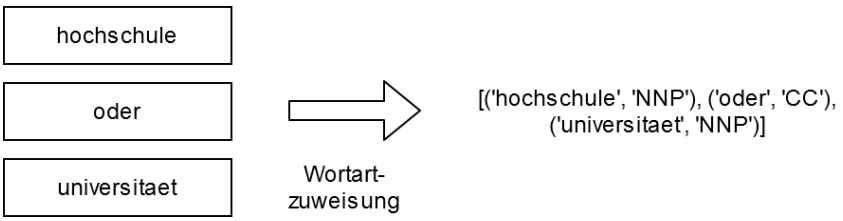


Abbildung 8: NLP Wortartzuweisung

5. Wortstammbildung (Stemming and Lemmatization)

Stemming und Lemmatization sind Textnormalisierungstechniken für natürliche Sprachen. In der Grammatik werden Wörter häufig so verändert, dass sie verschiedene grammatikalische Kategorien, wie Zeitform, Kasus, Person, Numerus, Geschlecht oder Stimmung ausdrücken können. Hierzu werden die Wörter mit einem Präfix, Suffix, Infix oder einem Vokalwechsel modifiziert. Durch Stemming und Lemmatization werden die einzelnen Wörter in ihre Grundform gebracht und auf den Wortstamm reduziert. Die Schritte Wortartzuweisung und Wortstammbildung reduzieren die Komplexität des Textes und erhöhen ihre Vergleichbarkeit. [19]

6. Worteinbettung

Wenn Wörter und Texte als Input für Algorithmen oder neuronale Netze verwendet werden sollen, müssen die Wörter durch Zahlen repräsentiert werden, damit sie von den verschiedenen Programmen verarbeitet werden können. Um diese Transformation von Wörtern in numerische Darstellungen umzusetzen, können die Methoden *Bag-of-Word* oder *Word2Vec* eingesetzt werden.

Bag-of-Word (BOW) dient zur Repräsentation von Wörtern durch Zahlen, damit diese als Input für Algorithmen oder neuronale Netze genutzt werden können. Bei der BoW-Methode wird nach der Bereinigung des Textes ein Wörterbuch aller im Text vorkommenden Wörter erstellt. Dabei wird eine eindeutige und aufsteigende ID an jedes Wort vergeben. Die erste ID ist die 0. Durch die vergebenen

ID und eine weitere Information über das Wort in Form einer Zahl wird ein zweidimensionaler Vektor des Wortes gebildet. Die zweite Dimension kann dabei eine der folgenden Informationen beinhalten:

- Vorkommen eines Wortes im Text (1 = enthalten, 0 = nicht enthalten)
- Absolute Häufigkeit eines Wortes
- Relative Häufigkeit eines Wortes

Bei der relativen Häufigkeit wird die zweite Dimension absteigend sortiert, wobei die niedrigste Zahl das Wort repräsentiert, welches am häufigsten in einem Text vorkommt. Das heißt, dass z. B. die Zahl 1 das Wort darstellt, welches am häufigsten vorkommt, die Zahl 2 dementsprechend das zweithäufigste Wort. [18]

Word2Vec ermöglicht die Repräsentation der Bedeutung der zuvor erstellten mathematischen Darstellung eines Wortes im Kontext zu anderen Worten. Es wird hierzu ein Algorithmus verwendet, welcher alle Input-Vektoren um die Information der umgebenden Wörter ergänzt und diese in einem multidimensionalen Raum anordnet. Wörter, welche häufig in einem ähnlichen Kontext genutzt werden, besitzen dann einen ähnlichen Vektor im Raum. Dieser Algorithmus verwendet das sogenannte *Unsupervised Learning*, welches ohne externe Vorgaben selbstständig eine Vektorrepräsentation für Wörter generiert. Ein Vektor kann je nach Trainingsmodell 100-300 Dimensionen haben und stellt zumindest ansatzweise die semantische Bedeutung des Wortes im Kontext dar. Der resultierende Vektor ist dabei abhängig von den Trainingsdaten des Trainingsmodells. Es können vordefinierte Datensätze wie beispielsweise das *GermanModel* oder eigene Texte verwendet werden. Die Vektoren ermöglichen arithmetische Operationen, wie das nachfolgende Beispiel zeigt. [18]

$$K\ddot{O}NIG - MANN + FRAU = K\ddot{O}NIGIN$$

So ergibt sich durch die Subtraktion des Vektors für das Wort *MANN* von dem Vektor des Wortes *KÖNIG* und die anschließende Addition des Vektors für *FRAU* der Vektor für das Wort *KÖNIGIN*. Dabei handelt es sich nicht um ein exaktes Ergebnis, aber der Ergebnisvektor dieser Operation ist dem Vektor des Wortes *KÖNIGIN* am ähnlichsten. Mit Hilfe von Word2Vec und den daraus resultierenden Vektoren können Inputdaten für aktuelle Anwendungen zur Analyse von Wörtern

im Kontext von Texten genutzt werden. Für viele Anwendungsfälle zur Textanalyse bietet Python-Bibliotheken. [18]

Die Tabelle 1 beinhaltet die im Projekt primär genutzten Python-Bibliotheken zur Verarbeitung der natürlichen Sprache. Deren Funktionsweise wird aufgrund der hohen Komplexität als Black-Box betrachtet.

Bibliothek	Beschreibung	Anwendungsfälle
Gensim	Dient zur Verarbeitung der natürlichen Sprache mit dem Schwerpunkt der Themenmodellierung. Gensim ermöglicht semantische Textanalysen, Inhaltsmodellierungen und einen semantischen Vergleich von Textkörpern. Darüber hinaus implementiert die Gensim-Bibliothek eine Implementierung der Word2Vec-Worteinbettung. [20], [21], [22]	• Zuordnung von IDs zu Wörtern • Transformation von Wörtern zu Vektoren • Vergleich anhand der numerischen Repräsentation (Vektoren), ob bestimmte Hitwords in Dokumenten enthalten sind • Bestimmen der ähnlichsten Wörter zu Nachhaltigkeitsbegriffen • Verwendung des *GermanModels*
Natural Language Toolkit (NLTK)	Das Natural Language Toolkit (NLTK) ist eine Plattform zur Erstellung von Python-Programmen, die zur Verarbeitung und Analyse der menschlichen Sprache dienen. Es bietet Schnittstellen zu verschiedenen lexikalischen Ressourcen. [23], [24]	• Entfernung von Stoppwörtern aus den Zeitungsartikeln

Bibliothek	Beschreibung	Anwendungsfälle
TextBlob	TextBlob ist eine Python-Bibliothek, welche das Natural Language Tool Kit (NLTK) verwendet. TextBlob ist eine einfache Bibliothek, die komplexe Analysen und Operationen auf textuellen Daten ermöglicht. [25]	• Stimmungsanalyse (Sentiment-Analyse) von Texten zur Entschlüsselung der Stimmung und Emotionen eines Zeitungsartikels • Analyse der Polarität, wobei das Ergebnis der Analyse einen Wert zwischen -1 und 1 einnimmt (eine Polarität von -1 definiert ein negatives Empfinden, 0 ein neutrales Empfinden und 1 hingegen ein positives Empfinden) • Analyse der Subjektivität, mit einem Ergebniswert zwischen 0 und 1 (eine Subjektivität von 0 impliziert, dass der Zeitungsartikel sachliche Informationen beinhaltet; -1 hingegen, dass der Zeitungsartikel eher persönliche Meinungen beinhaltet)

Tabelle 1: Verwendete Python-Bibliotheken

3 Projektmanagement

Ein Projekt ist ein ganzheitlich zielgerichtetes und einmaliges Vorhaben, welches sich durch abgestimmte Tätigkeiten und begrenzte Ressourcen kennzeichnet. Ressourcen sind im Kontext eines Projektes zeitlich, personell, finanziell und projektspezifisch-organisatorisch begrenzt. Weiterführend zeichnet sich ein Projekt durch die interdisziplinäre Bearbeitung und Durchführung im Team aus. [26]

Die Hausarbeit kennzeichnet sich durch die oben genannten Bedingungen als Projekt im Zuge des Masterstudiums. Als Kunde und Auftraggeber des Projektes fungieren die Hochschule Osnabrück in Form der Dozenten des Moduls *Performance Measurement und Big Data*. Die vorgegeben Projektziele machen das Projekt bewertbar, jedoch bringen diese auch immer einen Konflikt mit sich. Der Zielkonflikt des Projektmanagements besteht aus dem Ergebnis, der Zeit und den Ressourcen. Diese Faktoren beeinflussen die Erfüllung der Ziele, da eine vollständige Berücksichtigung nicht möglich ist. Dabei wird ein Gleichgewicht aller Faktoren angestrebt, um so ein qualitativ hochwertiges Ergebnis zu liefern, welches die Ziele unter Einhaltung der vorgeschriebenen Ressourcen erfüllt. [27] Um dies zu gewährleisten ist die Planung und Organisation des Projekts für das Team von großer Bedeutung.

„In der Projektpraxis existieren unterschiedliche Rollen und Organisationseinheiten, wie auch verschiedene Bezeichnungen.“ [27] Die Rollen werden vor der Bearbeitung des Projekts eindeutig verteilt und akzeptiert, damit während der Bearbeitung Unstimmigkeiten und Konfliktpotentiale vermieden werden.

3.1 Zeitmanagement

Die Bearbeitungszeit ist durch einen definierten Endtermin eine begrenzte Ressource des Projektes. Die termingerechte Erfüllung trägt außerdem maßgeblich zum Projekterfolg bei. Um Risiken von Beginn an zu vermeiden ist es wichtig, ein funktionierendes und transparentes Zeitmanagement zu betreiben. Die Meilensteine des Projektes werden zu Beginn festgelegt. Meilensteine sind wichtige und erfolgskritische Ereignisse im Projektmanagement, die den Verlauf des Projekts in überschaubare Abschnitte einteilen. Dadurch können auf dem Weg zur Zieler-

reichung die Zwischenziele überprüft werden. Im Zweifelsfall kann so die Richtung angepasst oder verändert werden. Hierdurch ist eine effiziente Überwachung gegeben. [27]

Eine genaue Übersicht der Termine und Aufgaben wird durch den Meilenstein-Balkenplan (s. Anhang G) visualisiert. Dazu werden die entwickelten Meilensteine in ein sog. Gantt-Diagramm überführt. Die X-Achse beschreibt den zeitlichen Fortschritt in Tagen. Auf der Y-Achse werden alle für das Projekt relevanten Ereignisse nach dem Fälligkeitsdatum sortiert. Durch das Tool des Meilenstein-Balkenplans ist die Visualisierung des Projektfortschritts nutzerfreundlich und transparent. Das Projektmanagement-Werkzeug zeigt übersichtlich den Bearbeitungsstatus des Projektes, welche Aufgaben tagesaktuell anstehen und zu welchem Zeitpunkt der nächste Meilenstein zu erfüllen ist. Es erleichtert dem Projektleiter außerdem die Definition von Arbeitspaketen und die zeitliche Einordnung dieser in den Verlauf des Projekts. [27]

3.2 Dokumenten- und Aufgabenmanagement

Um alle Informationen, Dokumente und Arbeitspakete konsistent zu speichern und jedem Mitglied des Projektteams den Zugriff zu ermöglichen, wird *Microsoft Teams* genutzt. Das Organisationstool dient der Aufgaben- und Dokumentenorganisation innerhalb des Projektteams. Hier erfolgt neben der Ablage aller Dokumente ebenfalls die direkte Definition und Zuweisung von Arbeitspaketen. Zur Versionsverwaltung des Software-Entwicklungsprojektes wird der netzbasierte Dienst *GitHub* verwendet.

4 Softwarearchitektur

Die Softwarearchitektur beschreibt die grundlegende Organisation der erstellten Software. Sie macht sichtbar welche Komponenten verwendet werden und welche Schnittstellen zwischen diesen bestehen. Dabei wird grundlegend zwischen den Funktionen (s. Abbildung 9) des *Scraping*, *Extracting* und *Analysing* unterschieden.

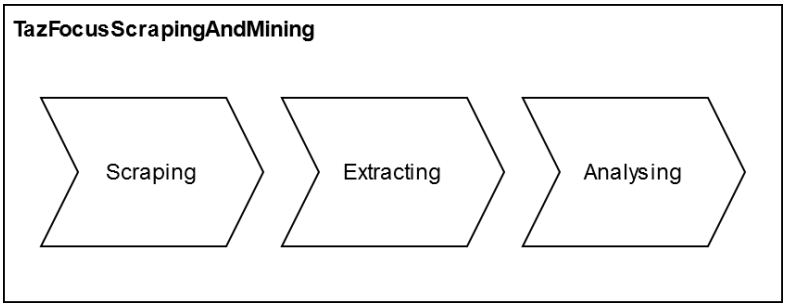


Abbildung 9: Taz / Focus Scraping und Mining

Die Funktionen des Scrapings werden durch die in der Abbildung 10 dargestellten Python-Skripte (Scraper.py, TazScraper.py, FocusScraper.py, DatabaseConnection.py) sowie der dokumentorientierten Datenbank *Scraper* (MongoDB) auf einem von der Hochschule zur Verfügung gestellten Server realisiert. Die MongoDB ermöglicht die Speicherung einer komplexen Datenhierachie sowie eine simple Abfrage von Daten mit Hilfe des *Mongo-Clients* der Python-Bibliothek pymongo.

Das Python-Skript *Scraper.py* speichert alle acht Stunden die rohen HTML-Quellcodes der Nachrichten-Webseiten Focus und taz. Dabei verwendet das Skript *Scraper.py* die jeweiligen Funktionen der Skripte *FocusScraper.py* und *TazScraper.py*. Diese beiden Skripte implementieren spezifische Funktionen zum Speichern des HTML-Codes der unterschiedlichen Nachrichtenseiten. Sie verwenden dazu einen *get-request* aus der HTTP-Bibliothek Request. Der durch den *Scraper.py* gespeicherte HTML-Quellcode wird in der Datenbank *Scraper* als JSON-Dateien persistiert. Das Python-Skript *DatabaseConnection.py* realisiert eine Datenzugriffsschicht, da es dem *Scraper.py* einen Zugriff auf die *Scraper*-Datenbank ermöglicht. Hierbei werden die rohen HTML-Quellcodes je nach Zeitung in den unterschiedlichen Collections *focusrawhtml* und *tazrawhtml* persistiert. Die folgenden Key-Value-Paare werden dazu pro Artikel gespeichert:

Key	Value
_id	enthält die einmalige URL des Artikels
rawhtml	enthält den HTML-Quellcode der URL

Tabelle 2: Key-Value-Paare der Entität der Collection tazrawhtml und focusrawhtml

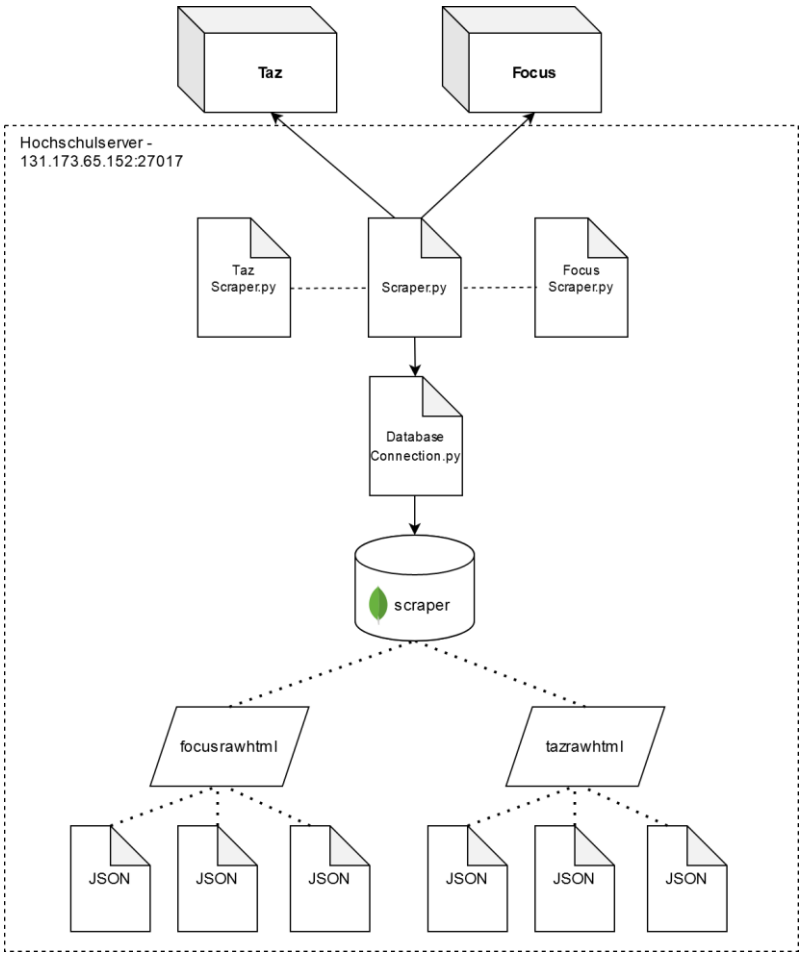


Abbildung 10: Softwarearchitektur des Projektes

Die Abbildung 11 zeigt eine beispielhafte Entität der Collection *tazrawhtml*.

```
{
  "_id": "https://taz.de/Militaerischer-Naturschutz-in-Kongo/
          !5727376/",
  "rawhtml": "<html lang="de" xmlns="http://www.w3.org/1999/xhtml"
             xmlns= [...]"
}
```

Abbildung 11: Entität der Collection *tazrawhtml*

Nachdem der Scraping-Prozess abgeschlossen ist, werden die Daten der Datenbank extrahiert. Die Abbildung 12 zeigt die Komponenten, welche die rohen HTML-Quellcodes extrahieren.

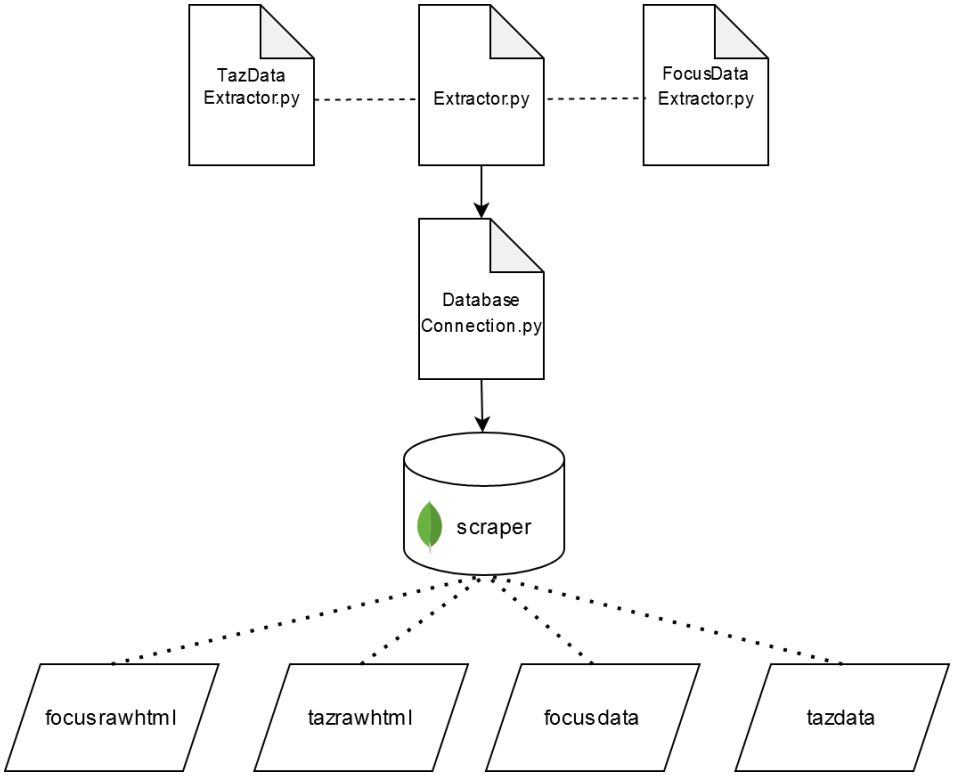


Abbildung 12: Komponenten zur HTML-Extraktion

Das Python-Skript *Extractor.py* greift auf spezifische Funktionalitäten der Skripte *TazDataExtractor.py* und *FocusDataExtractor.py* zurück. *DatabaseConnection.py* ermöglicht beim Extrahieren ebenfalls den Zugriff auf die persistierten Daten. Aus den rohen HTML-Quellcodes werden für jede Entität der Collections *focusrawhtml* und *tazrawhtml* die nachfolgenden Key-Value-Paare (s. Tabelle 3) extrahiert und in einer separaten JSON-Datei in den jeweiligen Collections *tazdata* und *focusdata* gespeichert.

Key	Value
_id:	einmalige URL des Artikels
date:	Datum des Artikels
author:	Autor des Artikels
title:	Überschrift und ggf. Unterüberschrift des Artikels
keywords:	Keywords des Artikels (nur bei der taz)
intro:	Kopf des Artikels mit den wichtigsten Informationen
body:	Hauptteil des Artikels mit den detaillierten Informationen
ressort:	Ressort in dem der Artikel publiziert wurde

Tabelle 3: Key-Value-Paare der Entität in der Collection tazdata und focusdata

Enthält der rohe HTML-Quellcode eines dieser Values nicht, so wird an dieser Stelle kein Wert hinterlegt. Die *DatabaseConnection.py* ermöglicht dem *Extractor.py* den Zugriff auf die *Scraper*-Datenbank. Die Abbildung 13 zeigt eine beispielhafte Entität der Collection *tazdata*.

```
{
    "_id": "https://taz.de/Konjunkturhilfen-fuer-Forstwirtschaft/
           !5731659/"
    "date": "27.11.2020",
    "author": "Ulrike Fokken",
    "title": "Konjunkturhilfen für Forstwirtschaft: Meister des
             Lobbyismus"
    "keywords": ["Waldschadensbericht", "Forstwirtschaft", "Ökolo
                gie"],
    "intro": "Die Forstwirtschaft hat es verstanden [...]"
    "body": "Konjunkturhilfen für Forstwirtschaft: Meister des
            Lobbyismus. Die Forstwirtschaft hat es verstanden, ihre
            [...]"
    "ressort": "Ökologie"
}
```

Abbildung 13: Entität der Collection *tazdata*

Die extrahierten Daten der Collections *tazdata* und *focusdata* bilden die Grundlage der Analyse, welche durch die in Abbildung 14 dargestellten Komponenten sichergestellt wird.

Bei der Analyse ermöglicht das Skript *DatabaseConnection.py* ebenfalls den Datenzugriff. Die Skripte *NLPEvaluation.py* und *SentimentalAnalysis.py* verarbeiten

für eine inhaltliche Analyse die Textkörper (body) der Collections *tazdata* und *focusdata*. Das Skript der *SentimentAnalyse.py* ermöglicht eine Bewertung hinsichtlich der Stimmung des Textkörpers (positiv oder negativ) und analysiert, ob dieser auf Meinungen oder Fakten basiert. *DataOperations.py* stellt dem *SentimentAnalyse.py* Skript grundlegende Funktionen zur Verfügung, um auf den Daten zu operieren. Das Skript *NLPEvaluation.py* dient zur Bestimmung der Häufigkeit und Ähnlichkeit von Wörtern in Textkörper (bodys). Hierzu greift es auf die Funktionalitäten des Skriptes *SpeachAnalysis.py* zurück, um die inhaltliche Analyse beispielsweise mit der Word2Vec Methode durchzuführen.

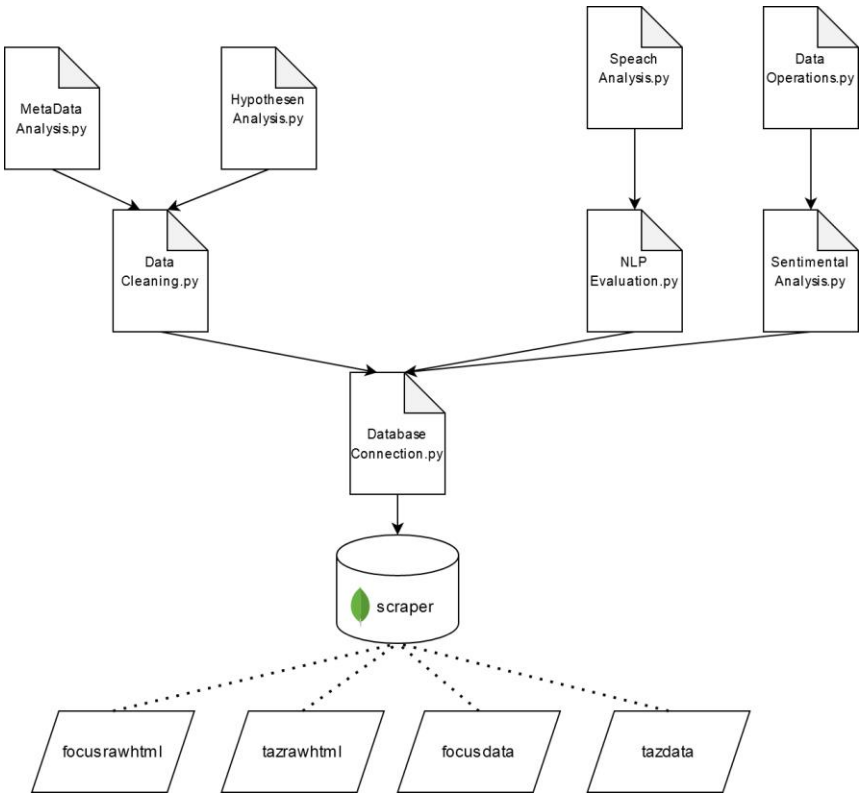


Abbildung 14: Komponenten zur Textanalyse

Die Skripte *MetaDataAnalysis.py* und *HypothesenAnalysis.py* verwenden das Skript *DataCleaning.py*, um eine Datenbereinigung der extrahierten Daten in der Collection *tazdata* vorzunehmen. *MetaDataAnalysis.py* dient zur Analyse von Metadaten der Artikel, die durch die taz veröffentlicht wurden. Beispiel hierfür sind: Artikelhäufigkeit pro Ressort oder Häufigkeit der Keywords. *HypothesenAnalysis.py* wertet die extrahierten Daten beider Zeitungen hinsichtlich der Hypothesen aus und visualisiert die Daten. Dabei verwendet es die Ergebnisse von *NLPEvaluation.py* und *SentimentAnalyse.py*, um mit den dort gewonnenen Erkenntnissen Hypothesen zu überprüfen.

5 Forschungsprozess

In Kapitel 5 wird der Forschungsprozess der vorliegenden Arbeit beschrieben. Im Kontext des Themas Nachhaltigkeit und Onlinezeitungen wird die Entstehung der Forschungsfrage erläutert. Basierend auf der Forschungsfrage werden Hypothesen entwickelt, die am Ende des Kapitels beantwortet werden. Der Forschungsprozess basiert auf einer quantitativen Datenanalyse, welche mittels der in Kapitel 4 vorgestellten Softwarearchitektur umgesetzt wird. Die für die Datenanalyse verwendetet statistischen Methoden werden ebenfalls in diesem Kapitel erläutert.

5.1 Forschungsfrage und Hypothesen

Nach der Verabschiedung der 17 Nachhaltigkeitsziele durch die Vereinten Nationen im Jahr 2016 wurden im Sommer 2019 erstmals Arbeitsberichte zum Stand der Umsetzung veröffentlicht. Diese Berichte beinhalten Informationen, Analysen, Daten sowie die Entwicklung über die vergangenen Jahre zu jedem einzelnen Nachhaltigkeitsziel. Veröffentlicht wurden die fünf Berichte durch die Vereinten Nationen, durch die Bertelsmann Stiftung und die EU-Statistikbehörde *Eurostat*. Die Berichte nahm das *Basel Institute of Commons and Economics* (BICE) als Anlass eine Analyse durchzuführen. Das Institut untersuchte die Anzahl der Nennungen der Ziele und Themen in den fünf Berichten und erstellte auf dieser Basis eine Rangliste unter Berücksichtigung des durchschnittlichen Ranglistenplatzes des jeweiligen Ziels bzw. Themengebiets (siehe Tabelle 4). Als Schlussfolgerung stellt das Institut eine ungleiche Berücksichtigung der Ziele und damit eine ungleiche Priorität bei der Umsetzung heraus. Demnach besitzen u. a. Themen wie Gesundheit, Energie, Klima, Wasser und Bildung eine weitaus höhere Priorität als saubere Energie, Landökosysteme, Ozeane, Meere und soziale Inklusion. In dem Zuge wird jedoch vom Institut ebenfalls kritisch angemerkt, dass sich die Ziele gegenseitig beeinflussen. So ist beispielsweise eine Bekämpfung von Armut ohne die Schaffung von Gleichberechtigung und sozialem Zusammenhalt nicht möglich. [28]

Nachhaltigkeitsziel	Rang	Durchschnittlicher Rang	Anzahl Erwähnungen
Gesundheit	1	3,2	1814
Energie, Klima, Wasser	2	4,0	1328, 1328, 1784
Bildung	3	4,6	1351
Armut	4	6,2	1095
Ernährung	5	7,6	693
Wirtschaftliches Wachstum	6	8,6	387
Technologie	7	8,8	855
Ungleichheit	8	9,2	296
Gleichstellung der Geschlechter	9	10,0	338
Hunger	10	10,6	670
Gerechtigkeit	11	10,8	328
Regierungsführung	12	11,6	232
Menschenwürdige Arbeit	13	12,2	277
Frieden	14	12,4	282
Saubere Energie	15	12,6	272
Landökosysteme	16	14,4	250
Ozeane, Meere, Meeresres-source	17	15,0	248
Soziale Inklusion	18	16,4	22

Tabelle 4: Umsetzung der Nachhaltigkeitsziele, in Anlehnung an [28]

Auf Basis der fünf Berichte zum Stand der Umsetzung der Nachhaltigkeitsziele sowie der Analyse des *Basel Institute of Commons and Economics* (BICE) ergibt sich für dieses Projekt die nachfolgende Forschungsfrage:

> *Welche Präsenz hat das Thema Nachhaltigkeit in der Berichterstattung von taz und Focus und wie ausgeglichen ist diese bezüglich der einzelnen Nachhaltigkeitsziele?*

Aufbauend auf dieser Forschungsfrage werden Hypothesen entwickelt, die im Laufe des Projekts durch Datenerhebung, -aufbereitung und -analyse überprüft und beantwortet werden.

Hypothese H1:
Die Berichterstattung zu den einzelnen Nachhaltigkeitszielen ist unausgeglichen.

Hypothese H2:
Die taz berichtet ausgeglichener zu den verschiedenen Nachhaltigkeitszielen als der Focus.

Hypothese H3:
Nachhaltigkeitsartikel sind bezüglich des Textumfangs länger.

Hypothese H4:
Die Berichterstattung zu Nachhaltigkeitsthemen ist eher negativ geprägt.

Hypothese H5:
Die Berichterstattung zu Nachhaltigkeitsthemen ist im Vergleich zur restlichen Berichterstattung eher subjektiv geprägt.

Bei den Hypothesen H1 und H2 handelt es sich um allgemein formulierte Hypothesen, die sich auf publizierte Artikel beider Tageszeitungen beziehen. Für die Hypothesen H3 bis H5 hingegen werden die Artikel beider Tageszeitungen miteinander verglichen. Somit ergibt sich für jede dieser Hypothesen eine Analyse für die Artikel der taz und eine Analyse für die Artikel des Focus.

5.2 Forschungsdesign

Zur Beantwortung der Forschungsfrage wird eine Mischung aus quantitativer und qualitativer Forschung durchgeführt. Die notwendigen Daten stammen aus den HTML-Quelltexten der Online-Zeitungen taz und Focus. Diese wurden im Erhebungszeitraum vom 01.11.2020 bis 31.12.2020 erfasst. Die aus den HTML-Quelltexten extrahierten Daten unterscheiden sich je nach Tageszeitung. Für die Stichprobe aus der taz werden je Artikel folgende Daten extrahiert:

Variable	Beschreibung	Skalenniveau
Autor	Name des Autors	Nominalskala
Datum	Veröffentlichungsdatum des Artikels	Intervallskala
Titel	Titel des Artikels	Nominalskala
Intro	Kurzfassung des Artikels	Nominalskala
Ressort	Ressort in welchem der Artikel veröffentlicht wurde	Nominalskala
Textkörper	Gesamter Text des Artikels	Nominalskala
Schlüsselwörter	Thematische Schlüsselwörter zum Thema des Artikels	Nominalskala, mehrfach

Tabelle 5: Variablen und Skalenniveaus / taz

Durch weitere Rechenoperationen werden daraus folgende Merkmale je Artikel konstruiert:

Variable	Beschreibung	Skalenniveau
Wortanzahl	Anzahl der Wörter im Textkörper	Verhältnisskala
Text-Polarität	Polaritätswert des Textes (Ermittelt durch Sentiment-Analyse) im Intervall [-1;1]	Verhältnisskala
Text-Subjektivität	Subjektivitätswert des Textes (Ermittelt durch Sentiment-Analyse) im Intervall [0;1]	Verhältnisskala
Übereinstimmung Nachhaltigkeit	Der Textkörper enthält mindestens eines der ähnlichen Wörter zum Begriff „Nachhaltigkeit“[4].	Dichotom
Übereinstimmung Nachhaltigkeitsziel (X)	Der Textkörper enthält mindestens eines der ähnlichen Wörter zum Begriff (ermittelt durch NLP) des jeweiligen Nachhaltigkeitszieles. z.B.: „X=Gesundheit“.[5]	Dichotom, mehrfach

Tabelle 6: Berechnete Merkmale der Artikel / taz

[4] Bestimmung der ähnlichen Wörter mittels NLP
[5] Bestimmung der ähnlichen Wörter mittels NLP. Liste der assoziierten Begriffe siehe Anhang C Liste NLP

Die Stichprobe aus dem Focus dient als Vergleichsgruppe für die Textanalyse. Es werden folgende Daten aus dem Rohtext extrahiert:

Variable	**Beschreibung**	**Skalenniveau**
Titel	Titel des Artikels	Nominalskala
Intro	Kurzfassung des Artikels	Nominalskala
Textkörper	Gesamter Text des Artikels	Nominalskala

Tabelle 7: Variablen und Skalenniveaus / Focus

Weitere Merkmale des Focus werden analog zu den Variablen aus Tabelle 6 erhoben. Zur Exploration des Datensatzes werden folgende univariat deskriptive statistische Methoden angewendet:

- Absolute und relative Häufigkeiten
- Stabdiagramme
- Lagemaße
- Streumaße
- Box-Plot

Zur bivariaten Deskription kommen folgende Methoden zum Einsatz:

- Kontingenztabelle
- χ^2-Koeffizient
- Lineare Regression

Zur Überprüfung der Hypothesen dienen diese induktiven statistischen Methoden:

- t-Test
- χ^2-Anpassungstest
- Einfaktorielle Varianzanalyse (ANOVA-Test)

Absolute und relative Häufigkeiten

Im Rahmen einer deskriptiven Datenanalyse werden i. d. R. absolute und relative Häufigkeiten aus erhobenen Datensätzen untersucht. Die absolute Häufigkeit beschreibt die Anzahl der einzelnen Ausprägungen in einem Datensatz. Die relative

Häufigkeit hingegen beschreibt den Anteil der Ausprägungen an der Gesamtanzahl und stellt damit eine prozentuale Angabe dar. [29]

Stabdiagramme

Die erhobenen Daten können zur Veranschaulichung und zum Vergleich der Daten untereinander grafisch dargestellt werden. Dies kann beispielsweise mittels eines Stabdiagramms erfolgen. Dabei erfolgt die Darstellung der Ausprägungen des Merkmals auf der horizontalen Achse. Auf der vertikalen Achse werden die absoluten oder relativen Häufigkeiten der jeweiligen Ausprägung in Form eines Stabes dargestellt. Die Darstellung eines Stabdiagramms ist beispielhaft in Abbildung 15 abgebildet. [29]

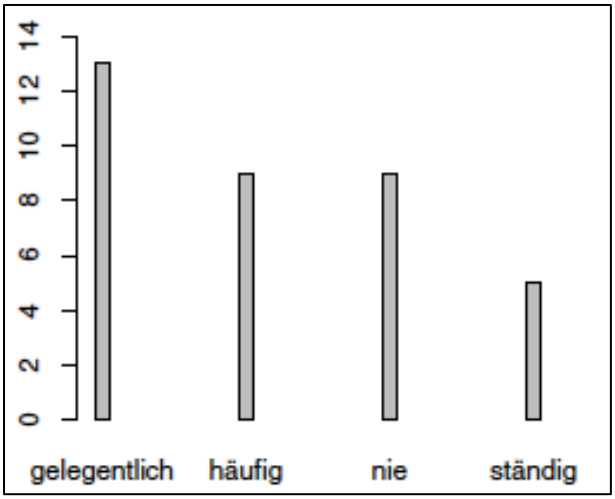


Abbildung 15: Darstellung Stabdiagramm [29]

Bei der Beschreibung von Verteilungen wird zwischen Lage- und Streumaßen unterschieden. Die Lagemaße dienen der Beschreibung des Zentrums der Verteilung. Hingegen wird durch die Streumaße dargestellt, inwieweit die Werte um das Zentrum streuen, ob die Verteilung symmetrisch ist oder ob es Ausreißer gibt. [29]

Lagemaße

Es sind unterschiedliche Lagemaße definiert, die für unterschiedliche Fragestellungen und Kontexte angewendet werden. Das arithmetische Mittel $\overline{x}$ ergibt sich aus der Aufsummierung aller Werte des Datensatzes, dividiert durch die Anzahl der Werte und wird im Kontext der Stichprobe eingesetzt. Dieses Maß ist gekennzeichnet von einer Empfindlichkeit auf Ausreißer, d. h. das arithmetische Mittel verändert sich merklich bei extremen Werten oder Abweichungen zum Rest des

Datensatzes. Im Kontext der Grundgesamtheit hingegen wird der Erwartungswert μ eingesetzt. Ein gegen Ausreißer resistenteres Lagemaß stellt das getrimmte Mittel dar. In diesem Fall werden die niedrigsten 10 % und die größten 10 % des Datensatzes herausgestrichen und aus den restlichen Daten das arithmetische Mittel berechnet. Dazu ist zuvor eine Sortierung des Datensatzes nötig. Der Median hat ebenfalls die Eigenschaft der Resistenz auf Extremwerte. So erfolgt die Platzierung des Medians in der Datenmitte, sodass die eine Hälfte der Daten oberhalb und die andere Hälfte der Daten unterhalb des Medians liegt. Der Modus wird bei kategorialen Merkmalen angewandt und gibt an, welche Ausprägung am häufigsten im Datensatz vorkommt. [29] Die Berechnung des arithmetischen Mittels, des Erwartungswertes und des Medians erfolgt wie nachfolgend dargestellt.

Arithmetisches Mittel $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Erwartungswert $\mu = \sum_{i\geq 1} x_i p_i$

Median $x_{med} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{für } n \text{ ungerade} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{für } n \text{ gerade} \end{cases}$

Streumaße

Die wichtigsten Streumaße sind die Standardabweichung und die Varianz. Mit Hilfe dieser Maße wird die Streuung der Daten um ihr arithmetisches Mittel $\overline{x}$ oder ihren Erwartungswert μ gemessen. Im Kontext der Stichprobe wird die Standardabweichung s angewandt und der Grundgesamtheit die Standardabweichung σ. [29] Die Berechnung der Standardabweichungen und Varianzen erfolgt wie nachfolgend dargestellt.

Standardabweichung $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$

$\sigma = \sqrt{\sum_{i\geq 1}(x_i - \mu)^2 f(x_i)}$

Varianz $s^2; \sigma^2$

Box-Plot

Zur grafischen Darstellung einer Verteilung wird häufig der Box-Plot angewandt. Durch die Darstellung werden verschiedene Lage- und Streumaße zusammengefasst und übersichtlich visualisiert. Somit ist eine schnelle Einschätzung zur

Streuung der Verteilung und ein Vergleich unter mehreren Verteilungen möglich. In Abbildung 16 ist der Aufbau eines Box-Plots beispielhaft dargestellt. Die Box des Box-Plots besteht von links nach rechts aus dem 25 %-Quantil, dem Median (50 %-Quantil) sowie dem 75 %-Quantil. Grundsätzlich werden anschließend auf beiden Seiten der Box Linien, auch Whisker genannt, bis zum minimalen und maximalen Wert der Verteilung gezeichnet. Eine Modifikation des Box-Plots liegt darin, zusätzlich das 5 %- und das 95 %-Quantil zu bestimmen und die Whisker lediglich bis zu diesen Grenzen zu zeichnen. Dabei erfolgt eine individuelle Kennzeichnung der Werte außerhalb der Grenzen, um Ausreißer deutlicher abzugrenzen. [29]

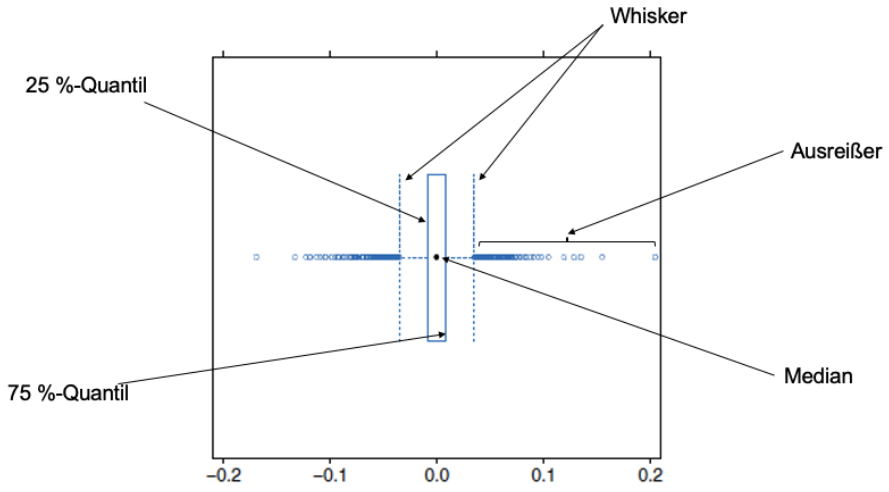


Abbildung 16: Darstellung Boxplot [29]

Die Berechnung der Quantile erfolgt wie nachfolgend dargestellt.

***p*-Quantile**

$$x_p = \begin{cases} x_{(\lfloor np \rfloor + 1)} & \text{für } np \text{ nicht ganzzahlig} \\ \frac{1}{2}(x_{(np)} + x_{(np+1)}) & \text{für } np \text{ ganzzahlig} \end{cases}$$

Kontingenztabelle

Eine Kontingenztabelle ist eine tabellenartig strukturierte Zusammenfassung von Häufigkeiten. Sie enthalten diese absoluten oder relativen Häufigkeiten aus der Kombination mehrerer Merkmalsausprägungen. Die Bezeichnung Kontingenz bedeutet in dem Zusammenhang das gemeinsame Auftreten zweier Merkmale. In dem Zuge werden mehrere Häufigkeiten miteinander verknüpfter Merkmale dargestellt. Zudem werden diese Häufigkeiten in der Kontingenztabelle durch die Randhäufigkeiten ergänzt. Die Größe der Kontingenztabelle wird durch die Anzahl der Zeilen und Spalten als „k × m“-Kontingenztabelle angeben. [29]

Die Form einer Kontingenztabelle für absolute Häufigkeiten ist in Abbildung 17 dargestellt. Dabei bezeichnen h_{ij} die absolute Häufigkeit der Kombination aus a_i und b_j, h_1. bis h_k. die Randhäufigkeiten des ersten Merkmals und $h_{\cdot 1}$ bis h_m. die Randhäufigkeiten des zweiten Merkmals. Die Kontingenztabelle stellt somit die gesamte Verteilung der beiden betrachteten Merkmale X und Y in absoluten Häufigkeiten dar. Bei der Betrachtung von relativen Häufigkeiten anstelle absoluter Häufigkeiten werden die einzelnen absoluten Häufigkeiten durch die Gesamtanzahl des Datensatzes dividiert. [29]

	b_1	...	b_m	
a_1	h_{11}	...	h_{1m}	$h_{1\cdot}$
a_2	h_{21}	...	h_{2m}	$h_{2\cdot}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
a_k	h_{k1}	...	h_{km}	$h_{k\cdot}$
	$h_{\cdot 1}$	...	$h_{\cdot m}$	n

Abbildung 17: Kontingenztabelle für absolute Häufigkeiten [29]

χ^2-Koeffizient, Kontingenzkoeffizient und korrigierter Kontingenzkoeffizient

Der χ^2-Koeffizient und der Kontingenzkoeffizient sind statistische Zusammenhangsmaße. Dabei werden die vorliegenden Häufigkeiten zweier Merkmale mit den, im Falle einer Unabhängigkeit zu erwartenden Häufigkeiten verglichen. Ist der errechnete χ^2-Wert groß, so besteht eine Abhängigkeit der Merkmale und ist der errechnete χ^2-Wert klein, dann besteht keine Abhängigkeit der Merkmale. Die Aussagekraft des χ^2-Koeffizienten ist jedoch gering, da die Werte von χ^2 von der Dimension der Kontingenztafel abhängen und somit eine geringe Vergleichbarkeit zwischen mehreren Datensätzen möglich ist. Um eine geringere Abhängigkeit von der Stichprobengröße zu erhalten, kann der Kontingenzkoeffizient eingesetzt werden. Damit zusätzlich eine Unabhängigkeit gegenüber der Dimension der Kontingenztafel besteht, wird der korrigierte Kontingenzkoeffizient eingesetzt. Die Berechnung des χ^2-Koeffizienten, des Kontingenzkoeffizienten und des korrigierten Kontingenzkoeffizienten ist nachfolgend dargestellt. [29]

χ^2-Koeffzient $\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(h_{ij}-\frac{h_{i\cdot}h_{\cdot j}}{n})^2}{\frac{h_{i\cdot}h_{\cdot j}}{n}},\ \chi^2 \in [0,\ \infty]$

Kontingenzkoeffizient $K = \sqrt{\frac{\chi^2}{n+\chi^2}}\ , K \in [0, \sqrt{\frac{M-1}{M}},$

$wobei\ M = \min\{k, m\}$

Korrigierter Kontingenzkoeffizient $K^* = \frac{K}{\sqrt{\frac{M-1}{M}}}, K^* \in [0,1]$

Lineare Regression

Die lineare Regression findet Anwendung, wenn gerichtete Zusammenhänge untersucht werden sollen. Dabei wird die Beziehung von zwei Merkmalen X und Y durch eine lineare Funktion angenähert beschrieben. Die Güte der errechneten Ausgleichsgeraden an die vorliegenden Daten kann dabei mit Hilfe des Bestimmtheitsmaßes beurteilt werden. Die Berechnung der linearen Regression und des Bestimmtheitsmaßes ist nachfolgend dargestellt. [29]

Lineare Regression $y_i = \alpha + \beta x_i + \epsilon_i, \qquad i = 1, \dots, n$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \quad \hat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

$\hat{\epsilon}_i = y_i - \hat{y}_i, \qquad i = 1, \dots, n \qquad mit\ \hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$

Bestimmtheitsmaß $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}, \quad es\ gilt: 0 \le R^2 \le 1, \quad R^2 = r_{XY}^2$

t-Test

Beim t-Test soll ein hypothetischer Erwartungswert μ_o mit dem tatsächlichen und bekannten Erwartungswert μ verglichen werden. Dabei wird anhand des Mittelwerts einer Stichprobe geprüft, ob der Mittelwert gleich dem eines vorgegebenen Wertes ist. Die Berechnung und Überprüfung erfolgen wie nachfolgend dargestellt. [29]

Hypothesen

(a) $H_o: \mu = \mu_o \quad H_1: \mu \neq \mu_o$

(b) $H_o: \mu \geq \mu_o \quad H_1: \mu < \mu_o$

(c) $H_o: \mu \leq \mu_o \quad H_1: \mu > \mu_o$

Teststatistik $T = \frac{\bar{x}-\mu_o}{s}\sqrt{n}$

Verteilung unter $\mu = \mu_o$ $t(n-1), f\ddot{u}r\ n \geq 30\ approximativ\ N(0,1)$

Ablehnungsbereich (a) $|T| > t_{1-\frac{\alpha}{2}}(n-1)$

(b) $T < t_{\alpha}(n-1) = -t_{1-\alpha}(n-1)$

(c) $T > t_{1-\alpha}(n-1)$

Einen Sonderfall stellt der Zweistichproben-t-Test dar. Dabei wird anhand der Mittelwerte zweier Stichproben geprüft, ob die Mittelwerte zweier Grundgesamtheiten gleich sind. Es unterscheidet sich im Vergleich zum Einstichproben-t-Test lediglich die Teststatistik und die Berechnung der Freiheitsgrade.

Teststatistik $T = \frac{\bar{x}_1 - \bar{x}_2}{s_p^* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

Verteilung unter $\mu = \mu_o$ $t(n_1 + n_2 - 2)$

χ^2-Aanpassungstest

Der χ^2-Anpassungstest dient der Prüfung, ob eine tatsächliche Verteilung einer vorgegebenen Verteilung entspricht. Dabei wird zunächst die Nullhypothese H_0 formuliert und unter Anwendung des χ^2-Anpassungstest geprüft. Die Berechnung und Überprüfung erfolgen wie nachfolgend dargestellt. [29]

Hypothesen $H_0: P(X = i) = \pi_i, i = 1, \dots, k$

$H_1: P(X = i) \neq \pi_i \; für \; mindestens \; ein \; i$

Teststatistik $\chi^2 = \sum_{i=1}^{k} \frac{(h_i - n\pi_i)^2}{n\pi_i}$

Verteilung unter H_0 $approximativ \; \chi^2(k-1) \; wenn \; n\pi_i \geq 1 \; für \; alle \; i, n\pi_i$

$\geq für \; mindestens \; 80\% \; der \; Zellen$

Ablehnungsbereich $\chi^2 > \chi^2_{1-\alpha}(k-1)$

Einfaktorielle Varianzanalyse

Durch die einfaktorielle Varianzanalyse können Unterschiede in den Erwartungswerten verschiedener Gruppen einer Verteilung beurteilt werden. Dabei dient die Varianzanalyse unterstützend bei der Entscheidung, ob die Unterschiede in den Mittelwerten der einzelnen Gruppen ausreichend sind, um davon auf Unterschiede in den dazugehörigen Grundgesamtheiten schließen zu können. Das Vorgehen zur einfaktoriellen Varianzanalyse ist nachfolgend dargestellt. [29]

Modell (I) $Y_{ij} = \mu_i + \epsilon_{ij}\,; \quad \epsilon_{ij} \sim N(0, \sigma^2), unabhängig\,;$

$i = 1, \dots, I\,;\; j = 1, \dots, n_i$

Modell (II) $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}\,;\; \sum_{i=1}^{I} n_i \alpha_i = 0\,;\; \epsilon_{ij} \sim N(0, \sigma^2), unabhängig\,;$

$i = 1, \dots, I\,; j = 1, \dots, n_i$

Die Schätzer für μ und α_i im Modell (II) sind gegeben als:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} Y_{ij} = \bar{Y}..$$

$$\hat{\alpha}_i = \bar{Y}_i. - \bar{Y}.. \quad mit \quad \bar{Y}_i. = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

Die Prüfgröße für das Testproblem

$$H_0: \alpha_1 = \dots = \alpha_I = 0 \quad gegen \quad H_1: mindestens\ zwei\ \alpha_i \neq 0$$

ist gegeben als

$$F = \frac{MQE}{MQR} = \frac{\sum_{i=1}^{I} n_i * (\bar{Y}_i. - \bar{Y}..)^2 / (I-1)}{\sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i.)^2 / (n-I)}$$

wobei H_0 zu verwerfen ist, falls

$$F > F_{1-\alpha}(I-1, n-I)$$

5.3 Durchführung der Datenanalyse

Im Zuge der Datenanalyse werden zunächst Metadaten der Stichprobe explorativ analysiert und ein Überblick über die Stichprobe erzeugt. Darauffolgend werden die Hypothesen H1 - H5 untersucht.

5.3.1 Metadaten der Stichprobe

Metadaten wurden im Wesentlichen für die Onlinezeitung taz erhoben. Der Focus dient hauptsächlich als Vergleichsgruppe für die Sprachanalyse. Im Folgenden wird die Exploration dieser Daten mit Blick auf die Forschungsfrage dargestellt. Ein Teilergebnis der Datensammlung für beide Onlinezeitungen ist in Tabelle 8 aufgeführt.

	taz	Focus
Anzahl der Artikel	1.801	13.379
Anzahl der Wörter im Textkörper	1.228.775	6.823.097
Durchschnittliche Anzahl der Wörter im Korpus/Artikel	682,27	509,99
Standardabweichung Anzahl der Wörter im Textkörper	361,29	621,04
Min. Anzahl der Wörter je Artikel	130	0
Max. Anzahl der Wörter je Artikel	3.443	10.812

Tabelle 8: Datensammlung taz und Focus

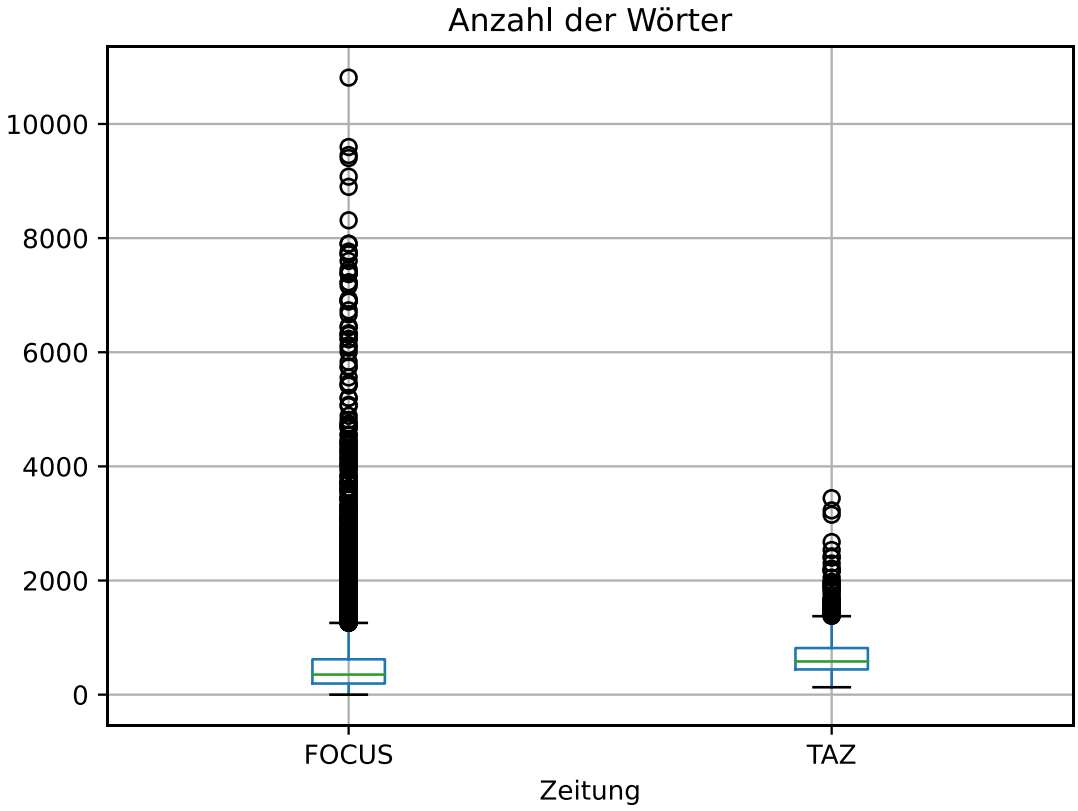


Abbildung 18: Boxplot - Wortanzahl - taz/Focus

Es zeigt sich, dass die Stichprobe für die Onlinezeitung Focus deutlich größer ist als die Stichprobe der Onlinezeitung taz. Für den Focus wurden ca. 7,5-mal so viele Artikel erhoben. Diese Beobachtung passt zur unterschiedlichen Größe und Reichweite beider Zeitungsformate. Im Focus wurden im gleichen Zeitraum mehr Artikel veröffentlicht. Es zeigen sich außerdem deutliche Unterschiede bei der Verteilung der Länge je Artikel beider Zeitungsformate. Im arithmetischen Mittel enthalten Artikel in der taz gut 30 % mehr Wörter als im Focus. Die Artikellänge

schwankt im Focus dafür erheblich mehr als in der taz, was aus der höheren Standardabweichung hervorgeht. Die Verteilungen beider Stichproben sind in Abbildung 18 als Box-Plots dargestellt.

Beide Verteilungen sind rechtsschief, wobei dies beim Focus stärker ausgeprägt ist. Der Median ist also kleiner als der Mittelwert. Das arithmetische Mittel der Wortanzahl wird durch Ausreißer, besonders lange Texte, nach oben hin beeinflusst. Beim Vergleich der Ausreißer beider Zeitungen zeigt sich ein deutlicher Unterschied. Der längste Artikel der taz hat knapp 4.000 Worte, wohingegen der längste Artikel im Focus[6] über 10.000 Wörter umfasst.

Für die taz wurden in dieser Studie je Artikel weitere Merkmale erhoben. Es handelt sich dabei um das Ressort, in dem der Artikel veröffentlicht wurde, den Autor des Artikels, die Schlüsselwörter des Artikels und das Veröffentlichungsdatum. Die Veröffentlichungen je Ressort sind im Stabdiagramm in Abbildung 19 dargestellt. Die Ressorts, welche dem Hauptressort „Öko“ zugeordnet werden, sind im Diagramm grün hervorgehoben. Es zeigt sich, dass die beiden meist frequentierten Ressorts „Deutschland“ und „Europa“ sind. Beide sind dem Hauptressort „Politik“ zugeordnet. Das Ökologieressort hat die drittmeisten Veröffentlichungen. Weitere Ressorts aus dem Hauptressort „Öko“, „Wissenschaft“, „Netzökonomie“, „Konsum“ und „Arbeit“ sind vergleichsweise niedrig frequentiert.

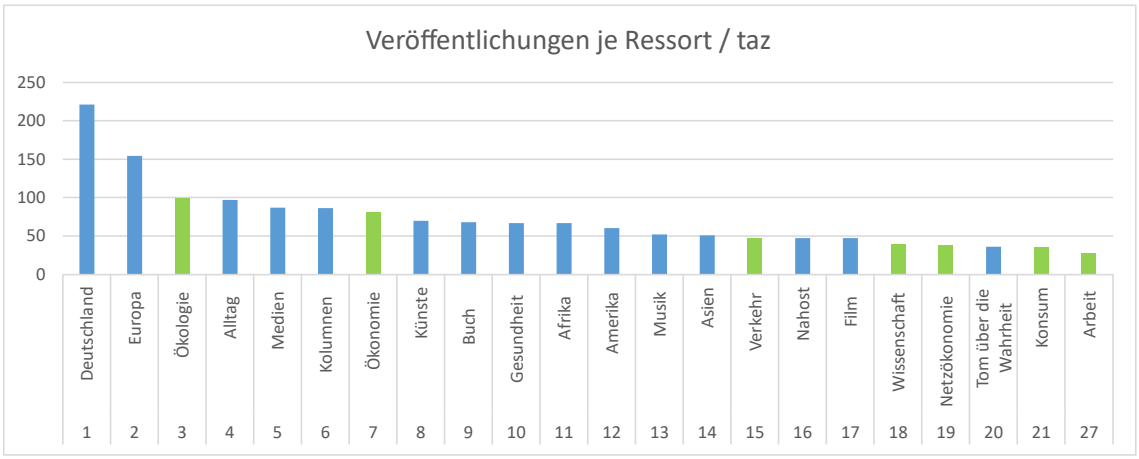


Abbildung 19: Häufigkeit der Veröffentlichung nach Ressort

Wird zusätzlich zur Veröffentlichung je Ressort die Dimension „Autor“ betrachtet zeichnet sich ab, dass in Ressorts, die viele Artikel veröffentlichen auch viele unterschiedliche Autoren publizieren. Im größten Ressort des Hauptressorts

[6] Es handelt sich dabei um einen Newsticker zur Fußball-Berichterstattung der 1. und 2. Bundesliga.

„Öko“ haben im Erhebungszeitraum 33 verschiedene Autoren publiziert, im Ressort Netzökonomie waren es hingegen 15 (siehe Abbildung 20 und Abbildung 21). Weiterhin ist zu beobachten, dass im Ressort „Netzökonomie“ dominierend von einer Person publiziert wird. Andere Autoren veröffentlichen nur vereinzelt. Im großen Ressort „Ökologie“ verhält sich dies etwas ausgeglichener, obwohl sich auch hier Autoren herauskristallisieren, die deutlich mehr veröffentlichen als andere. Im Anhang B sind die Verteilungen der Artikel je Autor für die übrigen fünf Ressorts des Hauptressorts „Öko“ einzusehen.[7] In Abbildung 22 sind die Autoren aufgeführt, welche die meisten Artikel in einem einzelnen Ressort veröffentlichen. In grün sind Autoren aus dem Hauptressort „Öko“ hervorgehoben. Hier stellt sich heraus, dass die Autorin *Svenja Bergt* die meisten Artikel für ein einzelnes Ressort publiziert hat. Dabei ist das betreffende Ressort – „Netzökonomie“ – im Vergleich eher niedrig frequentiert. Insgesamt zeichnet sich ab, dass in den Öko-Ressorts eher einzelne Autoren die Berichterstattung bestimmen.

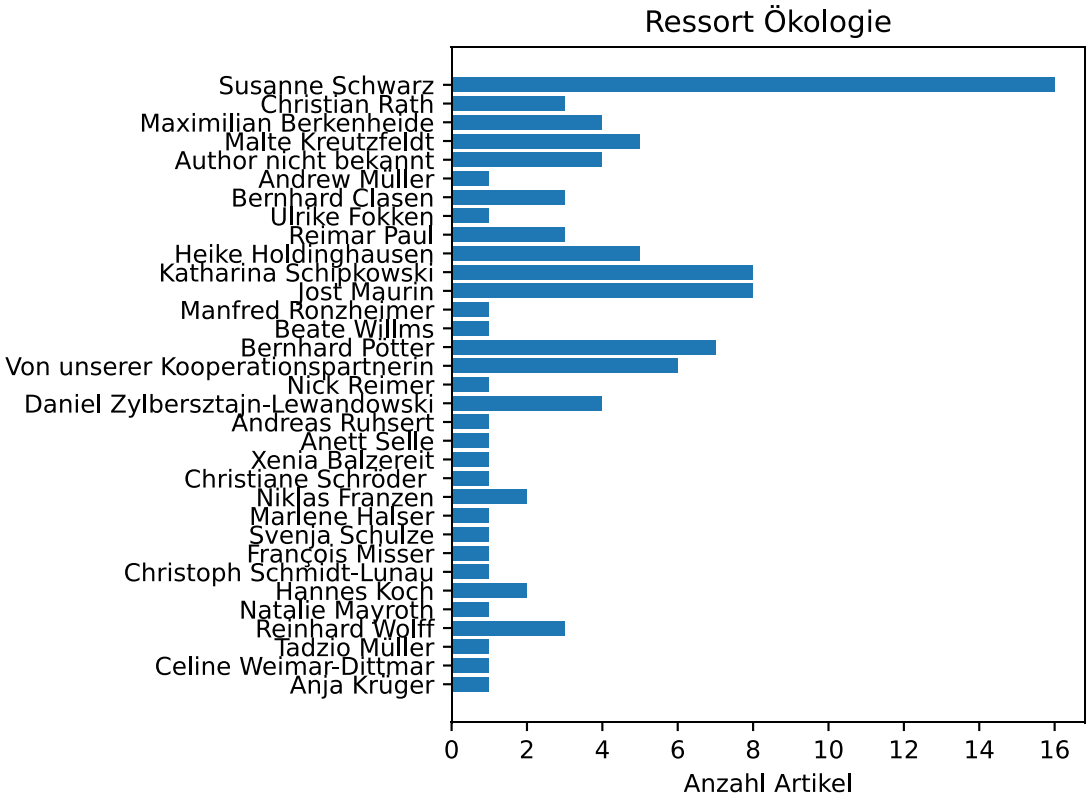


Abbildung 20: Anzahl publizierter Artikel je Autor - Ressort Ökologie / taz

[7] Die Ausprägung „Autor nicht bekannt“ resultiert aus der Bereinigung des Datensatzes. In diesem Fall konnten aus den Quelltexten keine Autoren entnommen werden.

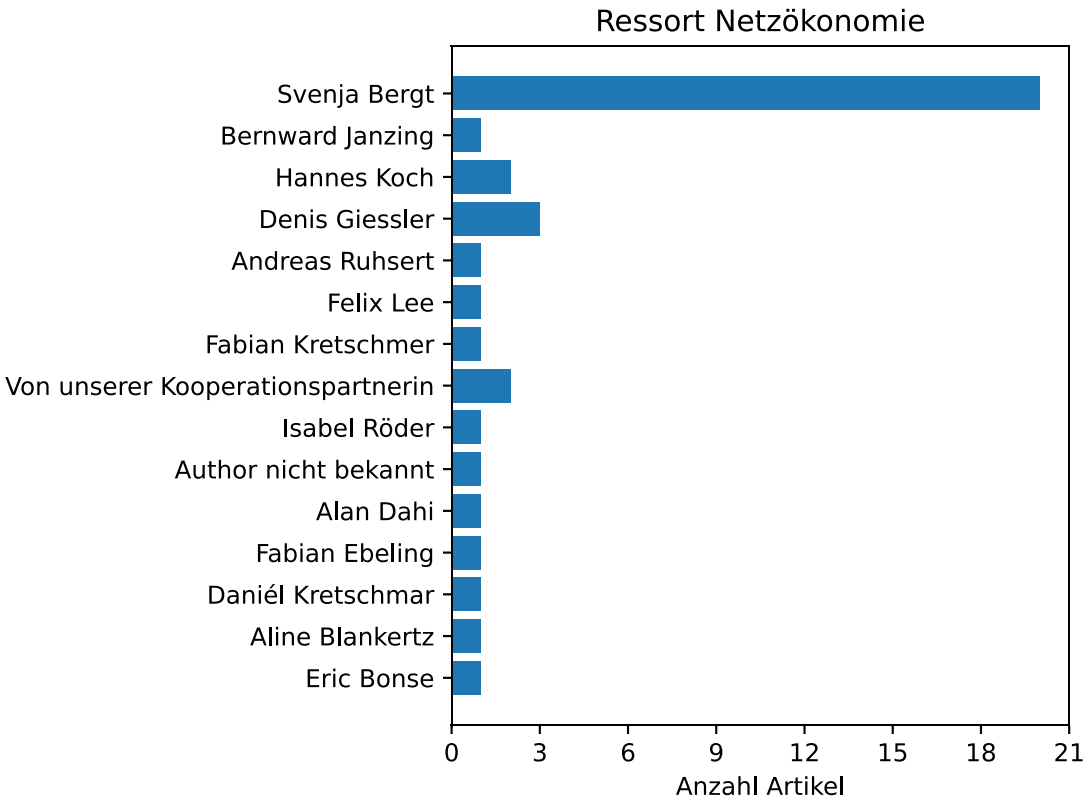


Abbildung 21: Anzahl publizierter Artikel je Autor - Ressort Netzökonomie / taz

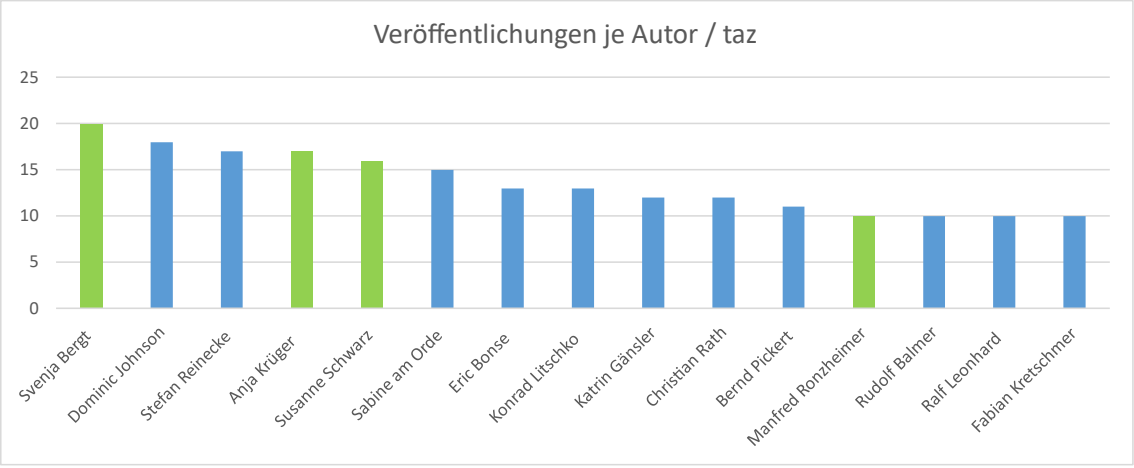


Abbildung 22: Anzahl publizierter Artikel je Autor – Gesamt / taz

In einer weiteren bivariaten Betrachtung können die durchschnittliche Wortanzahl der Artikel und das Ressort gegenübergestellt werden. Bei einer Betrachtung der vier großen Hauptressorts und einer Klassenbildung der Wortanzahl ergibt sich eine Kontingenztafel entsprechend der Tabelle 9. Es zeigt sich, dass die meisten Artikel in jedem Ressort zwischen 400 und 800 Wörter umfassen.

	Wortanzahl des Artikels				
Resort	**<400**	**400-800**	**800-1.200**	**>= 1.200**	**Summe**
Gesellschaft	60	216	79	81	**436**
Kultur	14	155	77	53	**299**
Öko	111	203	40	14	**368**
Politik	133	384	74	37	**628**
Summe	**318**	**958**	**270**	**185**	**1.731**

Tabelle 9: Kontingenztafel Artikellänge-Ressort / taz

Ob ein Unterschied zwischen den Mittelwerten der einzelnen Gruppen besteht, ist u. a. aufgrund der Zusammenfassung der Klassen nicht direkt abzusehen. Eine Varianzanalyse dient der Untersuchung der Mittelwerte in den einzelnen Gruppen. Im Gegensatz zur Betrachtung in der Kontingenztafel geht das Messniveau der Variable Wortanzahl nicht durch die Klassenbildung verloren. Testproblem und -ergebnis der durchgeführten Varianzanalyse sind in Tabelle 10 zusammengefasst. Demnach kann signifikant verworfen werden, dass kein Unterschied zwischen den Mittelwerten der einzelnen Ressorts besteht (sehr kleiner p-Wert).

Testparameter und Ergebnisse	
H_0	Es besteht kein Unterschied zwischen den Mittelwerten der Wortanzahl der einzelnen Ressorts.
H_1	Mindestens zwei Mittelwerte unterscheiden sich voneinander.
F	59,036
p	2,48E-36

Tabelle 10: Varianzanalyse Artikellänge-Ressort / taz

Ein weiteres Merkmal aus dem Datensatz der taz sind die Schlüsselwörter des Artikels. Diese geben in Stichworten den Inhalt des Artikels wieder und ermöglichen eine thematische Einordnung. Je Artikel gibt es mindestens ein Schlüsselwort; im Regelfall sind es mehrere. Abbildung 23 zeigt die 50 meistgenannten Schlüsselwörter in absteigender Reihenfolge. Begriffe, die thematisch mit einem der 17 Nachhaltigkeitsziele assoziiert werden können, sind grün hervorgehoben. In der Häufigkeitsverteilung der Schlüsselwörter spiegeln sich aktuelle Themen der Berichterstattung während des Erhebungszeitraumes wider. Beispielsweise

erfährt die Covid-19-Pandemie große mediale Aufmerksamkeit in der taz (Schlüsselworte: Corona-Maßnahmen, #1; Corona-Virus, #2; Corona-Impfstoff, #3; Lockdown, #8; etc.). Auch die US-Präsidentschaftswahl ist in der Datenreihe zu erkennen (Schlüsselworte: USA, #4; Donald Trump, #5; Joe Beiden, #11). Schlüsselworte, die dem Themenfeld Nachhaltigkeit zuzuordnen sind (grün), können in der Top-50-Darstellung eher im mittleren bis hinteren Bereich identifiziert werden (Schlüsselworte: Landwirtschaft, #18; Klimataz, #19; Menschenrechte, #22; etc.). Es ist zu hinterfragen, inwieweit Ereignisse wie die US-Wahl oder die Covid-19-Pandemie dieses Bild unter Anbetracht des Erhebungszeitraumes verfälschen. Es ist zu vermuten, dass zumindest die US-Wahl eine geringere Präsenz haben würde, wenn der Erhebungszeitraum ein größerer wäre.

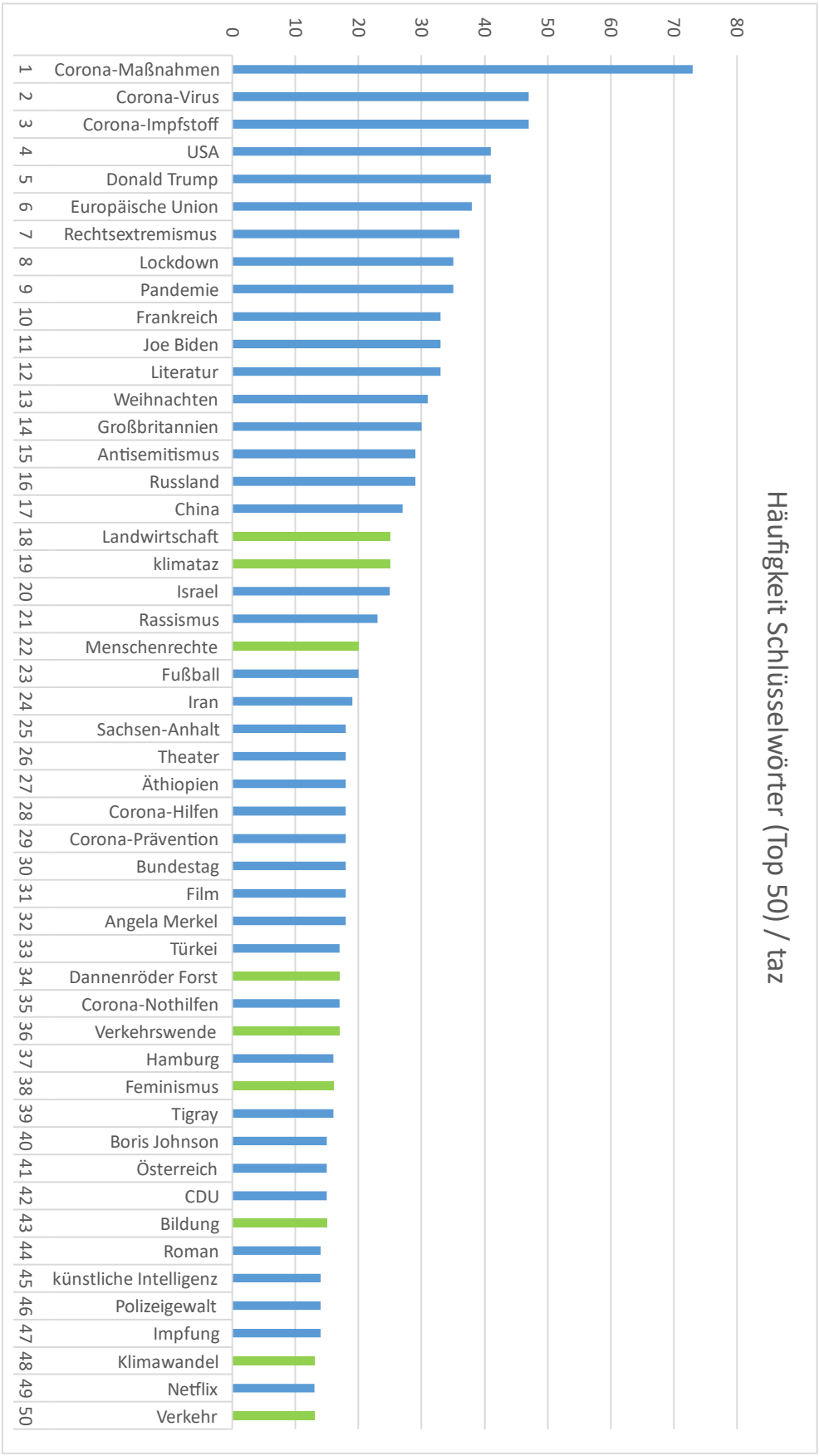


Abbildung 23: Häufigkeit von Schlüsselwörtern (Top 50) / taz

Abschließend soll der Zeitverlauf der Veröffentlichungen in der taz betrachtet werden. Zur Einordnung dieser Daten muss beachtet werden, dass die Datenerhebung mit dem Stichtag 01.11.2020 begonnen hat. Nichtsdestotrotz sind Artikel, die vor diesem Stichtag veröffentlicht wurden im Datensatz enthalten. Je nachdem wie regelmäßig im entsprechenden Ressort veröffentlicht wird, ist dieser retrospektive Zeitraum kleiner oder größer. In gering frequentierten Ressorts sind noch Artikel weit aus der Vergangenheit verfügbar. In Abbildung 25 ist die Zeitreihe der Veröffentlichungen pro Tag über alle Ressorts in der taz für die Monate Oktober, November und Dezember dargestellt. Der Stichtag des Erhebungsbeginns ist rot hervorgehoben. Es lässt sich erkennen, dass der Datensatz vor dem Stichtag deutlich ausgedünnt ist. Es existieren auch Artikel im Datensatz, die noch weit vor dem Monat Oktober datiert sind. Diese werden hier jedoch nicht mehr betrachtet. Der aussagekräftige Bereich der Zeitreihe beschränkt sich auf die Monate November und Dezember.

Bei einer separierten Betrachtung beider Monate zeichnet sich ab, dass die Menge an veröffentlichten Artikeln im Monat November ansteigt und über den Monat Dezember wieder abfällt. Mittels einer Regressionsanalyse (siehe Abbildung 24) lässt sich dieses Bild bestätigen. Die Regressionsgerade für den Monat November hat eine Steigung von 1,32, wohingegen die Gerade für den Dezember eine Steigung von -0,67 aufweist. Das Bestimmtheitsmaß der Regression für den Monat Dezember liegt allerdings nur bei 0,32, weshalb der abfallenden Gerade keine zu große Bedeutung zugemessen werden sollte. Der Abfall könnte durch die Feiertage am Ende des Monats zu erklären sein.

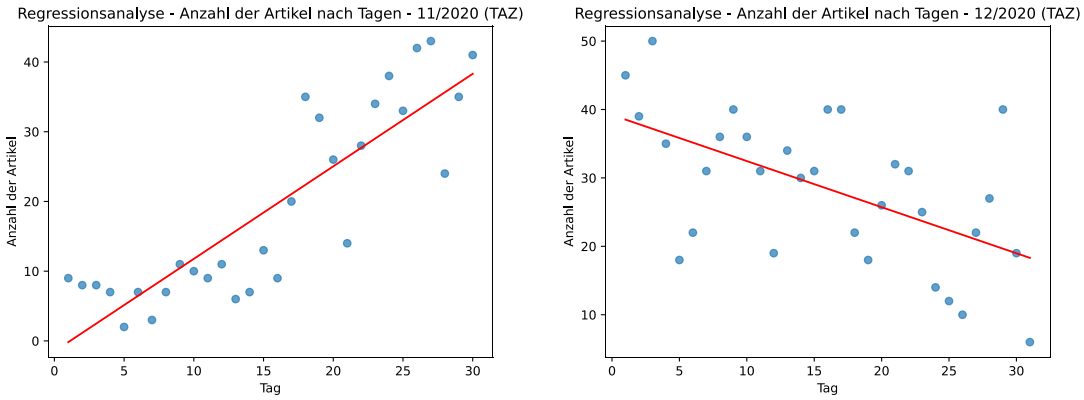


Abbildung 24: Regressionsgeraden der Veröffentlichungen je Monat (für Nov. und Dez.) / taz

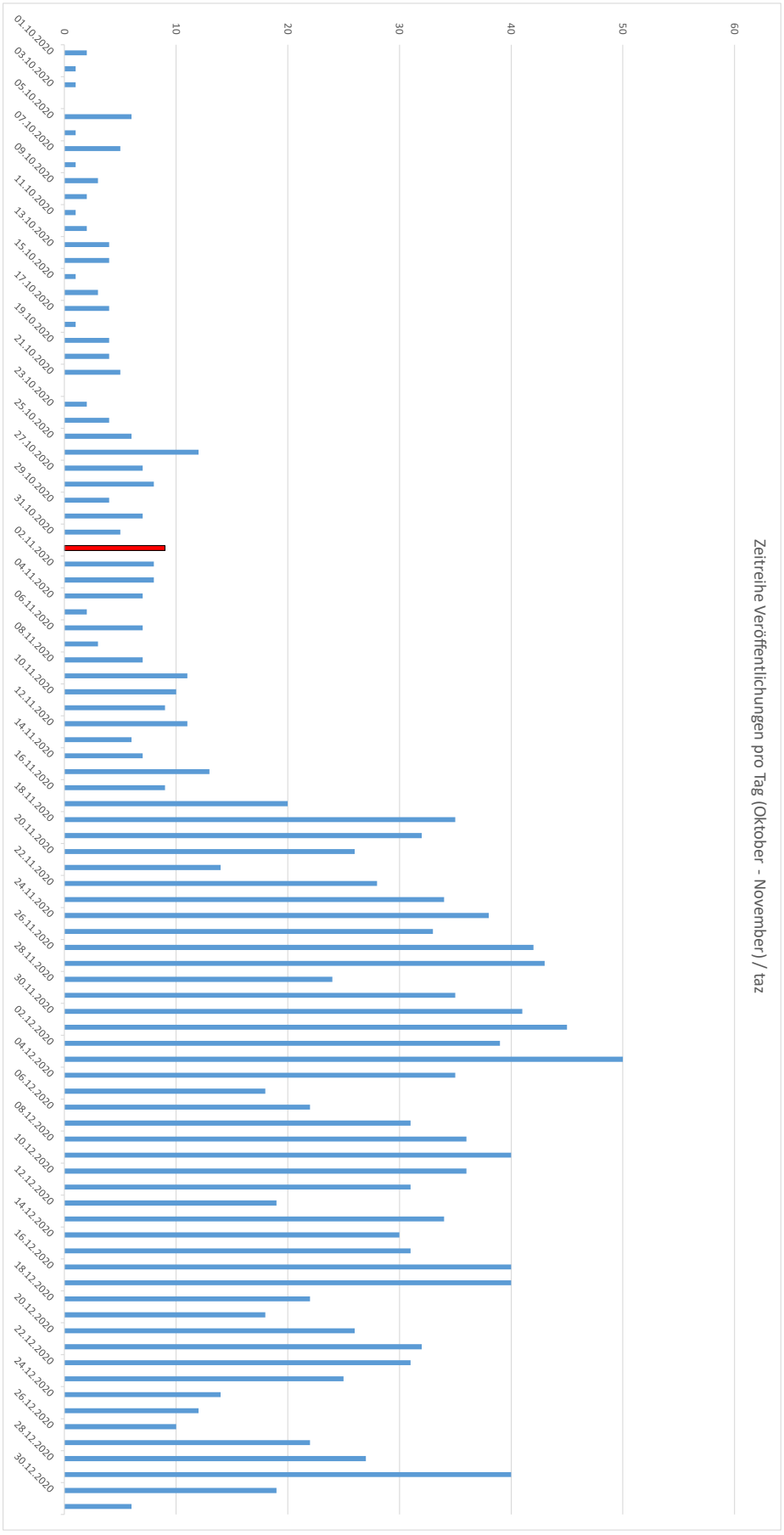


Abbildung 25: Zeitreihe der Veröffentlichungen pro Tag / taz

5.3.2 Überprüfung der Hypothesen

Hypothese H1:

Die Berichterstattung zu den einzelnen Nachhaltigkeitszielen ist unausgeglichen.

Zur Beantwortung dieser These werden die Textkörper auf die Verwendung bestimmter Wörter überprüft. Diese Wörter sollen ein Hinweis dafür sein, dass es in diesem Artikel inhaltlich um eines der Nachhaltigkeitsziele geht oder das Thema zumindest angeschnitten wird. Es soll also ein Indikator für Präsenz in der Berichterstattung des jeweiligen Zieles gesucht werden. Diese Listen von assoziativen Wörtern werden je Nachhaltigkeitsziel in zwei Schritten gebildet. Zunächst wird auf Basis der Stellungnahme der Bundesregierung zur Nachhaltigkeitspolitik [30] je ein Wort pro Ziel ausgewählt (im Folgenden: Schlagwort), welches möglichst genau mit dem Thema des Ziels zu assoziieren ist. Im zweiten Schritt werden mittels NLP die ähnlichsten Wörter zum Schlagwort ermittelt. Es wird anhand eines Vergleichs der Wortvektoren entschieden. Beispielhaft für das erste Nachhaltigkeitsziel „Armut in jeder Form und überall beenden“, stellt sich die Auswahl wie folgt dar:

Schlagwort: Armut → Ähnlichstes Wort: Arbeitslosigkeit

Übereinstimmung 77,5%

Die Auswahl erfolgt zunächst anhand des Modells aus dem Textkorpus der taz und des Focus. Aufgrund einer zu geringen Stichprobengröße sind die Ergebnisse jedoch nicht verwertbar. Daher wird ein vortrainiertes Modell, das *German-Model*, für die Auswahl der Wortlisten herangezogen. Das Ergebnis dieser Auswertung ist in Anhang C zu finden.

Alle Textkörper der Zeitungen werden auf das Vorhandensein von mindestens einem der Wörter überprüft. Für jedes Ziel wird gezählt, in wie vielen Artikeln mindesten ein Wort aus der Wortliste auftaucht. Daraus ergibt sich die Häufigkeitsverteilung für beide Zeitungen in Tabelle 11. Auf Grund der gesellschaftlichen Vorkommnisse im Erhebungszeitraum sind die Ergebnisse für die Ziele *Gesundheit und Wohlergehen* sowie *Nachhaltige Städte und Gemeinden* nur be-

dingt aussagekräftig, da ähnliche Wörter häufig auch in Verbindung mit der globalen Pandemie gebracht werden. Außerdem wird das 17. Nachhaltigkeitsziel nicht betrachtet, da keine aussagekräftigen Suchwörter ermittelt werden können.

Nachhaltigkeitsziel	**Absolute Häufigkeit**	**Anteilig an Artikeln ges.**
Gesundheit und Wohlergehen	2.900	19,1 %
Nachhaltige Städte und Gemeinden	2.206	14,5 %
Ozeane und Meere	1.293	8,5 %
Industrie, Innovation und Infrastruktur	1.216	8,0 %
Maßnahmen zum Klimaschutz	1.187	7,8 %
Sauberes Wasser und Sanitäreinrichtungen	993	6,5 %
Geschlechter Gleichheit	964	6,4 %
Hochwertige Bildung	520	3,4 %
Frieden, Gerechtigkeit und starke Institutionen	450	3,0 %
Keine Armut	381	2,5 %
Menschenwürdige Arbeit und Wirtschaftswachstum	311	2,0 %
Kein Hunger	289	1,9 %
Bezahlbare und saubere Energie	288	1,9 %
Ökosysteme an Land	225	1,5 %
Weniger Ungleichheiten	161	1,1 %
Nachhaltiger Konsum und Produktion	47	0,3 %

Tabelle 11: Erwähnung von Nachhaltigkeitszielen in Berichterstattung

In der Verteilung zeigt sich, dass manche Ziele zu einem gewissen Maß gleiche Präsenz in der Berichterstattung erfahren. Ziele wie *Weniger Ungleichheiten* und *Nachhaltiger Konsum und Produktion* hingegen, haben jedoch nahezu keine Präsenz. Aus der rein deskriptiven Betrachtung lässt sich schlussfolgern, dass nicht von einer Gleichverteilung der Berichterstattung über die einzelnen Ziele gesprochen werden kann. Diese Beobachtung bestätigt die Hypothese H1.

Hypothese H2:

Die taz berichtet ausgeglichener zu den verschiedenen Nachhaltigkeitszielen als der Focus.

Eine ausgeglichene Berichterstattung zu den einzelnen Nachhaltigkeitszielen impliziert, dass die Berichterstattung je Ziel ungefähr gleichverteilt ist. Wie schon den Ausführungen zu der Hypothese H1 zu entnehmen ist, kann den Daten beider Zeitungen kaum eine Gleichverteilung unterstellt werden. Dennoch soll untersucht werden, welche Zeitung einer ausgeglichenen Berichterstattung zu den einzelnen Zielen näherkommt. Dazu wird die vorliegende beobachtete Verteilung jeder Zeitung mit der zu erwartenden Verteilung bei gleichverteilter Berichterstattung verglichen (siehe Tabelle 12).

Nachhaltigkeitsziel	**Beobachtete Verteilung, taz**	**Annahme Gleichverteilung, taz**	**Beobachtete Verteilung, Focus**	**Annahme Gleichverteilung, Focus**
Gesundheit und Wohlergehen	518	183,5	2.382	655,93
Nachhaltige Städte und Gemeinden	483	183,5	1.723	655,93
Leben unter Wasser	214	183,5	1.079	655,93
Industrie, Innovation und Infrastruktur	279	183,5	937	655,93
Maßnahmen zum Klimaschutz	329	183,5	858	655,93
Sauberes Wasser und Sanitäreinrichtungen	152	183,5	841	655,93
Geschlechter Gleichheit	137	183,5	827	655,93
Hochwertige Bildung	173	183,5	347	655,93
Frieden, Gerechtigkeit und starke Institutionen	149	183,5	301	655,93

Nachhaltigkeits-ziel	Beobach-tete Verteilung, taz	Annahme Gleichver-teilung, taz	Beobachtete Verteilung, Focus	Annahme Gleichver-teilung, Focus
Keine Armut	139	183,5	242	655,93
Menschenwürdige Arbeit und Wirt-schaftswachstum	54	183,5	257	655,93
Kein Hunger	84	183,5	205	655,93
Bezahlbare und saubere Energie	99	183,5	189	655,93
Leben an Land	43	183,5	182	655,93
Weniger Ungleich-heiten	78	183,5	83	655,93
Nachhaltiger Kon-sum und Produk-tion	5	183,5	42	655,93

Tabelle 12: Vergleich beobachtete Verteilung und Gleichverteilung

Als Maßgröße der Ähnlichkeit zwischen beobachteter Verteilung und Gleichverteilung dient der Chi-Quadrat-Koeffizient. Aus den vorhandenen Verteilungen ergeben sich $\chi^2_{taz} = 1829{,}9$ / $\chi^2_{Focus} = 9731{,}7$. Es gilt somit $\chi^2_{0,95} << \chi^2_{taz} < \chi^2_{Focus}$. Beide Chi-Quadratwerte der Teststatistik sind deutlich größer als das 0,95-Quantil der Chi-Quadrat-Verteilung (s. Anhang F), weshalb ein Anpassungstest signifikant verworfen werden kann. Da $\chi^2_{taz} < \chi^2_{Focus}$ gilt, kann geschlussfolgert werden, dass die Verteilung der beobachteten Werte in der Stichprobe des Focus weniger einer Gleichverteilung ähnelt als die Verteilung der beobachteten Werte in der Stichprobe der taz. Somit kann die Hypothese H2 bestätigt werden.

Hypothese H3:

Nachhaltigkeitsartikel sind bezüglich des Textumfangs länger.

Als Indikator für die Artikellänge wird die Anzahl der Wörter im Textkörper gewählt. Für die Untersuchung dieser These werden aus allen Artikeln zwei Mengen gebildet. Eine Menge beinhaltet Artikel, die mit dem Begriff Nachhaltigkeit zu assoziieren sind. Dazu wird, nach dem Schema zur Überprüfung von Hypothese

H1, eine Wortliste mit ähnlichen Begriffen zum Wort Nachhaltigkeit mittels NLP gebildet (siehe Anhang C). Die durchschnittliche Anzahl der Wörter je Artikel wird daraufhin verglichen. Der Vergleich findet für beide Zeitungen getrennt statt.

taz	**Mit Nachhaltigkeit**	**Ohne Nachhaltigkeit**
Anzahl der Artikel	55	1.764
Anzahl der Wörter im Textkörper	40.835	1.187.940
Durchschnittliche Anzahl der Wörter im Textkörper je Artikel	742,5	680,4
Focus		
Anzahl der Artikel	172	13.207
Anzahl der Wörter im Textkörper	172.332	6.650.765
Durchschnittliche Anzahl der Wörter im Textkörper je Artikel	1.001,9	503,6

Tabelle 13: Analyse der Artikellänge - Nachhaltigkeit

Es zeigt sich, dass die Artikel mit Nachhaltigkeitsbezug in beiden Zeitungen durchschnittlich mehr Wörter enthalten als Artikel ohne Nachhaltigkeitsbezug. Dieser Unterschied ist beim Focus jedoch ausgeprägter als bei der taz. Zur Überprüfung der deskriptiven Statistik wird jeweils ein Zweistichproben-t-Test auf Gleichheit der Mittelwerte ($H_o: \mu_1 = \mu_2 \;\; / \;\; H_1: \mu_1 \neq \mu_2$) durchgeführt. Die Parameter und Ergebnisse beider Tests sind in Tabelle 14 und Tabelle 15 aufgelistet. Demnach kann die Hypothese – die durchschnittliche Anzahl der Wörter in Artikeln mit und ohne Nachhaltigkeitsbezug ist gleich – für die taz zum Signifikanzniveau von 5% nicht verworfen werden.[8] Der Unterschied der Mittelwerte ist bei gegebener Stichprobengröße zum Verwerfen der Hypothese nicht signifikant genug. Für den Focus hingegen kann die Hypothese zu nahe jedem Signifikanzniveau verworfen werden (p-Wert sehr klein).

[8] Vergleich mit dem entsprechenden Quantil der t-Verteilung in Anhang D. Bei weiteren t-Tests in diesem Kapitel erfolgt kein erneuter Verweis auf die Verteilungsparameter.

Testparameter und Ergebnisse	
H_0	Die durchschnittliche Anzahl der Wörter in Artikeln mit und ohne Nachhaltigkeitsbezug im Focus ist gleich.
H_1	Die durchschnittliche Anzahl der Wörter in Artikeln mit und ohne Nachhaltigkeitsbezug im Focus ist ungleich.
Df	1.799
t	1,25
p	2,98E-1

Tabelle 14: Zweistichproben-t-Test auf Gleichheit der Mittelwerte Wortanzahl / taz

Testparameter und Ergebnisse	
H_0	Die durchschnittliche Anzahl der Wörter in Artikeln mit und ohne Nachhaltigkeitsbezug im Focus ist gleich.
H_1	Die durchschnittliche Anzahl der Wörter in Artikeln mit und ohne Nachhaltigkeitsbezug im Focus ist ungleich.
Df	13.377
t	10,49
p	1,1E-25

Tabelle 15: Zweistichproben-t-Test auf Gleichheit der Mittelwerte Wortanzahl / Focus

Folglich kann die Hypothese H3 nur teilweise bestätigt werden. Für den Focus lässt sich statistisch signifikant zeigen, dass die Artikel mit Nachhaltigkeitsbezug länger sind als Artikel ohne Nachhaltigkeit. Für die taz ist dies auf Basis der vorhandenen Daten nicht signifikant darzulegen. In der deskriptiven Statistik ist ein Unterschied dennoch zu erkennen.

Hypothese H4:

Die Berichterstattung zu Nachhaltigkeitsthemen ist eher negativ geprägt.

Zur Überprüfung dieser Hypothese wird der Parameter *Polarität* der Sentiment-Analyse als Indikator verwendet. Dieser ist ein Maß dafür wie positiv oder negativ ein Text geschrieben ist und liegt im Wertebereich [-1, 1]. Wie auch der Untersuchung zur Hypothese H3 werden die Texte je Zeitung in zwei Mengen unterteilt.

Das Kriterium ist ebenfalls die Wortliste zum Begriff Nachhaltigkeit. Der Polaritätswert für jeden Artikel wird errechnet. Die daraus resultierenden Verteilungsparameter sind in Tabelle 16 aufgeführt.

taz	**Mit Nachhaltigkeit**	**Ohne Nachhaltigkeit**
Anzahl der Artikel	55	1.764
Durchschnittliche Text-Polarität	0,0616	0,0465
Standardabweichung der Text-Polarität	0,1058	0,1173
Focus		
Anzahl der Artikel	172	13.207
Durchschnittliche Text-Polarität	0,1235	0,0943
Standardabweichung der Text-Polarität	0,1076	0,1421

Tabelle 16: Analyse der Text-Polarität - Nachhaltigkeit

Für beide Zeitungen ist der Wert der Textpolarität sowohl in Artikeln mit Nachhaltigkeitsbezug als auch in Artikeln ohne Nachhaltigkeitsbezug leicht im positiven Bereich. Insgesamt zeigt sich, dass Texte im Focus einen etwas höheren Polaritätswert besitzen als die Texte in der taz. Bei beiden Zeitungen sind die Polaritätswerte jeweils bei Artikeln mit Nachhaltigkeitsbezug höher als bei Artikeln ohne Nachhaltigkeitsbezug. Eine negativ geprägte Berichterstattung würde durch einen Polaritätswert kleiner 0 indiziert werden. Daher werden zur Überprüfung der Hypothese zwei t-Tests ($H_o: \mu = \mu_0$ / $H_1: \mu \neq \mu_0$) mit $\mu_0 = 0$ durchgeführt.

Testparameter und Ergebnisse	
H_0	Der Mittelwert der Polarität ist bei taz Artikeln mit Nachhaltigkeitsbezug gleich 0.
H_1	Der Mittelwert der Polarität ist bei taz Artikeln mit Nachhaltigkeitsbezug ungleich 0.
Df	54
t	4,273
p	7,81E-5

Tabelle 17: t-Test auf Mittelwert der Polarität / taz

Testparameter und Ergebnisse	
H_0	Der Mittelwert der Polarität ist bei taz Artikeln mit Nachhaltigkeitsbezug gleich 0.
H_1	Der Mittelwert der Polarität ist bei taz Artikeln mit Nachhaltigkeitsbezug ungleich 0.
Df	171
t	15,004
p	5,16E-31

Tabelle 18: t-Test auf Mittelwert der Polarität / Focus

In beiden Tests kann die Hypothese signifikant verworfen werden (sehr kleiner p-Wert). Demnach kann auch die Hypothese H4 verworfen werden. Artikel über Nachhaltigkeitsthemen sind nicht negativ geprägt. Jedoch ist auch die restliche Berichterstattung beider Zeitungen im Schnitt nicht negativ geprägt.

Hypothese H5:

Die Berichterstattung zu Nachhaltigkeitsthemen ist im Vergleich zur restlichen Berichterstattung eher subjektiv geprägt.

Zur Überprüfung dieser Hypothese wird der Parameter *Subjektivität* der Sentiment-Analyse als Indikator verwendet. Dieser ist ein Maß dafür, wie objektiv oder subjektiv ein Text geschrieben ist und liegt im Wertebereich [0, 1]. Sehr objektive Texte haben einen Wert von nahezu 0. Sehr subjektive Texte haben einen Wert von nahezu 1. Nach dem gleichen Schema wie bei der Untersuchung zu Hypothese H4 werden Mengen gebildet. Die Verteilungsparameter des Subjektivitätswertes sind entsprechend in Tabelle 19 aufgeführt.

taz	**Mit Nachhaltigkeit**	**Ohne Nachhaltigkeit**
Anzahl der Artikel	55	1.764
Durchschnittliche Text-Subjektivität	0,076	0,0708
Standardabweichung der Text- Subjektivität	0,0403	0,0533
Focus		
Anzahl der Artikel	172	13.207
Durchschnittliche Text-Subjektivität	0,0802	0,0559
Standardabweichung der Text- Subjektivität	0,0573	0,0599

Tabelle 19: Analyse der Text-Subjektivität - Nachhaltigkeit

Es zeigt sich, dass die Texte in beiden Zeitungen eher objektiv verfasst sind. Der Wert der Textpolarität liegt nahe 0. Bei der taz besteht nur ein sehr geringer Unterschied der durchschnittlichen Subjektivität bei Texten mit Nachhaltigkeitsbezug und Texten ohne Nachhaltigkeitsbezug. Im Focus hingegen ist ein Unterschied erkennbar. Texte mit Nachhaltigkeitsbezug sind hier etwas subjektiver verfasst als Texte ohne Nachhaltigkeitsbezug. Anhand von Zweistichproben-t-Tests ($H_o: \mu_1 = \mu_2 \;\; / \;\; H_1: \mu_1 \neq \mu_2$) soll für beide Zeitungen untersucht werden, ob die Mittelwerte der Subjektivität der beiden Text-Mengen unterschiedlich sind.

Testparameter und Ergebnisse	
H_0	Der Mittelwert der Text-Subjektivität ist bei taz Artikeln mit und ohne Nachhaltigkeitsbezug gleich.
H_1	Der Mittelwert der Text-Subjektivität ist bei taz Artikeln mit und ohne Nachhaltigkeitsbezug ungleich.
Df	1.799
t	0,856
p	0,392

Tabelle 20: Zweistichproben-t-Test auf Gleichheit der Mittelwerte Text-Subjektivität / taz

Testparameter und Ergebnisse	
H0	Der Mittelwert der Text-Subjektivität ist bei Focus Artikeln mit und ohne Nachhaltigkeitsbezug gleich.
H1	Der Mittelwert der Text-Subjektivität ist bei Focus Artikeln mit und ohne Nachhaltigkeitsbezug ungleich.
Df	13.377
t	5,29
p	1,24E-07

Tabelle 21: Zweistichproben-t-Test auf Gleichheit der Mittelwerte Text-Subjektivität / Focus

Für die taz kann die Hypothese zum Signifikanzniveau von 5% nicht verworfen werden. Es besteht also kein statistisch signifikanter Unterschied des Mittelwertes der Subjektivität in beiden Stichproben. Beim Focus hingegen kann die Hypothese signifikant verworfen werden. Folglich ist die Hypothese H5 für den Focus zu bestätigen, für die taz jedoch nicht.

5.4 Untersuchungsergebnisse

In der Tabelle 22 werden die Ergebnisse der Hypothesenüberprüfung dargestellt.

Hypothese	Ergebnis der Hypothesenüberprüfung
H1	Die Berichterstattung zu den einzelnen Nachhaltigkeitszielen ist übergreifend unausgeglichen.
H2	Die *taz* berichtet ausgeglichener zu den verschiedenen Nachhaltigkeitszielen als der *Focus*.
H3 / Focus	Die Nachhaltigkeitsartikel im *Focus* sind länger als andere Artikel.
H3 / taz	Auf Basis der Stichprobe kann nicht signifikant bestätigt werden, dass die Nachhaltigkeitsartikel in der *taz* länger sind als andere Artikel.
H4	Nachhaltigkeitsartikel sind sowohl in der *taz* als auch im *Focus* nicht negativ geprägt.
H5 / Focus	Die Berichterstattung zu Nachhaltigkeitsartikeln ist im Vergleich zur restlichen Berichterstattung im *Focus* eher subjektiv geprägt.
H5 / taz	Auf Basis der Stichprobe kann nicht signifikant bestätigt werden, dass die Nachhaltigkeitsartikel in der *taz* subjektiver geprägt sind als die restlichen Artikel.

Tabelle 22: Ergebnis der Hypothesenüberprüfung

Hypothese H1 und H2 können bestätigt werden. Hypothese H3 und H5 sind lediglich für die Stichprobe des Focus zu bestätigen. Hypothese H4 ist für die Stichprobe der taz und des Focus zu verwerfen. Die eingangs aufgestellte Forschungsfrage lautet:

> *Welche Präsenz hat das Thema Nachhaltigkeit in der Berichterstattung von taz und Focus und wie ausgeglichen ist diese bezüglich der einzelnen Nachhaltigkeitsziele?*

Die Bestätigung der Hypothese H1 passt zu den Analyseergebnissen des *Basel Institute of Commons and Economics*. Die ungleiche Berücksichtigung der Nachhaltigkeitsziele findet sich auch in der unausgeglichenen Berichterstattung des Focus und der taz wieder. Allerdings muss eingeordnet werden, dass in dem vorliegenden Untersuchungszeitraum die Themen Gesundheit und Wohlergehen oder Nachhaltige Städte und Gemeinden durch andere Faktoren eine besondere Medienpräsenz erlangt haben. Ein herausstechender Unterschied zwischen den Untersuchungsergebnissen des BICE und der Analyse der Berichterstattung ist die Einordnung des Zieles „Ozean, Meere, Meeresressourcen“. Dieser Unterschied könnte auf die durch das NLP ermittelten ähnlichen Begriffe (Suchworte) zurückgeführt werden. In diesem konkreten Fall überschneiden sich die Suchwörter thematisch mit anderen Inhalten als mit Nachhaltigkeit; hier beispielsweise deutsche Küste und Urlaub. An dieser Stelle sind weitere Untersuchungen notwendig.

Durch die Bestätigung der Hypothese H2 wird die Vermutung unterstützt, dass die taz auf Grund ihrer politischen Ausrichtung und des redaktionellen Konzepts eine reflektierte Berichterstattung in Bezug auf das Thema Nachhaltigkeit liefert. Die Berichterstattung der taz zu den einzelnen Zielen ist nicht gleichverteilt. Jedoch ist die Berichterstattung des Focus unausgeglichener und weist im Vergleich zur taz eine stärkere Konzentration auf einzelne Schwerpunkte auf. Die verwendete Untersuchungsmethode ermöglicht keine quantifizierte Aussage, um welchen Faktor sich die Zeitungen bezüglich einer Gleichverteilung unterscheiden. In Bezug der Untersuchung dieser Hypothese ist die verfälschende Wirkung der überdurchschnittlichen Berichterstattungen zu den globalen Ereignissen zu berücksichtigen. Durch die größere Anzahl der Veröffentlichungen im Focus könnte dies einen überproportionalen Effekt haben.

Die Hypothese H3 wird für beide Zeitungen getrennt untersucht. Für die taz ist kein signifikanter Unterschied der Textlänge zwischen Artikel zum Thema Nachhaltigkeit und anderen Artikeln zu erkennen. Der Focus hingegen weist einen signifikanten Unterschied hinsichtlich der Textlänge auf. Die Analyse der Metadaten zeigt, dass die Textlänge des Focus deutlich stärker variiert als die der taz. Dies kann unteranderem durch einen anderen Anspruch an die Berichterstattung begründet werden. Die taz veröffentlich weniger Kurzmeldungen als der Focus. Dies wird durch die durchschnittlich größere Textlänge belegt. Artikel zum Thema Nachhaltigkeit sind typischerweise keine Kurzmeldungen, wodurch besonders im Focus der signifikante Unterschied zwischen den Textlängen entsteht. Auch in der taz ist ein Unterschied erkennbar, jedoch reicht dieser nicht aus, um einen signifikanten Unterschied zu belegen. Anhand der vorliegenden Ergebnisse wird vermutetet, dass das Thema Nachhaltigkeit in der taz eher Inhalt der täglichen Berichterstattung ist, wohingegen es sich im Focus von der typischen Berichterstattung abhebt; beispielsweise durch Sonderartikel oder Themenhefte.

Der Hypothese H4 geht die Vermutung voraus, dass Nachhaltigkeitsthemen mit Problemen und Missständen verbunden werden. Demnach wird in der Hypothese überprüft, ob die Berichterstattung negativ geprägt ist. Die Tests der Hypothesen zeigen, dass die Berichterstattung zum Thema Nachhaltigkeit nicht negativ ist. Allerding ist die Berichterstattung der Zeitungen insgesamt neutral bis leicht positiv. Somit zeigt sich, dass auch wenn Berichte zur Nachhaltigkeit auf Probleme oder Missstände hinweisen, sie im Stil der Zeitungen größtenteils wertungsfrei berichtet werden.

Mit der abschließenden Hypothese wird untersucht, inwiefern Nachhaltigkeitsthemen subjektiv berichtet werden. Aufgrund der Vermutung, dass die Berichterstattung einer Zeitung eher objektiv ist, wird überprüft, ob ein Unterschied zwischen Nachhaltigkeitsthemen und den restlichen Berichten besteht. Für die taz kann kein signifikanter Unterschied nachgewiesen werden. Innerhalb der Berichterstattung des Focus werden Nachhaltigkeitsthemen subjektiver dargestellt als die übrigen Themen. Zeitungs- und themenübergreifend ist die Berichterstattung jedoch sehr objektiv. Hierbei ist der Unterschied möglicherweise durch das redaktionelle Konzept zu begründen.

Die Forschungsfrage lässt sich zusammenfassend wie folgt beantworten: In der taz scheint das Thema Nachhaltigkeit eher Teil der allgemeinen Berichterstattung zu sein. Es wird vermutet, dass Nachhaltigkeitsthemen im Focus auf Grund der Textlänge nicht Teil der gewöhnlichen Berichterstattung sind. Artikel zur Nachhaltigkeit zeichnen sich nicht durch eine negative, sondern eher durch eine neutrale Wortwahl aus. Dabei ist die Berichterstattung insgesamt sehr objektiv. Im Focus ist ein leichter Trend zur Subjektivität erkennbar. Zu den 17 Nachhaltigkeitszielen der UN wird in beiden Zeitungen unausgeglichen berichtet. Diese Beobachtung ist beim Focus ausgeprägter.

6 Fazit

In der vorliegenden Arbeit wurde die Berichterstattung zum Thema Nachhaltigkeit in den Onlinezeitungen taz und Focus analysiert. Aus der Recherche zu diesem Thema geht hervor, dass der Nachhaltigkeitsbegriff aus den drei gleichwertigen Komponenten – Wirtschaft, Umwelt und Soziales – besteht. Basierend darauf haben Die Vereinten Nationen 17 gleichwertige Nachhaltigkeitsziele formuliert, die auf eine weltweite nachhaltige Entwicklung abzielen. Laut einer Analyse des *Basel Institute of Commons and Economics* werden diese Ziele jedoch ungleich berücksichtigt. Aus diesem Grund wurde der folgenden Forschungsfrage im Zuge der Arbeit nachgegangen:

> *Welche Präsenz hat das Thema Nachhaltigkeit in der Berichterstattung von taz und Focus und wie ausgeglichen ist diese bezüglich der einzelnen Nachhaltigkeitsziel?*

Zur Beantwortung der Frage wurde ein quantitativer Analyseansatz gewählt. Die HTML-Quellcodes beider Zeitungen wurden über einen Zeitraum von zwei Monaten in einer Datenbank gespeichert. Durch die Datenextraktion wurden relevante Merkmale wie beispielsweise Autor, Ressort und Veröffentlichungsdatum aus den HTML-Quellcodes herausgefiltert. Diese konnten mittels statistischer Methoden und Big Data Techniken analysiert werden. Außerdem wurden die Inhalte der Zeitungstexte unter Zuhilfenahme von Natural Language Processing semantisch untersucht. Es konnte dadurch u. a. das Subjektivitätsniveau und die Stimmung innerhalb der Artikel bestimmt werden. Die Datenerhebung sowie die Datenanalyse wurden als Softwareprojekt in Python implementiert.

Zur Beantwortung der Forschungsfrage wurden fünf Hypothesen abgeleitet, welche mittels deskriptiver statistischer Methoden und Hypothesentests untersucht wurden. Die Untersuchungsergebnisse zeigen, dass die Berichterstattung zu den einzelnen Nachhaltigkeitszielen in beiden Zeitungen unausgeglichen ist. Artikel zum Thema Nachhaltigkeit sind im Focus im Mittel länger als andere Texte. In der taz besteht diesbezüglich kein Unterschied. Beide Zeitungen berichten über Nachhaltigkeit neutral und objektiv; der Focus hat dabei einen Hang zur Subjektivität. Das redaktionelle Konzept sowie die politische und inhaltliche Ausrichtung beider Zeitungen wurden durch die Analysen deutlich.

Literaturverzeichnis

[1] taz, die tageszeitung, „Wir über uns“, *taz Info*, 2020. https://taz.de/Zahlen-und-Fakten/!106557/ (zugegriffen Jan. 04, 2021).

[2] Brockhaus, „die tageszeitung“, *Brockhaus Enzyklopädie Online*, ohne Datum. https://brockhaus.de/ecs/enzy/article/die-tageszeitung (zugegriffen Jan. 04, 2021).

[3] taz, die tageszeitung, „Redaktionsstatut“, *taz Hilfe*, 2008. https://taz.de/!114802/ (zugegriffen Jan. 04, 2021).

[4] Brockhaus, „Focus“, *Brockhaus Enzyklopädie Online*, ohne Datum. https://brockhaus.de/ecs/enzy/article/focus (zugegriffen Jan. 04, 2021).

[5] Statista, „Statista-Dossier zum Nachrichten-Magazin Focus“, *Statista - das Statistik Portal*, 2020. https://de.statista.com/statistik/studie/id/24141/dokument/focus-statista-dossier/ (zugegriffen Jan. 04, 2021).

[6] B. Kaltenhäuser, *Abstimmung am Kiosk: Der Einfluss der Titelseitengestaltung politischer Publikumszeitschriften auf die Einzelverkaufsauflage*, 1. Auflage. Wiesbaden: Deutscher Universitäts-Verlag / GWV Fachverlage, 2005.

[7] Jürgen Scharrer, „Focus traut sich mehr als gedacht“, *HORIZONT*, Nr. 04, 2010.

[8] Iris Pufé, *Nachhaltigkeit*, 3. München: UKV Verlagsgesellschaft, 2017.

[9] A. Kleine, *Operationalisierung einer Nachhaltigkeitsstrategie, Ökologie, Ökonomie und Soziales integrieren*. Wiesbaden: Gabler, 2009.

[10] Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung (BMZ), „Agenda 2030. 17 Ziele für nachhaltige Entwicklung“, *Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung (BMZ)*, ohne Datum. https://www.bmz.de/de/themen/2030_agenda/17_ziele/index.html (zugegriffen Jan. 05, 2021).

[11] Vereinte Nationen, „Sustainable Development Goals“, *Vereinte Nationen Development Programme*, ohne Datum. https://www.undp.org/content/undp/en/home/sustainable-development-goals.html (zugegriffen Jan. 05, 2021).

[12] E. Uzun, „A Novel Web Scraping Approach Using the Additional Information Obtained From Web Pages“, *IEEE Access*, Nr. 8, S. 61726–61740, 2020.

[13] L. Richardson, „Beautiful Soup Documentation“, *Crummy: The Site*, 2020. https://www.crummy.com/software/BeautifulSoup/bs4/doc/ (zugegriffen Jan. 19, 2021).

[14] T. Becker, *Prozesse in Produktion und Supply Chain optimieren*, 3. Berlin Heidelberg: Springer Vieweg, 2018.

[15] L. Bosankic, „Natural Language Processing für Topic Modeling in Python“, *solvistas*, 2019. https://www.solvistas.com/blog/python-nlp-pipeline-fuer-die-extraktion-von-themen-aus-nachrichten/ (zugegriffen Jan. 19, 2021).

[16] T. Stucki, S. D'Onofrio, und E. Portmann, *Chatbots gestalten mit Praxisbeispielen der Schweizerischen Post*. Wiesbaden: Springer Vieweg, 2020.

[17] A. Klofak, „Wie funktioniert Natural Language Processing in der Praxis? Ein Überblick“, *Data Sience Blog*, 2020. https://data-science-blog.com/blog/2020/01/24/wie-funktioniert-natural-language-processing-in-der-praxis-ein-uberblick-dr-aleksandra-klofat/ (zugegriffen Jan. 05, 2020).

[18] R. Becker, „Word Embeddings – Methoden zur Repräsentation von Wörtern in Algorithmen und Neuronalen Netzen“, *JAAI*, 2019. https://jaai.de/word-embeddings-worteinbettung-word2vec-glove-bag-of-words-1872/#tab-id-2 (zugegriffen Jan. 13, 2021).

[19] H. Jabeen, „Stemming and Lemmatization in Python“, *datacamp.com*, 2018. https://www.datacamp.com/community/tutorials/stemming-lemmatization-python (zugegriffen Jan. 19, 2021).

[20] R. Řehůřek, „Core Concepts“, *Gensim, topic modelling for humans*, 2020. https://radimrehurek.com/gensim/auto_examples/core/run_core_concepts.html#sphx-glr-auto-examples-core-run-core-concepts-py (zugegriffen Jan. 14, 2021).

[21] S. Prabhakaran, „Gensim Tutorial – A Complete Beginners Guide“, *Machine Learning Plus*, ohne Datum. https://www.machinelearningplus.com/nlp/gensim-tutorial/ (zugegriffen Jan. 15, 2021).

[22] B. Stecanella, „What is TF-IDF?“, *MonkeyLearn*, Mai 10, 2019. https://monkeylearn.com/blog/what-is-tf-idf/ (zugegriffen Jan. 15, 2021).

[23] NLTK Project, „Natural Language Toolkit“, *nltk.org*, 2020. https://www.nltk.org (zugegriffen Jan. 17, 2021).

[24] M. Usman, „Removing Stop Words from Strings in Python“, *Stack Abuse*, ohne Datum. https://stackabuse.com/removing-stop-words-from-strings-in-python/, (zugegriffen Jan. 17, 2021).

[25] S. Parthvi, „Takeaways from My First Data Science Internship“, *Towards Data Science*, 2020. https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524 (zugegriffen Jan. 15, 2021).

[26] Deutsches Institut für Normung (DIN), *DIN 69901-5: Projektmanagement - Projektmanagementsysteme - Teil 5: Begriffe*. Berlin: Beuth, 2009.

[27] H. Meyer und H.-J. Reher, *Projektmanagement: Von der Definition über die Projektplanung zum erfolgreichen Abschluss*, 2. Wiesbaden: Springer Fachmedien Wiesbaden, 2020.

[28] Basel Institute of Commons and Economics, „World Social Capital Monitor“. ohne Datum, Zugegriffen: Jan. 17, 2021. [Online]. Verfügbar unter: http://commons.ch/wp-content/uploads/Synopsis_SDG_Reports_Goals_Allocation_2019.pdf.

[29] L. Fahrmeir, C. Heumann, R. Künstler, I. Pigeot, und G. Tutz, *Statistik: Der Weg zur Datenanalyse*. Berlin, Heidelberg: Springer Spektrum, 2016.

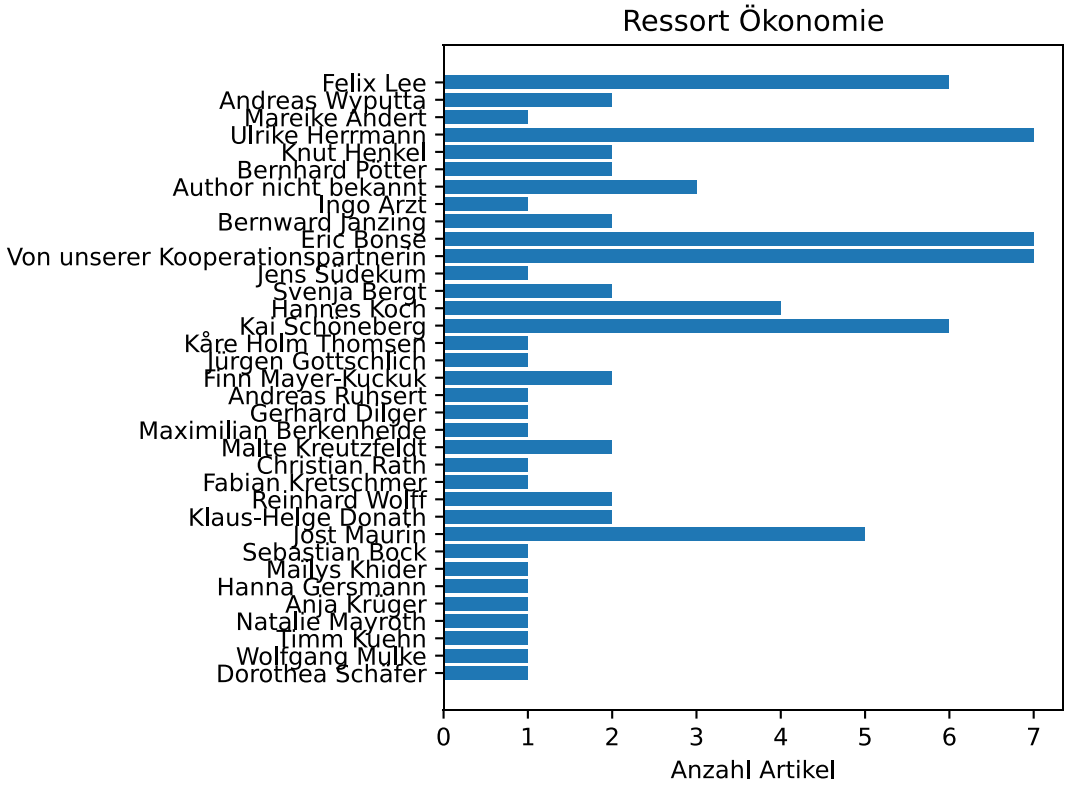
[30] Presse- und Informationsamt der deutschen Bundesregierung, „Die UN-Nachhaltigkeitsziele“, *Die Bundesregierung*, 2021. https://www.bundesregierung.de/breg-de/themen/nachhaltigkeitspolitik/die-un-nachhaltigkeitsziele-1553514 (zugegriffen Jan. 18, 2021).

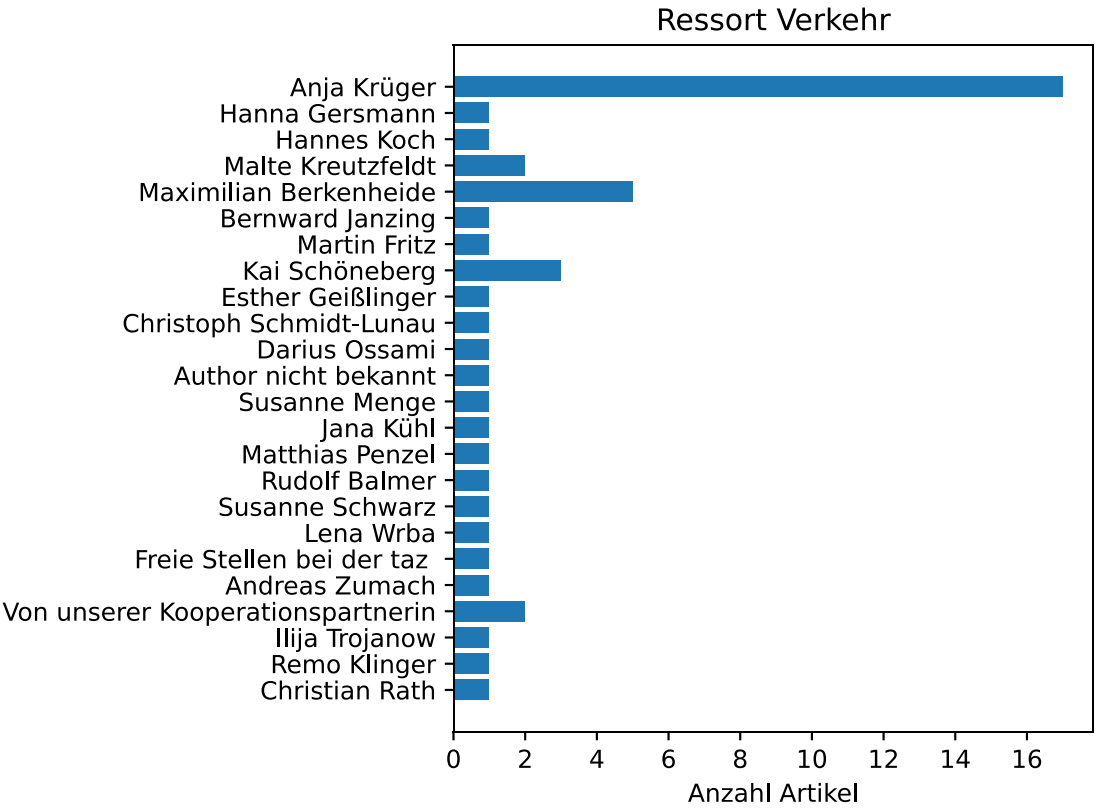
Anhang

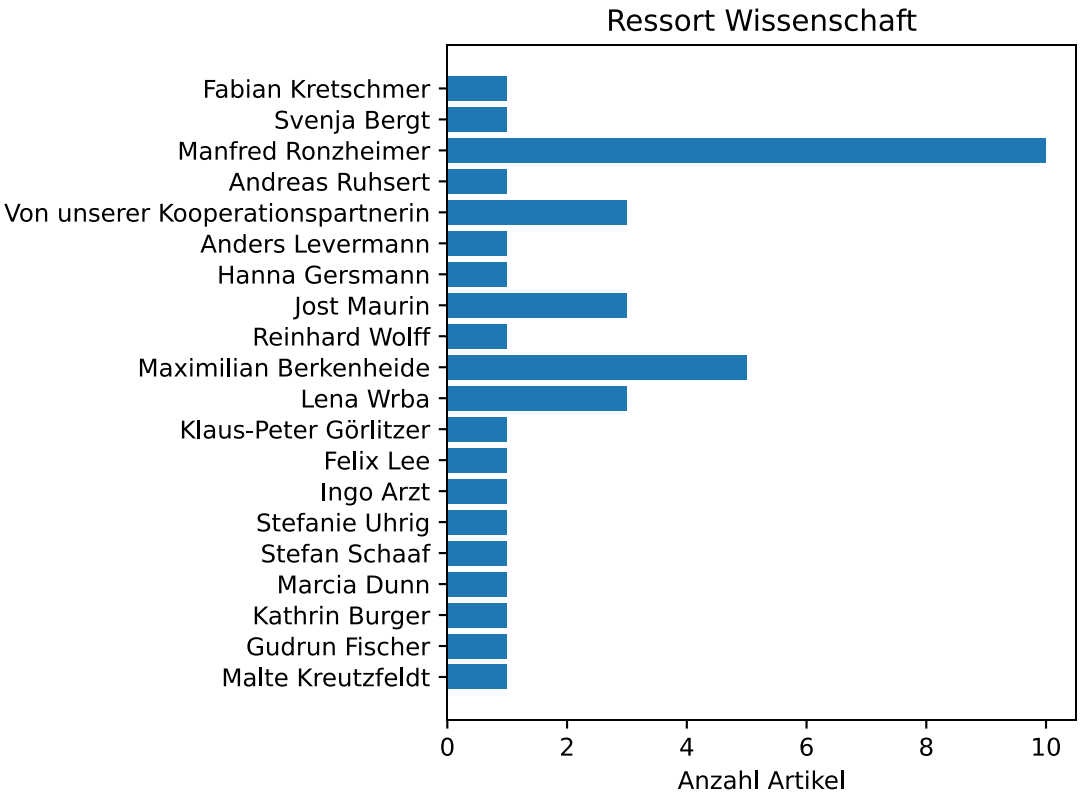
Anhang A: Ressorts der taz [1]

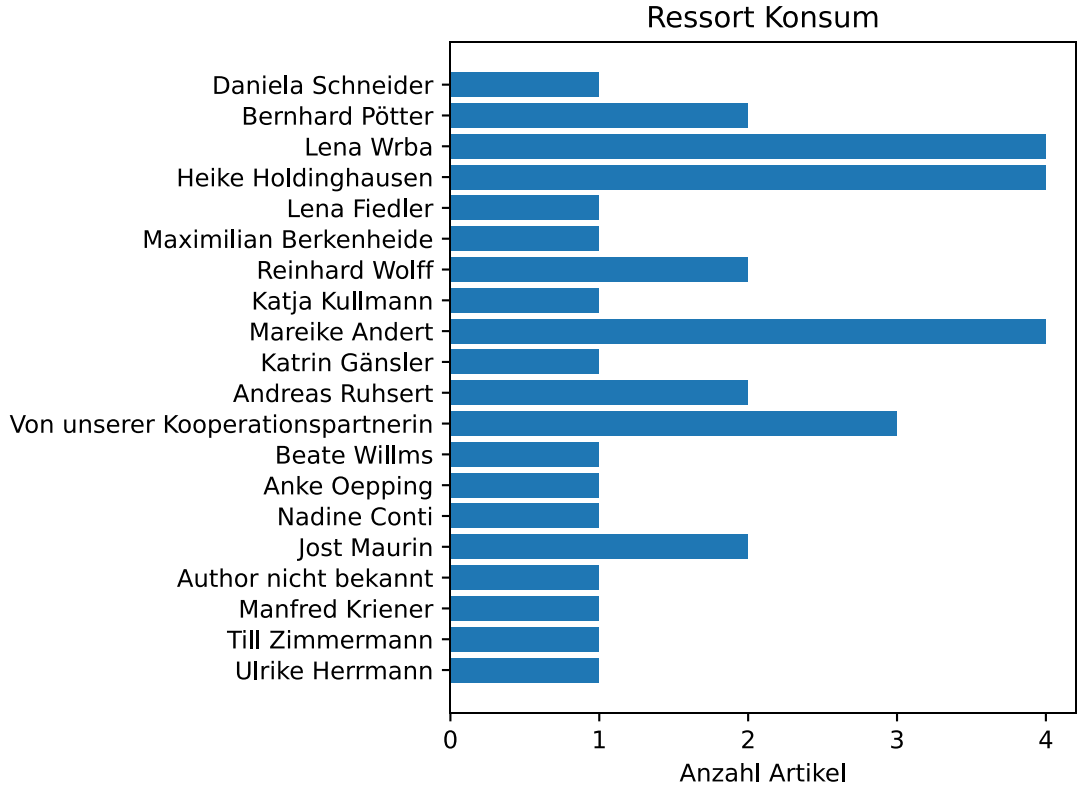
Hauptresort	Ressort
Politik	• Deutschland • Europa • Amerika • Afrika • Asien • Nahost • Netzpolitik
Öko	• Ökonomie • Ökologie • Arbeit • Konsum • Verkehr • Wissenschaft • Netzökonomie
Gesellschaft	• Alltag • Reprotage und Recherche • Debatte • Kolumnen • Medien • Bildung • Gesundheit • Reise
Kultur	• Musik • Film • Künste • Buch • Netzkultur
Sport	• Fußball • Kolumnen
Berlin	• Ohne Ressort
Nord	• Hamburg • Bremen • Kultur
Wahrheit	• Bei Tom • Über die Wahrheit

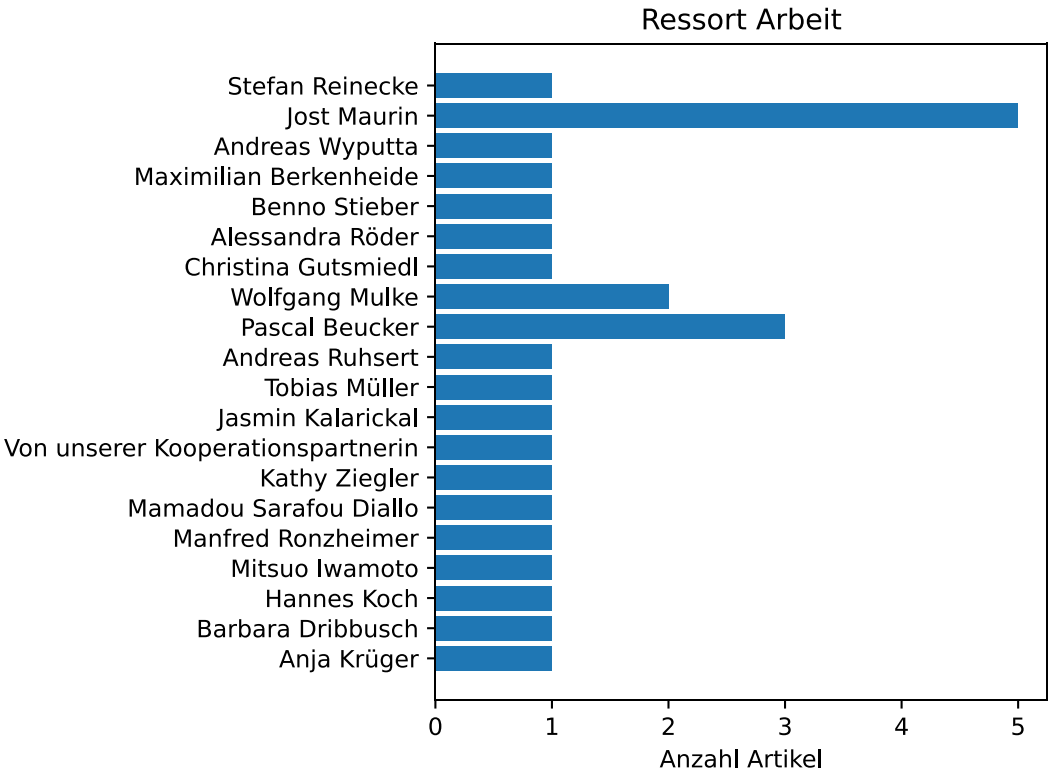
Anhang B: Artikel pro Autoren je Ressort (Öko)






Ressort Wissenschaft
Fabian Kretschmer
Svenja Bergt
Manfred Ronzheimer
Andreas Ruhsert
Von unserer Kooperationspartnerin
Anders Levermann
Hanna Gersmann
Jost Maurin
Reinhard Wolff
Maximilian Berkenheide
Lena Wrba
Klaus-Peter Görlitzer
Felix Lee
Ingo Arzt
Stefanie Uhrig
Stefan Schaaf
Marcia Dunn
Kathrin Burger
Gudrun Fischer
Malte Kreutzfeldt
0
2
4
6
8
10
Anzahl Artikel


Ressort Konsum
Daniela Schneider
Bernhard Pötter
Lena Wrba
Heike Holdinghausen
Lena Fiedler
Maximilian Berkenheide
Reinhard Wolff
Katja Kullmann
Mareike Andert
Katrin Gänsler
Andreas Ruhsert
Von unserer Kooperationspartnerin
Beate Willms
Anke Oepping
Nadine Conti
Jost Maurin
Author nicht bekannt
Manfred Kriener
Till Zimmermann
Ulrike Herrmann
0
1
2
3
4
Anzahl Artikel


Ressort Arbeit
Stefan Reinecke
Jost Maurin
Andreas Wyputta
Maximilian Berkenheide
Benno Stieber
Alessandra Röder
Christina Gutsmiedl
Wolfgang Mulke
Pascal Beucker
Andreas Ruhsert
Tobias Müller
Jasmin Kalarickal
Von unserer Kooperationspartnerin
Kathy Ziegler
Mamadou Sarafou Diallo
Manfred Ronzheimer
Mitsuo Iwamoto
Hannes Koch
Barbara Dribbusch
Anja Krüger
0
1
2
3
4
5
Anzahl Artikel

Anhang C: Auswertung NLP zu Nachhaltigkeitszielen

Nachhaltigkeitsziel	Schlagwort	Ähnliche Wörter (NLP)
Gesundheit und Wohlergehen	Gesundheit	Wohlbefinden, Ernaehrung, Ernährung, Umwelt, psychische Gesundheit, Bildung, gesunde Lebensweise, Nahrungsmittelsicherheit
Nachhaltige Städte und Gemeinden	Stadt	Kommune, Gemeinde, Kreisstadt, Stadtverwaltung, Nachbarstadt, Städte, Kreises, Landkreis, Landeshauptstadt
Ozeane und Meere	Meer	Ozean, Strand, See, Bucht, Nordsee, Kueste, Atlantik, Mittelmeer, Brandung, Ostsee
Industrie, Innovation und Infrastruktur	Infrastruktur	Verkehrsinfrastruktur, Infrastrukturen, marode Infrastruktur, Verkehrswege, Verkehrs-Infrastruktur, Investitionen Infrastruktur, Kitas, Ganztagsschulen, Gesundheitsversorgung, Strassen, Straßen, Brücken, Investitionen
Maßnahmen zum Klimaschutz	Klimaschutz	Umweltschutz, Umwelt, Klimaschutz, Energiesparen, erneuerbare Energien, Klima, Umweltpolitik, Energiepolitik, Energieeffizienz, alternative Energien
Sauberes Wasser und Sanitäreinrichtungen	Trinkwasser	Leitungswasser, Abwasser, Frischwasser, Trinkwasserversorgung, Grundwasser, gechlort, sauberes Wasser, Fremdwasser, Wasser
Geschlechter Gleichheit	Gleichstellung	Gleichberechtigung, Oeffnung Ehe, Öffnung Ehe, volle Gleichstellung, Lebenspartnerschaften, Ehe, Adoptionsrecht, homosexuelle Partnerschaft, Chancengleichheit, Gleichbehandlung, gleiche Bezahlung
Hochwertige Bildung	Bildung	leistbares Wohnen, bezahlbare Mieten, soziale Teilhabe, fruehkindliche Bildung, frühkindliche Bildung, Soziales, Integration, Bildungsgerechtigkeit, Teilhabe, leistbares
Frieden, Gerechtigkeit und starke Institutionen	Frieden	Friede, Weltfrieden, Frieden Palaestinensern, Versöhnung, Gerechtigkeit, dauerhafter Frieden, geeintes Europa, inneren Frieden
Keine Armut	Armut	Arbeitslosigkeit, Hunger, Ungleichheit, Elend, soziale Ausgrenzung, Ungerechtigkeit, soziale Ungleichheit, Analphabetismus, soziale Ungerechtigkeit,

Menschenwürdige Arbeit und Wirtschaftswachstum	Wirtschaftswachstum	Wachstum, Binnennachfrage, Wirtschaftsentwicklung, Exportwachstum, Kreditwachstum, BIP-Wachstum, privaten_Konsum, Bruttoinlandsprodukt, zweitgroessten Volkswirtschaft
Kein Hunger	Hunger	Durst, Hungers, Heisshunger, stillt, Hunger leiden, Armut, hungert
Bezahlbare und saubere Energie	Energiewende	Energiepolitik, Umsetzung Energiewende, erneuerbare Energien, Netzausbau, Klimaschutz, erneuerbar, Atomausstieg, ökologische Modernisierung
Ökosysteme an Land	Lebensraum	Lebensräume, Artenvielfalt, Pflanzen, Tierarten, Artenreichtum, idealen Lebensraum, Pflanzenarten, seltene Arten, Biotope, Wiesenvoegel
Weniger Ungleichheiten	Ungleichheit	Ungleichheiten, Einkommensungleichheit, soziale Ungleichheit, Einkommensunterschiede, Armut, soziale Spaltung, wachsende Kluft, Spaltung Gesellschaft, Arm Reich, Einkommensschere
Nachhaltiger Konsum und Produktion	Konsumverhalten	Kaufverhalten, Konsumgewohnheiten, Einkaufsverhalten, Verbraucherverhalten, Freizeitverhalten, Mobilitätsverhalten, Gesundheitsbewusstsein, Ernährungsgewohnheiten, Reiseverhalten, Trinkverhalten
Nachhaltigkeit	Nachhaltigkeit	Nachhaltiges Wirtschaften, Nachhaltigkeit, ökologische Nachhaltigkeit, Regionalität, Umweltschutz, Energieeffizienz, regionale Wertschöpfung, Ressourcenschonung, fairer Handel, Ressourceneffizienz, nachhaltiges Handeln, Nachhaltigkeit

Anhang D: t-Verteilung

Tabelliert sind die Quantile für n Freiheitsgrade. Für das Quantil $t_{1-\alpha}(n)$ gilt $F(t_{1-\alpha}(n)) = 1 - \alpha$.
Links vom Quantil $t_{1-\alpha}(n)$ liegt die Wahrscheinlichkeitsmasse $1 - \alpha$.

Ablesebeispiel: $t_{0.99}(20) = 2.528$

Die Quantile für $0 < 1 - \alpha < 0.5$ erhält man aus $t_\alpha(n) = -t_{1-\alpha}(n)$.

Approximation für $n > 30$: $t_\alpha(n) \approx z_\alpha$ (z_α ist das (α)-Quantil der Standardnormalverteilung)

n	0.6	0.8	0.9	0.95	0.975	0.99	0.995	0.999	0.9995
1	0.3249	1.3764	3.0777	6.3138	12.706	31.821	63.657	318.31	636.62
2	0.2887	1.0607	1.8856	2.9200	4.3027	6.9646	9.9248	22.327	31.599
3	0.2767	0.9785	1.6377	2.3534	3.1824	4.5407	5.8409	10.215	12.924
4	0.2707	0.9410	1.5332	2.1318	2.7764	3.7469	4.6041	7.1732	8.6103
5	0.2672	0.9195	1.4759	2.0150	2.5706	3.3649	4.0321	5.8934	6.8688
6	0.2648	0.9057	1.4398	1.9432	2.4469	3.1427	3.7074	5.2076	5.9588
7	0.2632	0.8960	1.4149	1.8946	2.3646	2.9980	3.4995	4.7853	5.4079
8	0.2619	0.8889	1.3968	1.8595	2.3060	2.8965	3.3554	4.5008	5.0413
9	0.2610	0.8834	1.3830	1.8331	2.2622	2.8214	3.2498	4.2968	4.7809
10	0.2602	0.8791	1.3722	1.8125	2.2281	2.7638	3.1693	4.1437	4.5869
11	0.2596	0.8755	1.3634	1.7959	2.2010	2.7181	3.1058	4.0247	4.4370
12	0.2590	0.8726	1.3562	1.7823	2.1788	2.6810	3.0545	3.9296	4.3178
13	0.2586	0.8702	1.3502	1.7709	2.1604	2.6503	3.0123	3.8520	4.2208
14	0.2582	0.8681	1.3450	1.7613	2.1448	2.6245	2.9768	3.7874	4.1405
15	0.2579	0.8662	1.3406	1.7531	2.1314	2.6025	2.9467	3.7328	4.0728
16	0.2576	0.8647	1.3368	1.7459	2.1199	2.5835	2.9208	3.6862	4.0150
17	0.2573	0.8633	1.3334	1.7396	2.1098	2.5669	2.8982	3.6458	3.9651
18	0.2571	0.8620	1.3304	1.7341	2.1009	2.5524	2.8784	3.6105	3.9216
19	0.2569	0.8610	1.3277	1.7291	2.0930	2.5395	2.8609	3.5794	3.8834
20	0.2567	0.8600	1.3253	1.7247	2.0860	2.5280	2.8453	3.5518	3.8495
21	0.2566	0.8591	1.3232	1.7207	2.0796	2.5176	2.8314	3.5272	3.8193
22	0.2564	0.8583	1.3212	1.7171	2.0739	2.5083	2.8188	3.5050	3.7921
23	0.2563	0.8575	1.3195	1.7139	2.0687	2.4999	2.8073	3.4850	3.7676
24	0.2562	0.8569	1.3178	1.7109	2.0639	2.4922	2.7969	3.4668	3.7454
25	0.2561	0.8562	1.3163	1.7081	2.0595	2.4851	2.7874	3.4502	3.7251
26	0.2560	0.8557	1.3150	1.7056	2.0555	2.4786	2.7787	3.4350	3.7066
27	0.2559	0.8551	1.3137	1.7033	2.0518	2.4727	2.7707	3.4210	3.6896
28	0.2558	0.8546	1.3125	1.7011	2.0484	2.4671	2.7633	3.4082	3.6739
29	0.2557	0.8542	1.3114	1.6991	2.0452	2.4620	2.7564	3.3962	3.6594
30	0.2556	0.8538	1.3104	1.6973	2.0423	2.4573	2.7500	3.3852	3.6460
∞	0.2533	0.8416	1.2816	1.6449	1.9600	2.3263	2.5758	3.0903	3.2906

Quelle: Fahrmeier et al., S.557

Anhang E: F-Verteilung

Tabelliert sind die rechtsseitigen Quantile für (n_1, n_2) Freiheitsgrade.
Für das Quantil $f_{1-\alpha}(n_1, n_2)$ gilt $F(f_{1-\alpha}(n_1, n_2)) = 1 - \alpha$. Links vom Quantil $f_{1-\alpha}(n_1, n_2)$ liegt die Wahrscheinlichkeitsmasse $1 - \alpha$.
Ablesebeispiel: $f_{0.99}(15, 8) = 5.5151$
Linksseitige Quantile: $f_\alpha(n_1, n_2) = \frac{1}{f_{1-\alpha}(n_1, n_2)}$

		n_2								
n_1	α	1	2	3	4	5	6	7	8	9
1	0.9	39.863	8.5263	5.5383	4.5448	4.0604	3.7759	3.5894	3.4579	3.3603
	0.95	161.45	18.513	10.128	7.7086	6.6079	5.9874	5.5914	5.3177	5.1174
	0.975	647.79	38.506	17.443	12.218	10.007	8.8131	8.0727	7.5709	7.2093
	0.99	4052.2	98.502	34.116	21.198	16.258	13.745	12.246	11.259	10.561
2	0.9	49.500	9.0000	5.4624	4.3246	3.7797	3.4633	3.2574	3.1131	3.0065
	0.95	199.50	19.000	9.5521	6.9443	5.7861	5.1433	4.7374	4.4590	4.2565
	0.975	799.50	39.000	16.044	10.649	8.4336	7.2599	6.5415	6.0595	5.7147
	0.99	4999.5	99.000	30.817	18.000	13.274	10.925	9.5466	8.6491	8.0215
3	0.9	53.593	9.1618	5.3908	4.1909	3.6195	3.2888	3.0741	2.9238	2.8129
	0.95	215.71	19.164	9.2766	6.5914	5.4095	4.7571	4.3468	4.0662	3.8625
	0.975	864.16	39.165	15.439	9.9792	7.7636	6.5988	5.8898	5.4160	5.0781
	0.99	5403.4	99.166	29.457	16.694	12.060	9.7795	8.4513	7.5910	6.9919
4	0.9	55.833	9.2434	5.3426	4.1072	3.5202	3.1808	2.9605	2.8064	2.6927
	0.95	224.58	19.247	9.1172	6.3882	5.1922	4.5337	4.1203	3.8379	3.6331
	0.975	899.58	39.248	15.101	9.6045	7.3879	6.2272	5.5226	5.0526	4.7181
	0.99	5624.6	99.249	28.710	15.977	11.392	9.1483	7.8466	7.0061	6.4221
5	0.9	57.240	9.2926	5.3092	4.0506	3.4530	3.1075	2.8833	2.7264	2.6106
	0.95	230.16	19.296	9.0135	6.2561	5.0503	4.3874	3.9715	3.6875	3.4817
	0.975	921.85	39.298	14.885	9.3645	7.1464	5.9876	5.2852	4.8173	4.4844
	0.99	5763.6	99.299	28.237	15.522	10.967	8.7459	7.4604	6.6318	6.0569
6	0.9	58.204	9.3255	5.2847	4.0097	3.4045	3.0546	2.8274	2.6683	2.5509
	0.95	233.99	19.330	8.9406	6.1631	4.9503	4.2839	3.8660	3.5806	3.3738
	0.975	937.11	39.331	14.735	9.1973	6.9777	5.8198	5.1186	4.6517	4.3197
	0.99	5859.0	99.333	27.911	15.207	10.672	8.4661	7.1914	6.3707	5.8018
7	0.9	58.906	9.3491	5.2662	3.9790	3.3679	3.0145	2.7849	2.6241	2.5053
	0.95	236.77	19.353	8.8867	6.0942	4.8759	4.2067	3.7870	3.5005	3.2927
	0.975	948.22	39.355	14.624	9.0741	6.8531	5.6955	4.9949	4.5286	4.1970
	0.99	5928.4	99.356	27.672	14.976	10.456	8.2600	6.9928	6.1776	5.6129
8	0.9	59.439	9.3668	5.2517	3.9549	3.3393	2.9830	2.7516	2.5893	2.4694
	0.95	238.88	19.371	8.8452	6.0410	4.8183	4.1468	3.7257	3.4381	3.2296
	0.975	956.66	39.373	14.540	8.9796	6.7572	5.5996	4.8993	4.4333	4.1020
	0.99	5981.1	99.374	27.489	14.799	10.289	8.1017	6.8400	6.0289	5.4671
9	0.9	59.858	9.3805	5.2400	3.9357	3.3163	2.9577	2.7247	2.5612	2.4403
	0.95	240.54	19.385	8.8123	5.9988	4.7725	4.0990	3.6767	3.3881	3.1789
	0.975	963.28	39.387	14.473	8.9047	6.6811	5.5234	4.8232	4.3572	4.0260
	0.99	6022.5	99.388	27.345	14.659	10.158	7.9761	6.7188	5.9106	5.3511
10	0.9	60.195	9.3916	5.2304	3.9199	3.2974	2.9369	2.7025	2.5380	2.4163
	0.95	241.88	19.396	8.7855	5.9644	4.7351	4.0600	3.6365	3.3472	3.1373
	0.975	968.63	39.398	14.419	8.8439	6.6192	5.4613	4.7611	4.2951	3.9639
	0.99	6055.8	99.399	27.229	14.546	10.051	7.8741	6.6201	5.8143	5.2565
11	0.9	60.473	9.4006	5.2224	3.9067	3.2816	2.9195	2.6839	2.5186	2.3961
	0.95	242.98	19.405	8.7633	5.9358	4.7040	4.0274	3.6030	3.3130	3.1025
	0.975	973.03	39.407	14.374	8.7935	6.5678	5.4098	4.7095	4.2434	3.9121
	0.99	6083.3	99.408	27.133	14.452	9.9626	7.7896	6.5382	5.7343	5.1779
12	0.9	60.705	9.4081	5.2156	3.8955	3.2682	2.9047	2.6681	2.5020	2.3789
	0.95	243.91	19.413	8.7446	5.9117	4.6777	3.9999	3.5747	3.2839	3.0729
	0.975	976.71	39.415	14.337	8.7512	6.5245	5.3662	4.6658	4.1997	3.8682
	0.99	6106.3	99.416	27.052	14.374	9.8883	7.7183	6.4691	5.6667	5.1114
13	0.9	60.903	9.4145	5.2098	3.8859	3.2567	2.8920	2.6545	2.4876	2.3640
	0.95	244.69	19.419	8.7287	5.8911	4.6552	3.9764	3.5503	3.2590	3.0475
	0.975	979.84	39.421	14.304	8.7150	6.4876	5.3290	4.6285	4.1622	3.8306

Quelle: Fahrmeier et al., S. 558

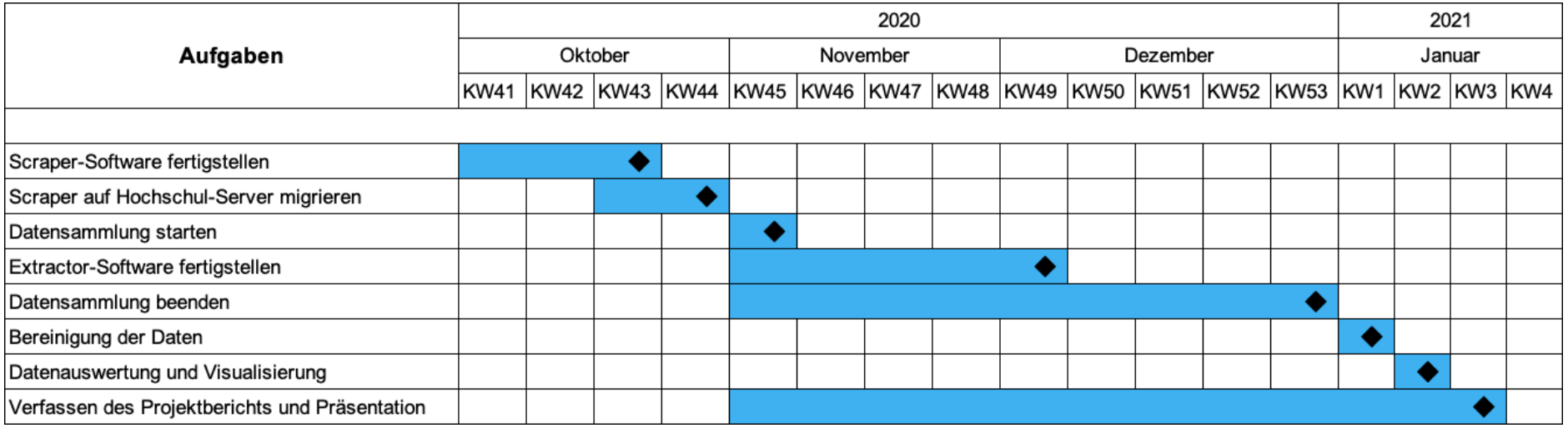
Anhang F: Chi-Quadrat-Verteilung

Tabelliert sind die Quantile für n Freiheitsgrade. Für das Quantil $\chi^2_{1-\alpha}(n)$ gilt $F(\chi^2_{1-\alpha}(n)) = 1-\alpha$.
Links vom Quantil $\chi^2_{1-\alpha}(n)$ liegt die Wahrscheinlichkeitsmasse $1-\alpha$.
Ablesebeispiel: $\chi^2_{0.95}(10) = 18.307$
Approximation für $n > 30$: $\chi^2_\alpha(n) \approx \frac{1}{2}(z_\alpha + \sqrt{2n-1})^2$
(z_α ist das α-Quantil der Standardnormalverteilung)

n	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
1	0.0002	0.0010	0.0039	0.0158	0.4549	2.7055	3.8415	5.0239	6.6349
2	0.0201	0.0506	0.1026	0.2107	1.3863	4.6052	5.9915	7.3778	9.2103
3	0.1148	0.2158	0.3518	0.5844	2.3660	6.2514	7.8147	9.3484	11.345
4	0.2971	0.4844	0.7107	1.0636	3.3567	7.7794	9.4877	11.143	13.277
5	0.5543	0.8312	1.1455	1.6103	4.3515	9.2364	11.070	12.833	15.086
6	0.8721	1.2373	1.6354	2.2041	5.3481	10.645	12.592	14.449	16.812
7	1.2390	1.6899	2.1674	2.8331	6.3458	12.017	14.067	16.013	18.475
8	1.6465	2.1797	2.7326	3.4895	7.3441	13.362	15.507	17.535	20.090
9	2.0879	2.7004	3.3251	4.1682	8.3428	14.684	16.919	19.023	21.666
10	2.5582	3.2470	3.9403	4.8652	9.3418	15.987	18.307	20.483	23.209
11	3.0535	3.8157	4.5748	5.5778	10.341	17.275	19.675	21.920	24.725
12	3.5706	4.4038	5.2260	6.3038	11.340	18.549	21.026	23.337	26.217
13	4.1069	5.0088	5.8919	7.0415	12.340	19.812	22.362	24.736	27.688
14	4.6604	5.6287	6.5706	7.7895	13.339	21.064	23.685	26.119	29.141
15	5.2293	6.2621	7.2609	8.5468	14.339	22.307	24.996	27.488	30.578

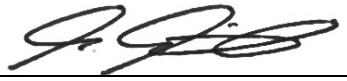
Quelle: Fahrmeier et al., S. 556

Anhang G: Meilenstein-Balken-Plan



Eidesstattliche Erklärung

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Lingen, 20.01.2021	
Ort, Datum	Unterschrift, M. Kuck
Lingen, 20.01.2021	
Ort, Datum	Unterschrift, M. Gierke
Lingen, 20.01.2021	
Ort, Datum	Unterschrift, F. Knese
Lingen, 20.01.2021	Niklas Fischer
Ort, Datum	Unterschrift, N. Fischer
Lingen, 20.01.2021	
Ort, Datum	Unterschrift, A. Lutterbeck