BEUTH HOCHSCHULE FÜR TECHNIK BERLIN
University of Applied Sciences

Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Beuth Hochschule für Technik - Fachbereich VIII:

Maschinenbau, Veranstaltungstechnik, Verfahrenstechnik

# Masterarbeit

angefertigt von

Lukas Sprengel

Matrikelnummer.: 842436

zur Erlangung des akademischen Grades

Master of Science (M. Sc.)

im Studiengang

Wirtschaftsingenieur/in - Energie und Umweltressourcen

## Thema:

# Explorative Datenanalyse von Kundenrezensionen hinsichtlich produktbezogener Nachhaltigkeitsaspekte mittels Text Mining Methoden

| | |
|---|---|
| Betreuer: | Prof. Dr.- Ing. Andreas Gerhard Goldmann |
| Betreuer extern: | Eduard Wagner (M. Sc.) |
| Eingereicht am: | 14.11.2018 |

BEUTH HOCHSCHULE FÜR TECHNIK BERLIN
University of Applied Sciences

Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

# Eidesstattliche Erklärung

Masterarbeit

Explorative Datenanalyse von Kundenrezensionen hinsichtlich produktbezogener Nachhaltigkeitsaspekte mittels Text Mining Methoden


von: Lukas Sprengel

Matrikelnummer: 842436


Ich versichere durch meine Unterschrift an Eides statt, dass ich die Arbeit selbständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen sind, habe ich als solche kenntlich gemacht.

Die Arbeit oder Auszüge daraus haben noch nicht in gleicher oder ähnlicher Form dieser oder einer anderen Prüfungsbehörde vorgelegen.

Ich weiß, dass bei Abgabe einer falschen Versicherung die Prüfung als nicht bestanden zu gelten hat.

Berlin, 14.11.2018

_____

*Unterschrift*

BEUTH HOCHSCHULE FÜR TECHNIK BERLIN
University of Applied Sciences

Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

# Abstract

Topic: Explorative data analysis of customer reviews regarding product orientated aspects of sustainability with the use of text mining methods

In this thesis a methodology is presented to explore a text dataset towards sustainability aspects of product design. The combined methods deal with the extraction, preprocessing, modelling, visualization and interpretation of text data. The dataset contains around 1.8 million words originating from customer reviews on the top sold smartphones on amazon.de due date 31st of July 2018. With the aim of generating information about feedback regarding sustainable product design, eco design directives were scanned and relevant keywords listed. Splitting and transforming the reviews into two vector space models, one based on very poor rating and one based on very good rating, enables to distinguish between sentimental context of the words. The keywords are queried e. g. by cosine similarity and the results, both in negative and positive sense are analyzed. New hypotheses and questions are opened for further research and possible applications are discussed.

Keywords: amazon product reviews, smartphones, German corpus, word embedding, word2vec, ecodesign directives, sustainable product design

Thema: Explorative Datenanalyse von Kundenrezensionen hinsichtlich produktbezogener Nachhaltigkeitsaspekte mittels Text Mining Methoden

In der vorliegenden Arbeit wird eine Methodik vorgestellt, mit der ein Textdatensatz explorativ auf Nachhaltigkeitsaspekte erforscht werden kann. Durch die Kombination verschiedener Methoden wird der gesamte Prozess von der Datenextraktion, -vorverarbeitung bis hin zur Modellbildung und Auswertung dargestellt. Die aus ca. 1,8 Millionen Wörtern bestehende Datengrundlage basiert auf Kundenrezensionen der meistverkauften Smartphones auf Amazon.de zum Stichtag 31.07.2018. Durch den gewählten Fokus auf Nachhaltigkeitsaspekten werden gezielt relevante Schlüsselwörter aus Ökodesignrichtlinien gesammelt und nach Relevanz geordnet. Anschließend können gezielt Zusammenhänge dieser Schlüsselwörter - bspw. durch die Kosinus – Ähnlichkeit - in zwei generierten Wordvektormodellräumen, gebildet aus Rezensionen der sehr schlechten und sehr guten bewerteten Produkte, abgefragt werden. Die auf diese Weise entdeckten Erkenntnisse werden diskutiert und mögliche neue Forschungsfelder formuliert. Außerdem wird die entwickelte Methodik diskutiert, Optimierungsvorschläge & -ideen und Anwendungsfelder aufgezeigt.

Keywords: Produktrezensionen, Smartphones, Korpus, Word Embedding, Vektorreprä sentation, word2vec, Ökodesignrichtlinie, Nachhaltiges Produktdesign

# I Inhaltsverzeichnis

I Inhaltsverzeichnis ..... IV
II Abbildungsverzeichnis ..... VI
III Tabellenverzeichnis ..... VIII
IV Formelverzeichnis ..... IX
V Abkürzungsverzeichnis ..... X
VI Glossar / Klärung relevanter Begriffe ..... XI
VII Symbolverzeichnis ..... XIII
1 Einleitung und Motivation ..... 1
2 Aufbau der Arbeit ..... 4
3 Grundlagen ..... 6
3.1 Aspekte des nachhaltigen Produktdesigns ..... 6
3.2 Big & Ambient Data auf Produktebene ..... 8
3.3 Explorative Datenanalyse nach Tukey ..... 9
3.4 Verortung der Text Mining Verfahren in den Datenwissenschaften (Data Science) 11
3.5 Methoden zur Datenextraktion ..... 14
3.6 Verfahren des Natural Language Processing ..... 16
3.6.1 Repräsentation von Wörtern als Vektoren (Word Embeddings) ..... 18
3.6.2 Dimensionsreduzierung ..... 22
3.6.3 Messung der Relevanz und Besonderheit von Wörtern ..... 23
3.6.4 Überblick über weitere Tools und Verfahren im Bereich Data und Text Mining 24
4 Methodische Vorgehensweise ..... 25
4.1 Extraktion von Keywords aus Ökodesigndirektiven ..... 26
4.2 Data & Text Mining: Datenbeschaffung, -verarbeitung & -visualisierung ..... 27
4.2.1 Datengrundlage schaffen: Datenextraktion ..... 28
4.2.2 Aufbereiten der Rezensionen mittels Natural Language Processing ..... 28
4.2.3 Term Frequency – Inversed Document Frequency ..... 31
4.2.4 Wortrepräsentationen mit Word2Vec ..... 33

4.3 Visualisierung durch Dimensionsreduzierung (t-SNE) ..... 34
5 Explorative Datenanalyse mittels Text Mining Methoden ..... 35
5.1 Extraktion von Produktrezensionen und Vorbereiten der Daten zur Analyse ..... 35
5.2 Analyse der geringdimensionalen Daten mittels Interaktiven Statistischen Grafiken ..... 37
5.3 Keywordextraktion auf Basis von Ökodesignrichtlinien ..... 39
5.4 Qualitatives Screening des Corpus nach beschreibender äquivalenter Umgangssprache ..... 42
5.5 Bestimmung der Relevanz der ausgewählten Keywords im Corpus ..... 44
5.6 Untersuchung der relevantesten Wörter mithilfe von Word Embeddings ..... 48
5.7 Ergebnisse Dimensionsreduzierung mit Tensorflow Projector ..... 51
6 Evaluierung und Diskussion der Ergebnisse ..... 54
6.1 Ergebnisse der Untersuchung mit Bezug zur Fragestellung ..... 54
6.2 Methodik ..... 54
6.2.1 Datenextraktion & Preprocessing ..... 54
6.2.2 Datenanalyse ..... 56
7 Fazit und Ausblick ..... 60
VIII Anhang ..... XIV
VIII.I Formatierte Keywordliste ..... XIV
VIII.II Codebausteine Webcrawler ..... XVII
VIII.III DVD / CD ..... XXI
8 Literaturverzeichnis ..... XXII

# II Abbildungsverzeichnis

Abbildung 2: Aufbau der Arbeit (eigene Darstellung) ..... 4

Abbildung 3: Stufen einer kontinuierlichen Optimierung hin zu nachhaltigen Produkten (Marwede und Jokinen 2018) ..... 6

Abbildung 4: Analyse des Russlandfeldzuges Napoleons 1812-183 (Minard, M. 1869).....10

Abbildung 5: KDD - Prozess in fünf Schritten (Fayyad et al. 1996) .....12

Abbildung 6: Venn Diagramm, das die Überschneidungen von Text Mining und sechs verwandten Disziplinen visualisiert (Miner et al. 2012) .....13

Abbildung 7: Charakteristika und Techniken des Maschinellen Lernens (MathWorks.com 2018).....17

Abbildung 8: Visualisierung des Konzepts einer Wertematrix (engl. co - occurence matrix) bei der Anwendung von word2vec (Mitra 2016).....20

Abbildung 9: Modellhafte Darstellung der entwickelten Wasserfall - Methodik der vorliegenden Arbeit (eigene Darstellung).....25

Abbildung 10: Beispiel für die Struktur eines XML - Baums (Codeausschnitt von amazon.com) .....15

Abbildung 11: Liste der der 50 Geschwisterknoten <li>, jedes Element entspricht einem Gerät (Codeausschnitt von amazon.com) .....15

Abbildung 12: Einlesen der Datenbank und Zuweisen der Rezensionen in Variable "onestar_reviews" (eigene Darstellung).....29

Abbildung 13: Zwei Funktionen zum Pre - Processing der Rezensionen (eigene Darstellung) .....30

Abbildung 14: Ablegen der generierten Listen in Datenbanken (eigene Darstellung).....31

Abbildung 15: Angepasster TF - IDF Code für das Relevanzranking der Keywords (Vembunarayanan 2013).....32

Abbildung 16: Parameterkonfiguration und Modellierung des Corpus (eigene Darstellung)..33

Abbildung 17: Gegenüberstellende Abfrage eines Zielwortes und Ausgabe der zehn ähnlichsten Wörter in den beiden Modellen (eigene Darstellung) .....34

Abbildung 18: Methoden zur Validierung eines Modells (Řehůřek 2014) .....34

Abbildung 19: Ausschnitt der entstandenen Rezensionsdatenbank mit einer Rezension je Zeile und insgesamt gesamt ca. 14770 Einträgen (eigene Darstellung) .....36

Abbildung 20: Anzahl Rezensionen, Sätze und Wörter innerhalb des Corpus (eigene Darstellung).....36

Abbildung 21: Stichprobenauswertung, 33 verschiedene Modelle (eigene Darstellung) .....37

Abbildung 22: Gegenüberstellung der durchschnittlichen Bewertung und dem Gerätepreis (eigene Darstellung mit Tableau®).....38

Abbildung 23: Keywords mit mehr als 40 Nennungen in sehr schlechten / ein Stern Rezensionen (eigene Darstellung) .....40

Abbildung 24: Keywords mit mehr als 100 Nennungen in sehr guten / fünf Sterne Rezensionen (eigene Darstellung) .....40

Abbildung 25: TF - IDF - Werte für die relevantesten Keywords im Teilcorpus (1 Stern) (eigene Darstellung).....47

Abbildung 26: TF - IDF - Werte für die relevantesten Keywords im Teilcorpus (5 Sterne) (eigene Darstellung).....47

Abbildung 27: Wörter im Umfeld des Wortes *batterie im 5 Sterne Corpus* (eigene Darstellung mit Tensorflow® Projector) .....51

Abbildung 28:Wörter im Umfeld des Wortes *display* im 5 Sterne Corpus (eigene Darstellung mit Tensorflow® Projector) .....52

Abbildung 29: Wörter im Umfeld des Wortes *rutschig* im 5 Sterne Corpus (eigene Darstellung mit Tensorflow® Projector) .....53

# III Tabellenverzeichnis

Tabelle 1: Selektierte und Extrahierte Tags, die als abhängige Spalten in die Datenbank ..... geschrieben werden (eigene Darstellung) .....35

Tabelle 2: Übersicht Ergebnisse der Datenextraktion (eigene Darstellung).....38

Tabelle 3: Ergebnisse des Durchsuchens der Corpora nach Kontext und umgangssprachlichen Wörtern (eigene Darstellung) .....42

Tabelle 4: Gegenüberstellung von den jeweils relevantesten Wörtern in sehr schlechten und sehr guten Rezensionen, gemessen am TF-IDF Wert in den Teilcorpora (eigene Darstellung) .....44

Tabelle 5: Keywords und ihre Kosinus Ähnlichkeiten in den Corpora (eigene Zusammenstellung).....48

# IV Formelverzeichnis

Formel 1: Kosinus - Ähnlichkeit zwischen den Vektoren i und j oder analog zweier Wortvektoren.....21

Formel 2: Wortanalogie durch Vektoroperationen (Mikolov et al. 2013).....21

Formel 3: Inversed Document Frequency.....23

# V Abkürzungsverzeichnis

| | |
|---|---|
| ASIN | Amazon Standard Identification Number |
| BITKOM | Bundesverbandes Informationswirtschaft, Telekommunikation und neue Medien e.V. |
| CBOW | Continous Bag of Words |
| CSS | Cascading Style Sheets |
| GAFA | Google, Amazon, Facebook & Apple |
| IoT | Internet of Things |
| IR | Information Retrieval |
| ISG | Interaktiven Statistischen Grafik |
| KDD | Knowledge Discovery in Databases |
| MIT | Massachusetts Institute of Technology |
| NLP | Natural Language Processing |
| NNLM | Neuronalen Netzen für Sprachmodellierung (engl. „Neural Networks Language Models”) |
| SEO | Search Engine Optimization |
| SQL | Structured Query Language |
| TDM | Text und Data Mining |
| TF – IDF | Term Frequency – Inversed Document Frequency |
| t - SNE | t-Distributed Stochastic Neighbor Embedding |
| VSM | Vector Space Model |
| XML | eXtensible Markup Language |
| Xpath | XPath Language |

# VI Glossar / Klärung relevanter Begriffe

Alle in Kapitel 6 *kursiv* gedruckten Wörter entstammen den selbst erstellten Keywordlisten bzw. dem Corpus.

| | |
|---|---|
| Algorithmus | Rechenauftrag für einen Computer, der auf selbstdefinierten Prozeduren, Schleifen und Variablen beruht. Oft werden Pakete kombiniert um neue Verfahrensweisen zu generieren. (Gabler 2018a) |
| Bag of Words Model | Zusammengefasste Wortliste, oft mit Häufigkeiten aller Wörter eines definierten Corpus (Harris 1954) |
| Corpus | Liste aller Wörter in einem Datensatz (Hardeniya 2015) |
| Documents | In dieser Arbeit ist ein document gleichzusetzen mit einer Rezension. Allgemein gesprochen ist ein document im Bereich des Natural Language Processing (NLP) eine Abfolge von Wörtern innerhalb eines Gesamtkontextes. (Manning et al. 2008) |
| Information Retrieval | Wissenschaftsgebiet, in dem mittels strukturierter Methoden und Vorgehensweise aus großen Textmengen z. B. Muster zu aufzudecken und dadurch Informationen und schließlich Wissen zu generieren. (Fuhr 2010) |
| Lemmatisierung | Verschiedene Wortformen (Konjugationen, Deklinationen) werden zu einem eindeutigen Wortstamm umformuliert, sodass z.B. die statistische Auswertung vereinfacht und genauer wird. Beispiel: *geh* (*gingen*, *gegangen*, *gehe*) (Manning et al. 2008) |
| Machine Learning | Ein Überbegriff im Bereich der Informations- & Mustererkennung in Datenbanken. Hauptsächlich geht es um das Generieren von Wissen aus Erfahrung mittels Algorithmen (MathWorks.com 2018) |
| Open - Source | Frei zugängliche Software (- bibliotheken), die dem Entwickler zur Nutzung und Veränderung zur Verfügung stehen. (Gabler 2018b) |
| Syntax | sprachenabhängiger Satzbau, sowohl für gesprochene, geschriebene als auch Programmiersprachen (Fuhr 2010) |
| Semantik | Die Bedeutung hinter den Wörtern. Für die Computerlinguistik teils sehr schwer zu erfassen: Ironie, Sarkasmus, aber auch einfache Mehrdeutigkeiten (Fuhr 2010) |

| | |
|---|---|
| Stemming | Ein Token wird auf den Wortstamm reduziert, um verschiedene Erscheinungsformen erfassen zu können. Im Gegensatz zur Lemmatisierung wird beim Stemming der hintere Teil des Tokens entfernt, ohne auf eine „Lemming“ – zurückzugreifen. (Manning et al. 2008) |
| Tokenisierung | Im Bereich der Textverarbeitung spricht man von „Tokenization“, einem Abgrenzen einer Buchstabenfolge in klassifizierende Sektionen (Tokens). Die Tokens dienen meist als Grundlage für weiter Prozessschritte. (Trim 2013) |
| Vokabular | „Wortschatz“ eines Corpus, Wortliste aller im Corpus vorkommenden Wörter, ohne Dopplungen (Hardeniya 2015) |

# VII Symbolverzeichnis

| | |
|---|---|
| K | Gesamtheit an Keywords |
| k | Teilmenge von K |
| N | Grundgesamtheit (hier Anzahl Wörter im Corpus) |
| n | Teilmenge von N |
| $df_k$ | Document Frequency von Wort k |
| $idf_k$ | Inversed Document Frequency von Wort k |

# 1 Einleitung und Motivation

Der 7. Januar 2007 hat rückblickend, wie von dem verursachenden Erfinder prognostiziert, die Welt verändert. An diesem Tag wurde von der Firma Apple das erste Smartphone ohne Tastatur und mit intuitiv bedienbarem Touchscreen angekündigt. Das Gerät vereinte erstmals Funktionen in einem Telefon, für die bis dato auch die Anschaffung weiterer Geräte notwendig war: Fotografie, Musik hören, Digitales Lesen, Mobile Internetnutzung, Nachrichtendienstleistungen, Kalenderfunktion und viele mehr. Dieser zusätzliche Nutzen für jeden einzelnen ist unschlagbar effizient und fortschrittlich, ermöglicht ein Smartphone doch heutzutage z. B. unter verhältnismäßig geringem finanziellen Aufwand den Zugang zu weltweit verfügbarem Wissen, Bildung und einer Vielzahl an „kostenlosen“ Diensten.

Man hätte vermuten können, dass durch diese Verdichtung von Features auch das Potential der Ressourcenschonung erhöht ist, da theoretisch weniger technische Geräte insgesamt produziert werden müssten.

Die Entwicklung der Verkaufszahlen der letzten Jahre hat aber gezeigt, dass in den letzten fünf Jahren jährlich rund 200 - 300 Millionen Smartphones mehr verkauft wurden als im jeweiligen Vorjahr. Im Jahr 2019 sollen es erstmals mehr als 1,5 Milliarden verkaufte Geräte werden. Nach zehn Jahren waren insgesamt sieben Milliarden Geräte auf dem Planeten im Umlauf. Äußerst kritisch ist in diesem Zusammenhang u. a. zu sehen, dass die $CO_2$ Emissionen während der Produktion je Gerät stetig ansteigen, was auf die immer komplexer und dichter veredelte Mikroelektronik zurückzuführen ist. (Apple Inc. 2017; Gartner 2018; IDC 2018a, 2018b)

In Kombination mit einer vielfach unterstellten, sich wegen teilweisen oder vollständigen Soft- oder Hardwaredefekts, verkürzenden Nutzungsdauer, bestenfalls aber wegen einer Weitergabe / -verkaufs des Geräts wird deutlich, wie immens die Umwelteinwirkungen dieser Produktgruppe sind und wie sinnvoll und notwendig die Bestrebungen hin zu einer möglichst langen Nutzung von Smartphones und generell elektrischen Geräten sind, welches durch intelligentes, nachhaltiges Produktdesign gefördert werden kann. (Jaeger-Erben und Hipp 2017)

So führen die Vereinfachungen und Annehmlichkeiten im Leben jedes Einzelnen von Smartphones und generell Elektrogeräten in globaler Perspektive zu einem erheblichen Eingriff auf ökologischer und auch sozialer Ebene, wenn auf Themen wie die Arbeitsbedingungen beim Abbau von seltenen Erden oder der Endmontage der Geräte aufmerksam gemacht wird und auf die gleiche Stufe mit den vermeintlich positiven ökonomischen Entwicklungen gestellt werden. (Jardim 2017; tagesschau.de 2018; Stiftung Warentest 2018)

Es ist zu erwähnen, dass die Geräte tendenziell effizienter werden, aber die zu beobachtende Entwicklung durch die immer steigende Nachfrage nach neuen und ressourcenintensiveren Geräten als klassischen Rebound - Effekt zu betiteln ist. (Rogall 2009)

Diese rasante globale Entwicklung lässt sich vermutlich nicht bremsen oder durch nationale Markteingriffe im Sinne einer ökologisch orientierten Gesetzgebung verhindern, sodass mit dieser immensen Beschleunigung in gleichem Maße der Ressourcenverbrauch in Form von kritischen Materialien und elektrischer Energie für die Versorgung der Infrastruktur, Produktion etc. außerordentlich ansteigen wird. Es liegt demnach an den Herstellern der Geräte, die den globalen Markt beliefern, diese möglichst zirkulär und damit umweltfreundlicher zu gestalten und auf eine Verringerung des ökologischen Impacts hinzuarbeiten. (Morley et al. 2018)

Positive Beispielunternehmen (z. B. Fairphone, Puzzlephone, Shiftphone), die sich den Problemen annehmen und praxistaugliche, performante Geräte anbieten, formieren sich zur Zeit, um im Sinne einer nachhaltigen Produktentwicklung Smartphones so „fair“ wie möglich produzieren und somit die Idee der nachhaltigen Entwicklung zu verbreiten.

Da Smartphones ihrer Natur wegen ständig mit dem Internet verbunden sind, wundert es nicht, dass allein in den Jahren 2013 und 2014 ca. 90 % des bis damals weltweiten Datenhaushaltes neu produziert wurden und wird prognostiziert wird, dass bis zum Jahr 2025 163 Zettabytes – zehn Mal mehr Daten als noch im Jahr 2016 - in der „datasphere“ erzeugt werden. So sollen täglich durchschnittlich rund 4.800 Interaktionen pro Person über verschiedene Schnittstellen – z. B. Internet of Things (IoT) fähige Geräte – über das Internet und zu Großteilen mit Unternehmen stattfinden. (IBM 2014)

Es kann also davon ausgegangen werden, dass sich die abzeichnende Entwicklung von einem erheblich steigenden Ressourcenverbrauch, letztlich in Form von Produktionsmaterialien und Primärenergieträgern noch um einiges beschleunigen wird.

Diese Entwicklung der Daten- und Informationsflut nutzend, wird in dieser Arbeit eine generische Methodik vorgestellt, mit der basierend auf verschiedenen Methoden der Computerlinguistik, explorativ Daten analysiert werden können. Der Motivation folgend, dass nach einer repräsentativen Studie 60 % der europäischen Bevölkerung Onlineempfehlungen vertrauen und wertschätzen (Kreutzer 2017), werden in dieser Arbeit reale Produktrezensionen von aktuellen Smartphonemodellen analysiert. Dabei soll folgende Fragestellung untersucht werden:

> Welche ausgewählten Nachhaltigkeitskriterien für Elektronikgeräte können in Form von Schlüsselwörtern in einem Textkorpus aus Kundenrezensionen auf Produkt- und Herstellerebene gefunden werden und in welchem Zusammenhang werden diese Wörter genannt?

Um diese Frage zu beantworten werden Unterfragen formuliert, die im Laufe der Arbeit durch Literaturrecherche und empirische Untersuchungen untersucht und beantwortet werden:

> Wo und wie werden Nachhaltigkeitskriterien für Elektronikgeräte definiert? (u. a. Kapitel 3.1, 4.1)
>
> Welche Methoden gibt es im Bereich des Text Mining Verfahren, insbesondere um semantische Ähnlichkeiten zwischen Wörtern herauszufinden? (u. a. Kapitel 3.4, 4.2)
>
> Wie können Kundenrezensionen quantitativ auswertbar gemacht werden? (u. a. Kapitel 3.5, 4.2, 5)
>
> Wie können die Ergebnisse aufbereitet werden? (u. a. Kapitel 3.5, 4.2, 5)
>
> Lassen sich herstellerabhängig bestimmte Eigenschaften, positiv wie negativ, mittels Nachhaltigkeitskeywords herausfiltern? (u. a. Kapitel 5, 6)
>
> Wie können Datenquellen miteinander verbunden werden, damit zeitliche und preisliche Komponenten mit Rezensionsinhalten verbunden werden können? (Kapitel 6, 7)

Ziel der Arbeit ist es herauszufinden, in wie weit die Kombination von Text Mining Methoden wie z. B. Wortvektormodellierung auch mit relativ kleinen Datensätzen befriedigende Ergebnisse liefern kann und auf diesem Weg brauchbare Informationen generiert werden können, die in eine nachhaltige Produktentwicklung & -design einfließen können. Dazu wird eine Vorgehensweise entwickelt, angewendet und deren Ergebnisse diskutiert. Außerdem können, aufbauend auf den Ergebnissen, Hypothesen und Fragestellungen zur weitergehenden Forschung entwickelt werden.

# 2 Aufbau der Arbeit

In der vorliegenden Arbeit werden Text Mining Methoden auf reale Produktrezensionen angewendet, um Zusammenhänge zwischen nachhaltigen Produktdesign und Kundenansprüchen und -wünschen sichtbar zu machen und weitergehende Thesen und Fragestellungen aufzustellen und zu entwickeln. Die Arbeit folgt dabei einer Kapitelstruktur von theoretischen Grundlagenstand, Methodische Vorgehensweise, Empirie, Diskussion, Fazit und Ausblick.

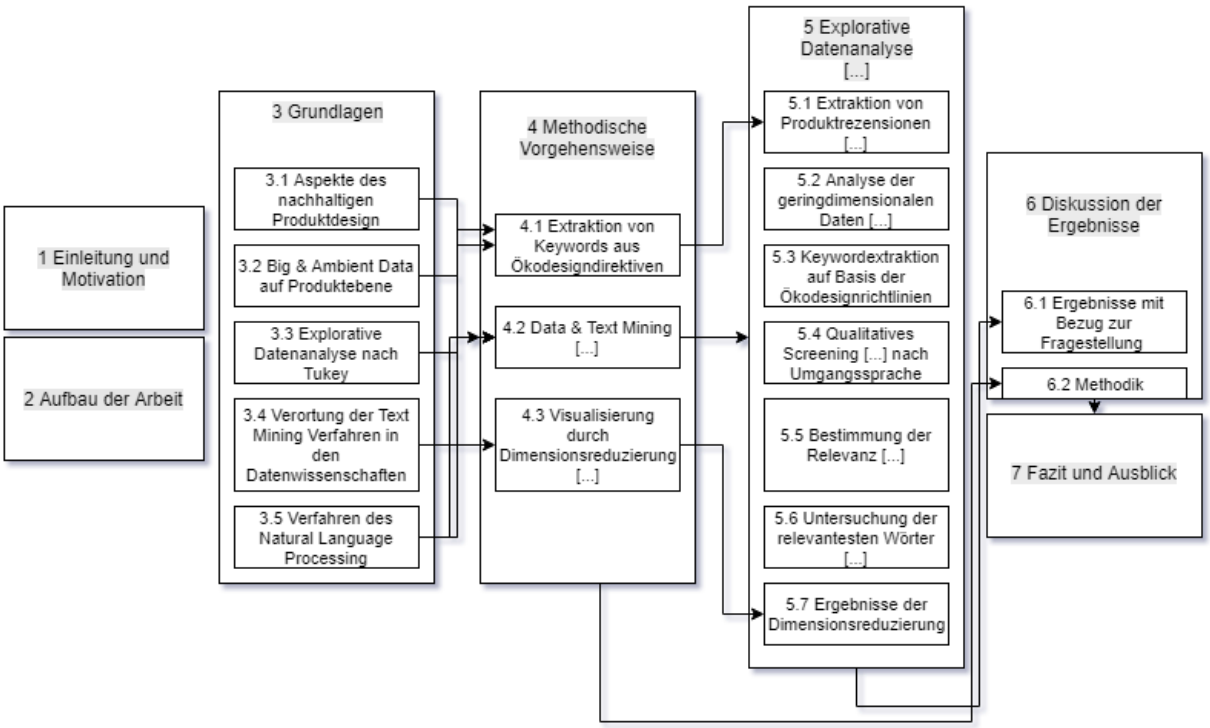


**Abbildung 1: Aufbau der Arbeit (eigene Darstellung)**

Diesem skizzierten Bezugsrahmen (Abbildung 1) folgend werden zunächst die für das Verständnis des Themas notwendigen Grundlagen vermittelt (Kapitel 3). Dabei geht es um die Verflechtung verschiedener Wissenschaftsdisziplinen. Zum einen werden Aspekte des nachhaltigen Produktdesigns erklärt und inhaltlich darauf aufbauende praxisrelevante Ökodesignrichtlinien vorgestellt (Kapitel 3.1)

Anschließend wird eine Brücke geschlagen zur IT-gestützten Datenanalyse. Es wird, aufbauend auf der im vorigen Kapitel genannten globalen Entwicklungen, ein Überblick gegeben über aktuelle Entwicklungen von frei verfügbaren, von Rezensenten generierten Produktdaten. (Kapitel 3.2).

Darauf folgt die Vorstellung einer möglichen Herangehensweise, um große Mengen an Daten zu analysieren: Die Explorative Datenanalyse nach Tukey (Kapitel 3.3). Weiterhin wird ein Überblick über die Wissenschaftsgebiete gegeben, in denen die vorgestellten Text Mining Verfahren Anwendung finden. (Kapitel 3.4)

Im letzten Grundlagenkapitel geht es speziell um für die Arbeit relevante Verfahren des Natural Language Processing, wie dem Word2Vec Modell, Möglichkeiten der Dimensionsreduzierung von Vektorrepräsentationen und einem Verfahren zur Relevanzmessung (Kapitel 3.6)

Dieses dargestellte Wissen wird in den folgenden Kapiteln zu einer entwickelten Methodik verarbeitet, mit der systematisch Nachhaltigkeitsaspekte in einem Corpus abgefragt werden können. Kapitel 4 beginnt mit einer Übersicht und der Beschreibung eine Methode, eine Keywordliste auf Basis von Ökodesignrichtlinien zu generieren (Kapitel 4.1). Anschließend wird die weitere Vorgehensweise der relevanten Text und Data Mining Verfahren entsprechend der Methodik erklärt (Kapitel 4.2) und die Möglichkeiten der Visualisierung aufgezeigt (Kapitel 4.3).

Im empirischen Teil finden sich die Ergebnisse der Methoden und Tools der bis dahin bekannten Struktur folgend (Kapitel 5.1 - 5.7: Keywordlisten, statistische Auswertungen, Sortierung der Keywords, Wortvektor – Modellauswertung, ausgewählte 2D Ansicht des Vektorraums).

Die Diskussion und Evaluierung der Ergebnisse mit Bezugnahme auf die ursprüngliche Fragestellung finden schließlich in Kapitel 6.1 statt. Die Methode wird in Kapitel 6.2 diskutiert und erörtert.

Abschließend werden die Ergebnisse zusammengefasst und Fragestellungen für die Zukunft offengelegt (Kapitel 7). Im Anhang ist außerdem die vollständige Keywordliste zu finden sowie die für die Datenbeschaffung benötigten die Codebausteine in der Programmiersprache Python zu finden.

# 3 Grundlagen

## 3.1 Aspekte des nachhaltigen Produktdesigns

Entsprechend der allgemeinen drei Nachhaltigkeitsstrategien Effizienz, Konsistenz, Suffizienz (Rogall 2009) wird mit folgendem Modell ein Weg zu einem nachhaltigen Produktdesign konzipiert:

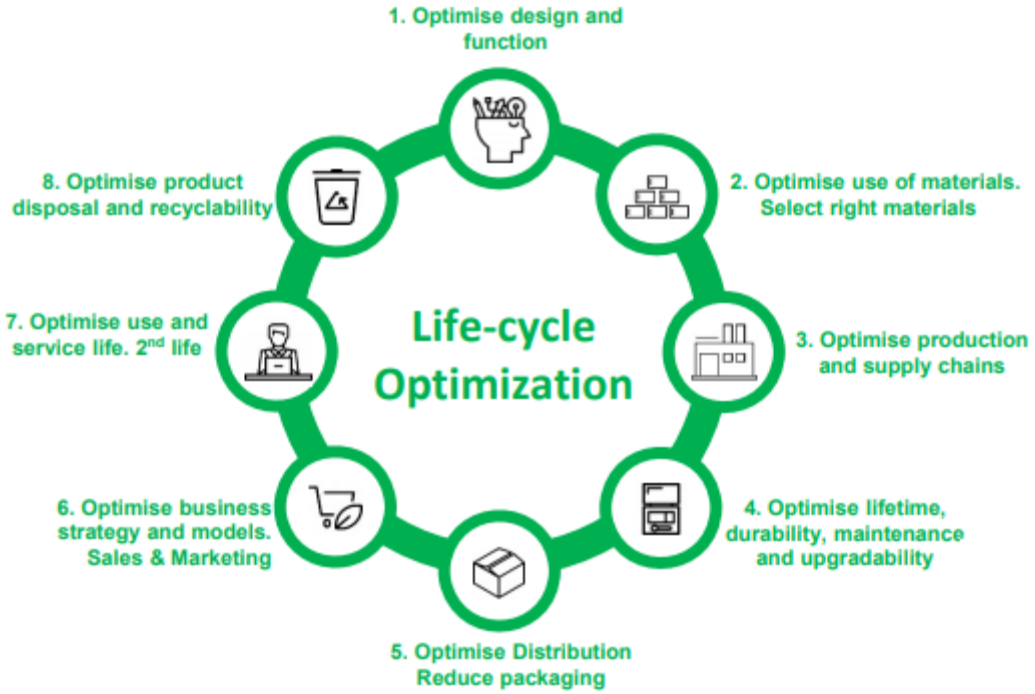


**Abbildung 2: Stufen einer kontinuierlichen Optimierung hin zu nachhaltigen Produkten (Marwede und Jokinen 2018)**

In den sich wiederholenden acht Prozessschritten wird der gesamte Lebenszyklus von Produkten ausgehend vom Design, den Materialien, den Produktionsprozessen und Lieferketten untersucht. Weiterhin wird untersucht, wie die Nutzungsdauer, Haltbarkeit sowie Wartungs- & Upgrademöglichkeiten verlängert bzw. vergrößert werden können. Auch werden Potenziale für die Reduzierung der Distribution und Verpackung von Gütern gesucht, deren Nutzungsphasen sowie die Beseitigung und Recyclingfähigkeit. Nicht zuletzt wird ein Augenmerk auf die Marketing- & Vertriebsstrategien und die Geschäftsmodelloptimierung gelegt. Es handelt sich also in einem gewissen Rahmen um ein holistisches Konzept bei dem verschiedene Blickwinkel und Stakeholder mit in eine Produktentwicklung im Sinne der drei Nachhaltigkeitsstrategien mit einbezogen werden.

Ein Vorschlag, auf Produktdesignebene ein Umdenken und Einlenken zu erreichen, sind beispielsweise sechs formulierte Strategien bzw. Produkteigenschaften, die an der TU Delft formuliert wurden nach Bakker::

| | *Original* (Bakker 2014)*:* | *Übersetzt:* |
|---|---|---|
| 1. | Product attachment & trust | Produktverbundenheit & -vertrauen |
| 2. | Product durability | Produkthaltbarkeit / -robustheit |
| 3. | Standardization & compatibity | Standardisierung & Kompatibilität |
| 4. | Ease of maintenance and repair | Einfachheit der Wartung und Reparatur |
| 5. | Upgradebality & adaptability | Erweiterungsfähigkeit & Adaptivität |
| 6. | Dis- & reassembly | (De-) Montagefähigkeit |

Es wird demnach vermutet, dass durch eine Verbundenheit zu Produkten auch eine nichtmonetäre Wertsteigerung entsteht und Konsumenten achtsamer mit ihren Geräten umgehen. Durch ein hohes Standardisierungslevel können bzw. konnten Fehler in der Produktion und schließlich in den Produkten verringert werden, aber auch die Austauschbarkeit und Kompatibilität von Bauteilen zwischen Gerätegenerationen kann gewährleistet werden. Um dies und bestenfalls eine Erweiterungsfähigkeit durch modulare Bauweise zu ermöglichen, müssten einfach zu wartende bzw. zu reparierende demontagefähige Geräte entwickelt werden. (Bakker 2014)

Diese Strategien und „Produktfähigkeiten“ finden sich direkt oder indirekt ebenfalls in den weltweit verfügbaren Ökodesigndirektiven wieder, siehe z. B. Blauer Engel (RAL gGmbH 2013), TCO (TCO Certified 2018) die in der Produktentwicklung als Leitfaden genutzt werden können. In dem Gros der Richtlinien werden Vorgaben und Vorschläge zu Bereichen des gesamten Lebenszyklus auf Produktebene gegeben: Materialien, Energieverbrauch, End of Life (EoL) Management, Verlängerung der Nutzung und Verpackung. Außerdem werden auf Unternehmensebene u. a. unternehmerische und operative Tätigkeiten, Produktion und Innovationsmanagement im Hinblick auf nachhaltige Verhaltensweisen angestoßen. (Underwriters Laboratories Inc. 2018)

Deren Wortlaute werden im Rahmen der vorliegenden Arbeit, zusammen mit adaptierter Umgangssprache, in Kundenrezensionen von Smartphones mithilfe einer entwickelten Text Mining Methodik untersucht. Dies dient dazu, Muster aufzudecken, was Kunden in dem Zusammenhang mit nachhaltigem Produktdesign äußern und wie positiv oder negativ dieser Äußerungen sind.

## 3.2 Big & Ambient Data auf Produktebene

„Ambiente Daten sind digitale Daten, die durch Kopie oder durch Aufzeichnung von ohnehin stattfindenden Prozessen erhoben werden."

(Thede 2014)

Global agierende Onlineunternehmen wie die GAFA (Google, Amazon, Facebook, Apple) aber auch Twitter oder eBay erzeugen täglich immense Mengen an Daten, die zu großen Teilen z. B. über Schnittstellen, sogenannte Application Programming Interfaces (APIs) von den Unternehmen zugänglich gemacht werden. Des Weiteren besteht die Möglichkeit, mithilfe sogenannter „Webcrawler" Elemente von Webseiten strukturiert in sogenannte (Non-) Structured Query Language (SQL) - Datenbanken herunterzuladen und darüber zu verfügen.

So wurde in der Vergangenheit beispielsweise ca. 35 Millionen Produktrezensionen von Amazon von (2009 – 2014) von der Stanford University (Lekovec 2013) und ca. 142 Millionen Rezensionen vom größeren Zeitraum von 1996 – 2014 durch strukturiertes Web Crawling für Forschungszwecke heruntergeladen. Die Bereiche Bücher, Elektronik und Filme & TV weisen demnach die meisten Rezensionen auf. (McAuley 2014)

Daneben existieren weitere Emittenten von Datenstreams, wie z. B. Wetter - & Umweltdaten oder Börsendaten, die teils gegen Lizenz, aber zu Großteilen auch frei verfügbar sind. (Press 2013)

Wie im Titel dieser Arbeit festgehalten, wird sich auf den Teil der frei verfügbaren Kundenrezensionen konzentriert. Auf Herstellerebene wird das Phänomen seit wenigen Jahren auch als „Review and Rating Management" im Bereich des Dialogmarketings bezeichnet. (Kreutzer 2017)

Aus Herstellersicht sind diese Kundenrezensionen neben der potenziellen Umsatzsteigerung auch Quelle für direktes Feedback von den Konsumenten. Negative Bewertungen beinhalten oft Informationen zu technischen Mängeln, nicht eingehaltenen Leistungen oder unfreundlichem Kundenservice. Diese Erkenntnisse können als kostengünstige Ausgangsbasis für Verbesserungen dienen und sollten zu entsprechender Stelle kommuniziert werden (z. B. Service, Produktentwicklung, Versand). (Kreutzer 2017)

Den immensen Einfluss von Onlinebewertungsportalen auf die Kaufentscheidung durch „Social Proof" und sogenanntem „electronic Word of Mouth" (Rensink 2013) ausblendend, wird

in dieser Arbeit aus Produktdesigner– bzw. Herstellersicht untersucht, ob die vom Kunden generierten Textdaten in möglicherweise nachhaltigere folgende Generationen der Produkte einfließen können. Dieses Konzept, es schließt auch andere vom Kunden generierten Daten ein, wird in der Literatur als „Prosumerism“ oder auch „Arbeitender Kunde“ bezeichnet. (Beyreuther et al. 2012)

Aufgrund der Wichtigkeit von Produktrezensionen im Verkaufsprozess haben sich Hersteller im Laufe der Zeit auch Nebenerscheinungen und Marktpotentiale von „incentivised“ Rezensionen nicht entgehen lassen. Dazu sei an dieser Stelle erwähnt, dass Amazon diese vergüteten Rezensionsleistungen stark reglementiert und eigene Lösungen gegen diese Phänomene entwickelt und in diesem Sinne neben eigenem Imageschutz auch Käuferschutz betreibt. Diese Entwicklungen führten hingegen dazu, dass seitens der Hersteller und Händler immer mehr Aufwand betrieben wurde, um die „fake reviews“ zu verbergen. (McCabe 2018)

Es gibt allerdings auch Bestrebungen von unabhängigen und kommerziellen Anbietern, die versuchen mithilfe einer Kombination von diversen transparenten Faktoren auf Basis der frei verfügbaren Kundendaten (z. B. Anzahl von Rezensionen, Wortvarianz, Anzahl gelöschter Rezensionen, Markenloyalität, etc.) die durchschnittliche Bewertung der Produkte realistischer zu gestalten. (Review Honesty Inc. 2018)

## 3.3 Explorative Datenanalyse nach Tukey

„Exploratory Data Analysis can never be the whole story,
but nothing else can serve as the foundation stone as the first step.“

(Tukey 1977, S. 3)

Dieser Aussage folgend wird in dieser Arbeit ein Datensatz systematisch beschrieben und analysiert. Übergeordnetes Ziel ist es, neue Erkenntnisse und Informationen in einem bisher wenig oder nicht auf diese Art und Weise bearbeiteten Forschungsfeld zu gewinnen. Dabei werden Annahmen über die Ursachen und Gründe der beobachteten Daten gebildet, Grundannahmen für spätere Interferenzstatistik, also dem Aufrufen und Behandeln von sich ergebenden Hypothesen (schließende Statistik) durch weitergehende Umfragen oder Designs von Experimenten entwickelt. (Diehl und Arbinger 2001)

Analog zur deskriptiven Statistik werden dabei vorliegende Daten zunächst statistisch erfasst und eingegliedert. Allgemeine Begriffe wie Grundgesamtheit, Stichprobe, uni- und multivariate

Merkmalsträger, Ausprägungen, diskret und stetige Merkmale gelten ebenso wie die Klassierung der Daten in die Skalenniveaus (nominal-, ordinal-, intervall- oder verhältnisskaliert) mit den daraus folgenden Möglichkeiten der Merkmalsauswertung. (Fahrner 2011)

Die verfügbaren und angewandten Verfahren entwickeln sich ähnlich schnell wie die Computer- und Softwaretechnologie, sodass mithilfe sogenannter Interaktiven Statistischen Grafiken (ISG) relativ zügig brauchbare graphische Auswertungen zu realisieren sind. (Fahrmeir et al. 2016)

Davon ausgehend, dass grundsätzlich ein geringes Wissen über Zusammenhänge von Daten besteht, da diese aus verschiedenen Quellen und Zusammenhängen stammen können, sei es bei der Vorgehensweise erwünscht, wenn nicht auch notwendig, bewusst subjektive Entscheidungen und Annahmen zu wählen, um Muster aufzudecken. Man spricht auch von numerischer bzw. zählender und graphischer „Detektivarbeit“. (Tukey 1977)

Ein bekanntes (Ur-) Beispiel aus der Literatur, das den Kriterien explorativen Vorgehens entspricht, ist die Infografik von Charles Minard aus dem Jahr 1813:

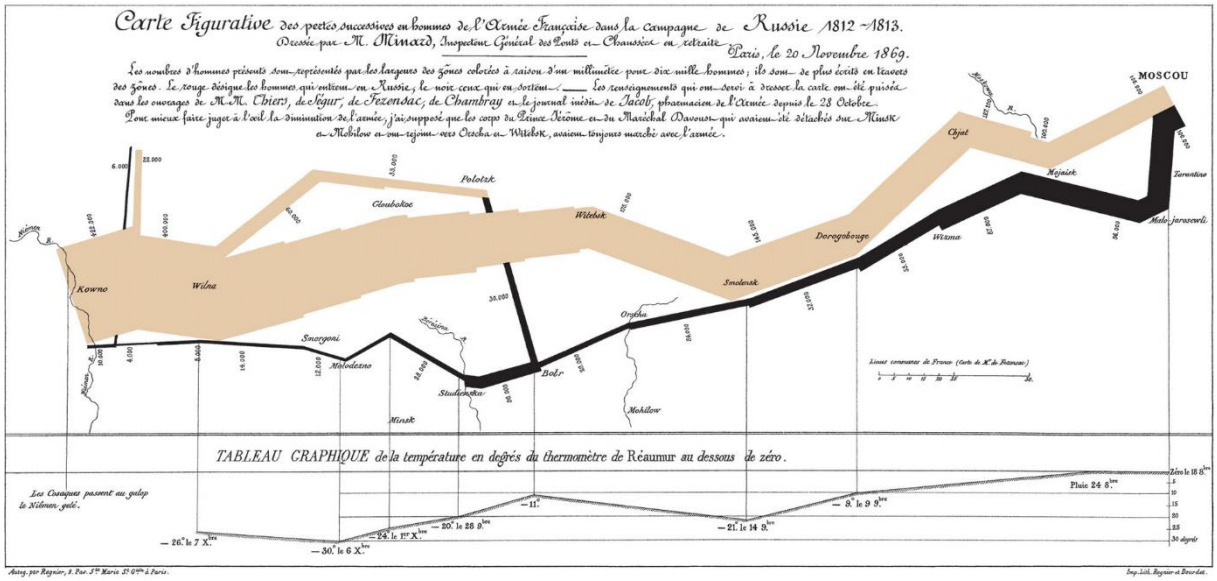


**Abbildung 3: Analyse des Russlandfeldzuges Napoleons 1812-183 (Minard, M. 1869)**

Hier werden verschiedenste Datenmerkmale miteinander verknüpft und dargestellt. Neben der Anzahl an Soldaten werden Ort- & Zeit, Temperatur, und Reiserichtung als Datenpunkte miteinander verknüpft und graphisch dargestellt. So kann vermutet (Hypothese) werden bzw. wird impliziert, dass mit dem immer stärkeren Absinken der Temperatur (Kausalität) ab dem Herbst in Moskau auch die Niederlage und der Rückzug der französischen Truppen eingeläutet wird.

In der heutigen Zeit wird diese Art von Darstellung als Sankey - Diagramm verstanden, bei dem Datenausprägungen zu z. B. Energieflüssen und deren -verlusten (bauteilabhängig in

technischen Systemen) und zeitlich bzw. in Form von Volumen – bzw. Massenströmen dargestellt werden.

Andere Visualisierungsmöglichkeiten sind auch einfache Balkendiagramme, Boxplots und Zeitreihen, sowie Stamm & Blatt Diagramme, Chernoff Gesichter, Mosaicplots, Parallelkoordinatenplots, welche das sogenannte *Highlighting,* also dem klar strukturierten, gezieltem Vermitteln von Informationen mit sich bringen.(Tukey 1977)

## 3.4 Verortung der Text Mining Verfahren in den Datenwissenschaften (Data Science)

„We are data rich, but information poor.“

– Jiawei Han (Han und Kamber 2010)

Wie in Kapitel 3.2 angedeutet, entsteht zur Zeit eine erhebliche Menge an Daten, die sowohl strukturiert in Form von Datenbanken als auch unstrukturiert in Form von Text-, Bild- und Tondaten im virtuellen Raum mehr oder weniger frei zur Verfügung steht. Die Informationstechnisch (IT-) gestützte Auseinandersetzung mit dem als „Big Data“ bekannt gewordene Phänomen wird als Data Science bezeichnet. (Mohanty 2015)

Diese heterogenen Daten können von verschiedenen wissenschaftlichen Disziplinen durch ausgewählte, oft generische Methoden und Werkzeuge erhoben und untersucht werden. Es wird daher nicht mehr nur von reinen statistischen Methoden gesprochen, sondern der Synthese der traditionellen Wissenschaften, die durch die Data Science unterstützt wird. Algorithmisierte Datenverarbeitung findet heute u. a. Anwendung in Computer Science, Linguistik, Ökonometrie, Soziologie und anderen Wissenschaften. (Dhar 2013)
In einer Studie zum Bedarf und den Anforderungen an Ressourcen für Text und Data Mining (TDM) konnte herausgefunden werden (Sens et al. 2015), dass Forscher aus verschiedensten Wissenschaftsbereichen TDM – Methoden nutzen und der Bedarf als steigend und der Nutzen als sehr hoch eingeschätzt wird. Problematisch sei allerdings die datenschutzrechtliche Lage, die bisher nicht ausreichend geklärt sei. (Coulet 2018)

Einen nicht nur ökonomisch wesentlich größeren Impact haben von Kunden generierte Daten, die heute fortlaufend von IT - Unternehmen erfasst werden, deren Geschäftsmodelle häufig auf aktiv (z. B. Rezensionen) oder auch passiv geteilten Nutzerdaten (z. B.: Bewegungsspuren auf Webseiten oder Sensordaten vom Smartphone bauen. (Sievers 2017))

Das generische Vorgehen in der „Data Science“ oder oft auch „Knowledge Discovery in Databases“ (KDD) beschreibt eine teilautomatisierbare, explorative Analyse und das Modellieren von großen Datenbeständen. Es wird als organisierter Erkenntnisprozess dargestellt, in dem

valide, neue, hilfreiche und verständliche Muster aus großen und komplexen „Datenseen“ extrahiert werden. Data Mining sei dabei der Kern des KDD Prozesses, bei dem mithilfe von Algorithmen Daten in Modellen abgebildet werden, aus denen bis zu dem Zeitpunkt unbekannte Muster und Informationen erkannt werden können, welches sich zu Wissen entwickeln kann. (Maimon und Rokach 2010)

Abbildung 4 zeigt die grundsätzliche strukturierte Vorgehensweise zur Erkenntnisgewinnung aus größeren Datenmengen. Innerhalb der Stufen gelten zahlreiche Entscheidungsketten:

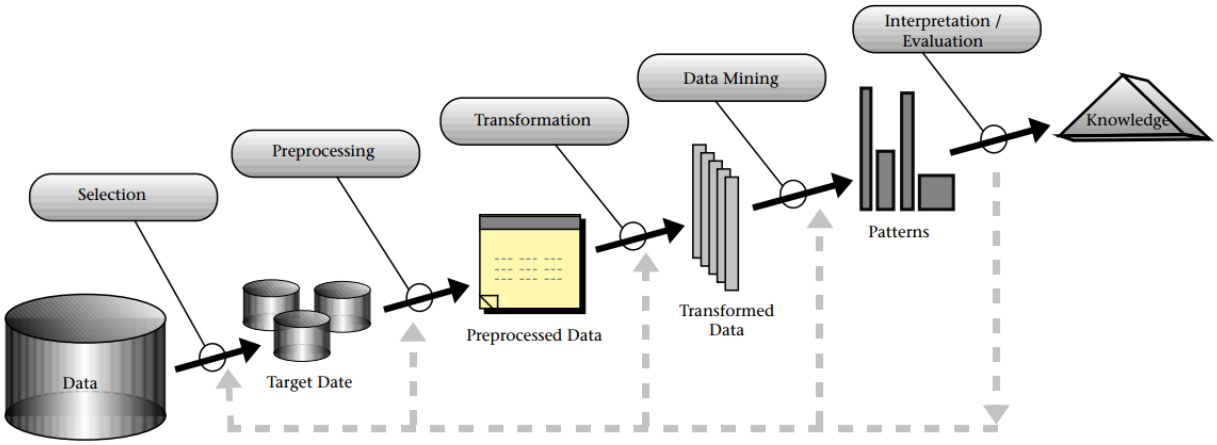


**Abbildung 4: KDD - Prozess in fünf Schritten (Fayyad et al. 1996)**

| | |
|---|---|
| Selektion: | Auswahl verfügbarer Daten und Übertragung in die Datenbank entsprechend des definierten Untersuchungsziels |
| Datenvorverarbeitung: (Pre - Processing) | Datenbereinigung, Fehlerbeseitigung und Ergänzung fehlender oder widersprüchlicher Daten |
| Transformation: | Datenumwandlung in Abhängigkeit des gewählten Data Mining Verfahren (z. B. Klassenbildung, Umwandlung in Booleanwerte) |
| Data Mining: | Mustersuche und Entwicklung von Modellen |
| Interpretation / Evaluation: | Güteüberprüfung der Resultate |

Dieser Prozess und die Subprozesse werden iterativ durchgeführt, bis brauchbare Ergebnisse entstehen. In der Literatur wird Data Mining vielfach unterschiedlich definiert. Gemeinsam ist aber, dass der Begriff als Schnittmenge von Disziplinen wie Künstliche Intelligenz, Statistik, Maschinelles Lernen, Soft Computing, Datenbanken und Visualisierungstechniken zu verstehen ist und in Wechselwirkung mit denselben steht, bzw. sich den Methoden in Form von Algorithmen und Arbeitsvorschriften zur Verarbeitung der Daten bedient. (Han und Kamber 2010; Zaki et al. 2010; Drees 2016; Morik et al. 2012; Miner et al. 2012; Khasanah et al. 2018)

Als wichtiges Gebiet soll außerdem das Information Retrieval (IR) genannt werden, das von Erkenntnissen des Data Mining profitiert. Methoden in diesem Gebiet bilden beispielsweise die grundlegende Funktionsweise für effiziente Suchmaschinen oder generell der Präsentation von Daten und Informationen für den Anwender. Verfahren wie Vektorraum Modellierung (engl. Vector Space Models (VSM)) und Clusteranalysen finden hier Anwendung, um z. B. ähnliche Suchergebnisse zu gruppieren. Text Mining und Web Mining sind zwei Spezialisierungen des Data Mining, die eng mit dem Information Retrieval verbunden sind. (Grossman und Frieder 2004)

Im Gebiet des Text Mining gibt es Methoden und Vorgehensweisen, die in allen Disziplinen verwendet werden, und welche generisch auf das jeweilige Untersuchungsziel anwendbar sind. So findet man beispielsweise Machine Learning Methoden im Bereich des Natural Language Processing mit denen, basierend auf statistischen Verfahren Daten aufbereitet und analysiert werden, um eine Problem- oder Fragestellung zu bearbeiten. Dies macht folgende Abbildung sichtbar:

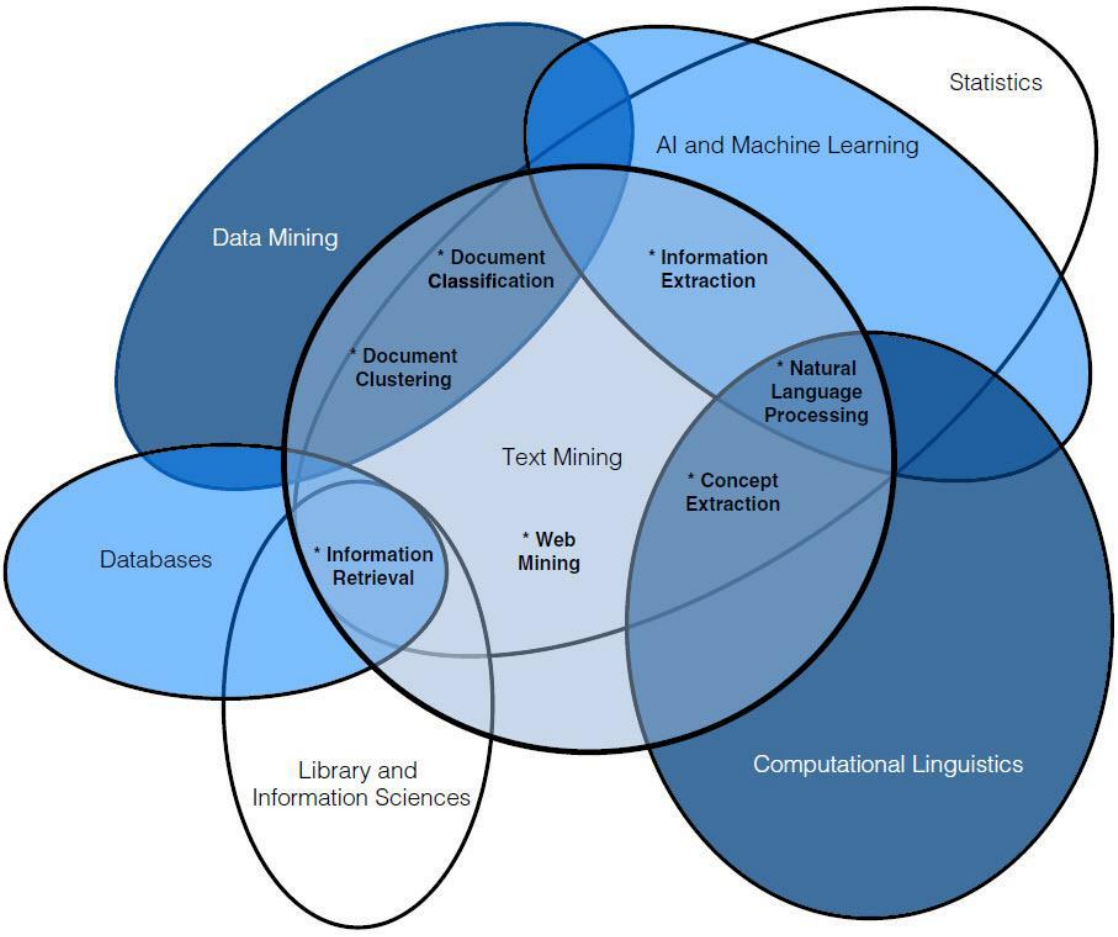


**Abbildung 5: Venn Diagramm, das die Überschneidungen von Text Mining und sechs verwandten Disziplinen visualisiert (Miner et al. 2012)**

Es wird deutlich, dass Text Mining zum einen sehr viele Wissenschaftsbereiche betrifft, aber auch, dass die geltenden Methoden, einmal in der Struktur und Anwendung verstanden und angewendet, auch in vielen anderen Bereichen anwendbar sein können und sich aufgrund der allgemein gehaltenen Programmierungen sehr gut Synergien zwischen den Disziplinen ergeben können. In Kapitel 4.2 werden aufgrund der Vielzahl der Möglichkeiten, nur die für das Untersuchungsziel dieser Arbeit relevanten Prozesse und Algorithmen genauer erläutert.

Forschungsschwerpunkte in dem Gebiet Text Mining findet man z. B. an der Stanford University und anderen kalifornischen Universitäten, die in geographischer Nähe zum sogenannten Silicon Valley und damit auch den IT - Konzernen Google, Amazon, Facebook & Apple (GAFA) liegen. Es sind ebenfalls Veröffentlichungen aus China, Indien und Deutschland (Institut für Maschinelle Sprachverarbeitung - Universität Stuttgart) zu finden, die oft sprachenunabhängig die Entwicklungen vorantreiben. (Allahyari et al. 2017)

Zwei allgemeine Probleme, die es bei der Anwendung der Methoden des Textmining bewältigen gilt, sind zum einen die Mehrdeutigkeit von Wörtern, die oftmals auch nur im Satzzusammenhang eindeutig wird. Zum anderen, dass es Wörter gibt, die die gleiche Bedeutung haben und synonym verwendet werden (Gaikwad et al. 2014). Außerdem muss in jedem Fall ein individuelles Pre - Processing für die neu zu untersuchenden Dokumente bzw. des entstehenden Corpus stattfinden.

## 3.5 Methoden zur Datenextraktion

Beim Text & Data Mining werden beispielsweise auf Daten aus dem World Wide Web, die in sogenannter eXtensible Markup Language (XML) in einer Baumstruktur auf vielen Webseiten zu finden ist, zugegriffen. Die XML – Struktur erlaubt es auf Host- bzw. Betreiberseiten Daten über Platzhalter auf der Seite zu platzieren. Die Sprache ist sowohl von Menschen als auch von Maschinen lesbar. (W3C 2018)

Mithilfe der XML Path Language (Xpath) oder alternativ auch Befehlen aus der Cascading Style Sheets (CSS) - Selektoren können die Pfade der Platzhalter und deren Inhalte addressiert und lokalisiert werden. Diese Pfade sind mit der Windows Ordnerstruktur vergleichbar, werden unterteilt in Knoten: *parent, children, siblings, ancestors, descendants* und können mit einheitlich geltender Syntax angesprochen werden. (W3C 2014)

Veranschaulicht werden kann das Vorgehen anhand eines Codebeispiels:

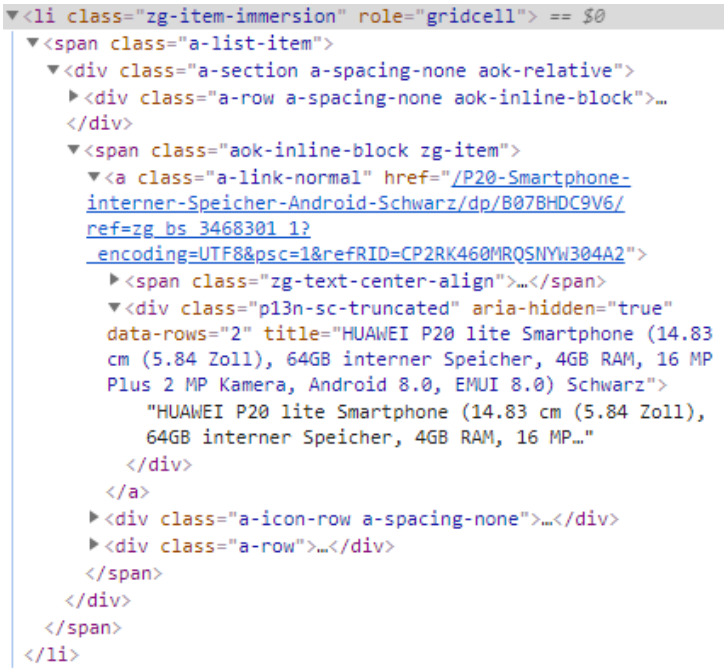


**Abbildung 6: Beispiel für die Struktur eines XML - Baums (Codeausschnitt von amazon.com)**

In Abbildung 7 wird die Struktur eines solchen Baumes sichtbar. Durch ein Aneinanderhängen der Pfadbausteine, wie *//li/span/div/span/a/div/text()* oder generischer *//div[contains(@class, "p13n-sc-truncated")]/text() (suche div container, der eine Klasse „p13n-sc-truncated“ enthält und gib den Text wieder)* kann beispielsweise der in Anführungszeichen, schwarz gedruckte Text oben ausgewählt werden. Auf dieser Beispielseite hat der List - Tag *<li>* 50 Geschwisterknoten – entsprechend der 50 Produkte auf dieser Seite - auf derselben Hierarchieebene, wie in Abbildung 7 zu sehen.

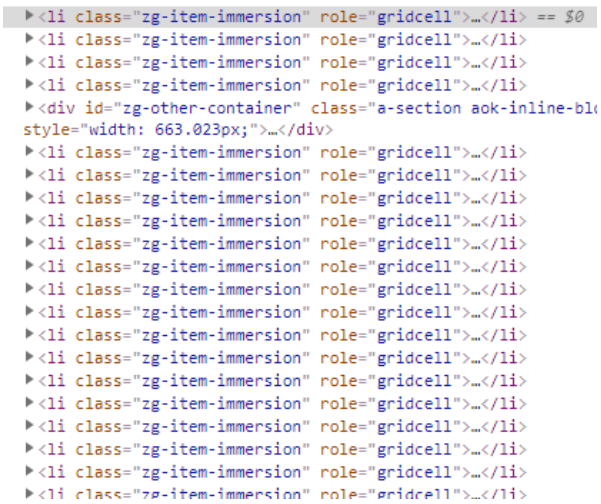


**Abbildung 7: Liste der der 50 Geschwisterknoten <li>, jedes Element entspricht einem Gerät (Codeausschnitt von amazon.com)**

Auf diese Weise können prinzipiell alle gewünschten Informationen, in diesem Fall auch Features genannt, direkt und knotenspezifisch adressiert werden und als Liste von sogenannten Response - Objekten z. B. mithilfe des Scrapy Frameworks extrahiert werden.

Das Scrapy Framework ist ein open source, auf Python basierendes Bezugssystem, mit dem effizient, einfach und auf sehr vielfältige Weise Daten von Webseiten heruntergeladen werden können. Sogenannte „spider“ laufen entsprechend der programmierten Selektoren, die schleifenweise aufgerufen werden, über die entsprechenden Webseiten und legen die Response - Objekte in Zwischenspeicher und nach weiterer Formatierung in Datenbanken ab. (Dale 2016; scrapinghub.com 2014; Kouzis-Loukas 2016)

Es können entlang des gesamten XML - Baumes einer Website relevante Daten extrahiert werden. Denkbar und für das Forschungsziel relevant sind auf einer Amazon - Seite z. B. Amazon Standard Identification Number (ASIN), das Reviewdatum, der Verkaufsstart auf Amazon, der Autor (ReviewAuthor), der Rezensionstitel (ReviewTitle) und der Reviewtext (ReviewBody). Diese Features können in diesem Kontext in eine Structured Query Language (SQL) Datenbank in beliebigen Datenformate (.csv, .json oder .xml) gespeichert werden.

Neben dem Scrapy Framework existieren außerdem einige Online Tools wie Octoparse.com oder aber das Python Paket BeautifulSoup, welches aber im Vergleich zu Scrapy wesentlich weniger flexibel und robust ist.

## 3.6 Verfahren des Natural Language Processing

Spezifische Verfahren zur Textanalyse hält das sogenannte Natural Language Processing (NLP) bereit. Wie in obiger Abbildung zu sehen ist, handelt es sich dabei um ein Wissenschaftsfeld, das sich im Bereich der Informationswissenschaften und -technik mit Einflüssen aus dem Bereich der Sprachwissenschaften und der Computerlinguistik entwickelt hat, in dem auch Methoden aus dem Bereich der Künstlichen Intelligenz entwickelt und adaptiert werden. Ein Ziel des NLP ist es, Anwendungen zu entwerfen, die die Interaktion zwischen Mensch und Computer vereinfacht, indem menschliche Sprache als Medium genutzt wird. (Pustejovsky und Stubbs 2012)

Es wurden beispielsweise im Laufe der Zeit spezifische Verfahren entwickelt, um Textdaten IT-gerecht aufzubereiten. In dem Leitfaden über Big-Data des Bundesverbandes Informationswirtschaft, Telekommunikation und neue Medien e.V. (BITKOM) werden die Aufgaben des NLP in **dokumenten**-, sprach- und domänenspezifische Verarbeitung aufgeteilt. In dem ersten Gebiet finden sich Teilaufgaben wie der „**Bereinigung und Normalisierung der Daten**“ und der

„Anreicherung der Dokumente mit Metadaten“. Des Weiteren werden Aufgaben wie „Sprachenerkennung, Satzsegmentierung, **Tokenisierung**, Wortstammreduktion (stemming), Lemmatisierung, Part - of - Speech Tagging, Parsing, Koreferenzauflösung“ und der „Erkennung genereller Eigennamen“ genannt. Zuletzt, im Bereich der spezifischen Domänen bzw. Anwendungen stehen Aufgaben wie domänenspezifische Eigennamenextraktion, **Stoppwortlisten**, **Topic-Modell**, Faktenextraktion“ und „Relationsextraktion“. (Weber 2014)

Die **fettgedruckten** Aufgaben bzw. Verfahren werden im Rahmen dieser Arbeit zunächst erklärt, in einer entwickelten Methodik angewandt und anschließend diskutiert.

In einigen dieser Aufgaben werden auch Algorithmen aus dem Bereich des Maschinellen Lernens genutzt. Eine grobe Struktur zur Einordnung der üblichen Techniken ist in der folgenden Abbildung gezeichnet:

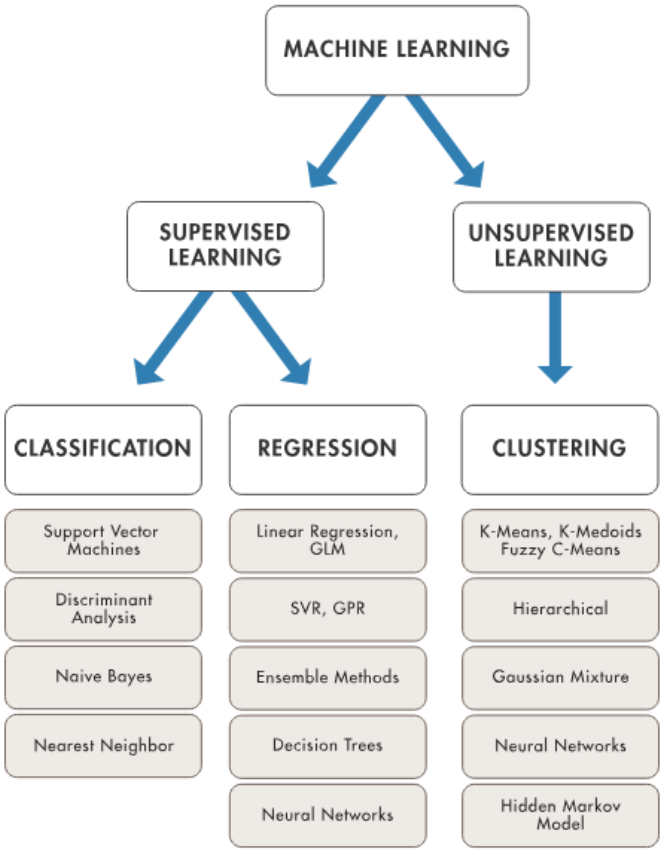


**Abbildung 8: Charakteristika und Techniken des Maschinellen Lernens (MathWorks.com 2018)**

Demnach wird maschinelles Lernen in überwachtes und nicht - überwachtes Lernen eingeteilt, manchmal wird zusätzlich in bestärkendes Lernen unterteilt (Tokic 2013). Überwachtes Lernen wird genutzt, wenn Eingabedaten und Abhängigkeiten der Daten (Ausgabedaten) bekannt sind

und daraus ein Modell zu entwickeln, das möglichst verlässliche Vorhersagen über neue Eingabedaten treffen soll. So kann z. B. zur Vorhersage des Risikos des Erkranken an einer bestimmten Krankheit ein Modell trainiert werden, in welches Anamnesedaten von vielen erkrankten und gesunden Patienten integriert werden, um auf den Daten basierend prognostizieren zu können, ob ein bisher nicht erfasster Patient einer erhöhten Gefahr der Erkrankung unterliegt. (MathWorks.com 2018)

Bei unüberwachtem (engl. „unsupervised“) Lernen werden häufig durch mathematische Operationen (z. B. Gradientenabstiegsfunktion), lokale und globale Extrempunkte gesucht, sodass sich mit steigender Anzahl an Iterationen bestimmte Unterschiede zwischen den Datenpunkten herauskristallisieren und sich bestenfalls Gruppen bilden, deren jeweilige Merkmalsträger Gemeinsamkeiten aufweisen (Zaki et al. 2010, S. 4–15).

Der in dieser Arbeit verfolgte Ansatz ist im Bereich nicht - überwachtes Lernen mit Neuronalen Netzen für Sprachmodellierung (NNLM) (engl. „Neural Networks Language Models”) verortet (Jansen 2018), welche als effizienter gelten als bis dahin bestehende Verfahren oder mit denselben als sogenannte „downstream tasks“ kombiniert werden können. Ausgewählte Beispiele dafür sind Topic Modeling oder k – Means - Clustering. Topic Modeling beschäftigt sich mit der automatisierten Erstellung von einer bestimmten Anzahl von Überschriften in einem Corpus, während k - Means - Clustering eine Einteilung aller Merkmalsträger in k verschiedene Gruppen vornimmt. (Liu et al. 2015)

### 3.6.1 Repräsentation von Wörtern als Vektoren (Word Embeddings)

„You shall know a word by the company it keeps” (J.R. Firth 1957)

Diese Aussage entspricht dem Kern der sogenannten distributionellen Hypothese (engl. „Distributional Hypothesis“) (Harris 1954), dass unterschiedliche Wörter mit ähnlicher Umgebung auch eine ähnliche Bedeutung haben. Mit dem 2013 veröffentlichten word2vec Algorithmus ist es möglich geworden, dies effizient auf Corpora von mehreren hundert Millionen Wörtern Umfang anzuwenden (Mikolov et al. 2013).

Bis zum Erscheinen dieses von Google veröffentlichten Modells wurde Natural Language Processing hauptsächlich durch die oben genannten Verfahren und Techniken der einzelnen Wörter vollzogen, meistens ohne den Kontext der einzelnen Wörter zu berücksichtigen. *Apple* und *Samsung* bekämen beispielsweise individuelle IDs, aber deren Zusammenhänge und beschreibenden Wörter, welche mit den Smartphone Herstellern in Verbindung gebracht werden, bleiben bei der Interpretation durch reine Bag of Words (BoW) Modelle, in denen z. B. Worthäufigkeiten festgehalten werden, verborgen. Hier setzt das in verschiedenen Aufgaben im Be-

reich der Computerlinguistik vielfach verwendete Umwandeln der Wörter in Vektor- bzw. Matrixdarstellung, im ersten Schritt mittels sogenanntem „One - Hot Encoding“ an. Damit wird versucht, wichtige syntaktische und semantische Informationen der Wörter untereinander zu quantifizieren. (TensorFlow.org 2018)

Nach der Transformation in Vektoren können mithilfe klassischer mathematischer Operationen der Linearen Algebra, z. B. durch Errechnen von Vektorprodukten oder Kosinus – Ähnlichkeiten, Aussagen über die Abhängigkeit der Wörter getätigt werden. Da die Erläuterung dieser Operationen im Rahmen dieser Arbeit nur rudimentär erfolgen kann, sei z. B. auf ein Machine Learning ausgerichtetes Lehrbuch des Massachusetts Institute of Technology (MIT) verwiesen, in dem tiefergehende Fragestellungen, sowie algebraische Grundlagen anschaulich behandelt werden. (Goodfellow et al. 2016)

Verschiedene auf dem word2vec Algorithmus aufbauende Publikationen, versuchen neben neuen Erkenntnissen auch die Funktionsweise des Algorithmus besser zu veranschaulichen und greifbar zu machen. (Goldberg und Levy 2014; Rong 2016; Jansen 2018)

Basierend auf den Regeln der Linearen Algebra, kann demnach ein Vektorraum bzw. daraus folgend eine Wertematrix mit beliebig vielen *m* Wörtern (Zeilen) und *n* Dimensionen (Spalten) erstellt werden. Jedes Wort im Corpus wird im Modell von einem Vektor repräsentiert, der sich in n Richtungen und verschiedene Werte annehmen kann, z. B.:

*Wort* (0,1,0,1,0,0,1) mit n = 7

Diese Richtungen (= Dimensionen) können grundsätzlich verschiedene Eigenschaften, auch Features genannt, repräsentieren, z. B.:

1. Vorhandensein des Wortes in dem Teil des Corpus bzw. Dokuments (0 oder 1) z. B. beim oben genannten „one – hot -encoding“ in einer m x m - Matrix
2. Nähe zu einem anderen Wort, abhängig von dem betrachteten Umfeld ( - x bis + x)
3. Vorkommen von bestimmten Mustern (z. B. 2 oder mehr aufeinanderfolgende Zeichen, Wörter)
4. Frei zu wählender Merkmalsträger (z. B. Substantiv, Verb, Adjektiv, Verb, blau, rot, kleiner 1 m, größer 1 m)
5. Ggfs. weitere (Mitra und Craswell 2017)

Um den Fokus auf den zweiten Punkt zu lenken, und damit die Vorgehensweise von Word2Vec an einem Beispiel zu erläutern, wird angenommen (Mitra 2016):

Dokument 1: [„seattle“, „seahawks“, „jerseys“]
Dokument 2: [„seattle“, „seahawks“, „highlights“]
Dokument 3: [„denver“, „broncos“, „jerseys“]
Dokument 4: [„denver“, „broncos“, „highlights“]

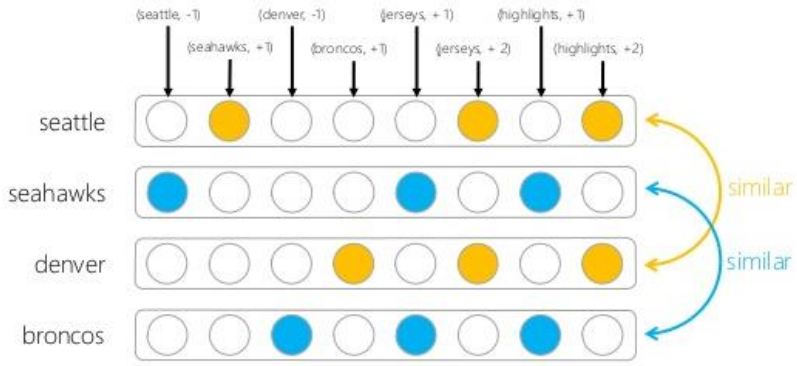


**Abbildung 9: Visualisierung des Konzepts einer Wertematrix (engl. co - occurence matrix) bei der Anwendung von word2vec (Mitra 2016)**

Obige Abbildung zeigt das Vorgehen bei der Erstellung einer sogenannten „co - occurrence“ Matrix. Zunächst werden alle im Corpus vorkommenden Wörter (m) in einer Spalte links aufgelistet. Die restlichen Spalten bzw. Dimensionen (n) ergeben sich aus der Entfernung der jeweiligen Wörter voneinander, also ob ein Wort gar nicht oder mit ein oder zwei Wörtern Abstand mit dem aktuellen Wort genannt wurde. Bei einem Vergleich fällt dann auf, dass die in dem Beispiel jeweils blauen und orangefarbigen häufig dieselben „Nachbarn“ haben, also in gewisser Weise ähnlich sind. (Jansen 2018)

Auf diese Weise kann je nach Größe des Corpus und des Vokabulars eine sehr große m x n - Wertematrix entstehen, abhängig von der Anzahl der verschiedenen Wörter im Corpus und der Größe des betrachteten Umfeldes. Da der Umgang damit sehr rechenaufwendig ist, werden die n Dimensionen mit Werten von 0 oder 1 mittels Exponentialfunktion (Softmax Funktion) normalisiert. Jeder Dimension wird demnach ein eindeutiger Wert zwischen 0 und 1 entsprechend seiner Codierung zugeordnet. (Rong 2016)

Neben weiteren Schritten auf mathematischer Ebene, wie der Entscheidung zwischen zwei Verfahrensvorschriften innerhalb des Algorithmus zur Abfrage und Bildung der Kontextvektoren (Continous Bag of Words oder Continous (CBOW) oder Skip - Gram und Leistungsverbesserungen (Hierarchical Softmax, Negative Sampling), auf die hier nicht weiter eingegangen werden können, werden folgende zur Modellbildung notwendige Parameter vom Anwender vordefiniert werden (Jansen 2018):

- Kontextgröße: Anzahl von Wörtern vor und hinter dem selektierten Wort.
- Mindestnennungen: Anzahl der Nennungen, ab der das Wort mit in das Vokabular des Modells aufgenommen wird.
- Teilprobenfrequenz: Besonders häufig vorkommende Wörter werden geringer gewichtet (empfohlener Wert $10^{-5}$).
- Größe des Negative Samplings: Effizienzsteigernder Ansatz, s. Mikolov 2013
- Anzahl der Dimensionen: Tendenziell bringen mehr Dimensionen bessere Ergebnisse
- Anzahl von Wiederholungen: Abhängig von Modellgröße und Rechenkapazität

Mit dem berechneten Modell, auch Vector Space Model (VSM) (Dubin 2004) genannt, wird nun jedes Wort im Vokabular durch einen bestimmten Vektor mit n Dimensionen repräsentiert. Mithilfe der Kosinus Ähnlichkeit kann die Ähnlichkeit zwischen zwei Vektoren und damit Wörtern berechnet werden:

$$sim(\vec{v}_i, \vec{v}_j) = cos(\vec{v}_i, \vec{v}_j) = \frac{\vec{v}_i^{\intercal} \vec{v}_j}{\|\vec{v}_i\| \|\vec{v}_j\|} \quad (1)$$

So können z. B. nicht nur die zehn ähnlichsten, also nächstliegenden Vektoren zu einem Input-Vektor ausgelesen werden, sondern auch durch Vektoraddition und -subtraktion die Güte des Modells überprüft werden. In der ursprünglichen Veröffentlichung wurden mehrere 100 Millionen Wörter von Wikipedia modelliert und es konnten verschiedenste Wortanalogien „errechnet“ werden, die qualitativ auf ihre Sinnhaftigkeit überprüft werden kann. Eine bekannte und viel zitierte Operation ist nach Mikolov et al. 2013:

$$v_{king} - v_{man} + v_{woman} = v_{queen} \quad (2)$$

Seit der Entwicklung von word2vec hat es viele Publikationen gegeben, die auf dem Prinzip der Vektorrepräsentationen aufbauen. Diese sollen grundlegend der Vollständigkeit erwähnt werden, aber finden in dieser Arbeit keine Verwendung, da hier zunächst überprüft werden soll, welche Ergebnisse erzielt werden, wenn der Kontext von ausgewählten Wörtern in zwei verschiedenen Corpora analysiert wird.

An der Stanford University wurde das für bestimmte Fälle effizientere GloVe (kurz für: Global Vectors) für Vektorrepräsentationen entwickelt, welches den Kontextvektor eines Wortes mithilfe von Häufigkeiten erhebt (Pennington et al. 2014). Außerdem wurde das word2vec Modell wenig später durch das doc2vec (auch: sent2vec oder paragraph2vec) Modell erweitert, bei

dem zusätzlich zur Wortebene auch die gesamte Dokumentenstruktur mitberücksichtigt und entsprechend ausgewertet wird (Le und Mikolov 2014). Das von Facebook veröffentlichte FastText Modell arbeitet hingegen auf Buchstabenebene und erzeugt Vektoren auf Basis von Buchstabenpaaren oder -folgen, sogenannte n - grams. So kann auf Basis eines zur Verfügung gestellten Modells (überwachtes Lernen) wesentlich schneller eine Klassifizierung von Wörtern stattfinden und es können wesentlich genauere Vektoren für ein spezifisches Wort ermittelt werden. Dies ist vor allem interessant für Sprachen in denen die Wörter sich je nach Konjugation oder Deklination stark verändern (Bojanowski et al. 2016). Aufbauend auf den bis dato veröffentlichten Arbeiten, wurde letztes Jahr das starSpace Modell vorgestellt, mit dem analog zum FastText Modell willkürliche kleine Bausteine „vektorisiert“ werden können und in Bezug zueinander gesetzt werden. Anwendung kann dies bei einer Vielzahl von Aufgaben, wie Suchmaschinenranking auf Basis von Suchbegriffen oder Empfehlungen aufgrund von konsumierten Produkten und Dienstleistungen finden. (Wu et al. 2017).

Seit diesem Jahr bietet das Unternehmen Amazon auch die Dienstleistung „Sagemaker“, an mit der Word Embeddings und andere Methoden auf Höchstleistungsservern vollzogen werden können. (Gupta und Khare 2018)

### 3.6.2 Dimensionsreduzierung

Im Feld des Information Retrievals (IR) liegen neben dem word2vec Algorithmus z. B. auch die Verfahren t – Stochastic Neighbor Embedding (t-SNE) und Term Frequency – Inverted Document Frequency.

Ein Wordvektor eines Word2Vec Modell wird aus bis zu 1.000 abstrakten Dimensionen repräsentiert, deren Bedeutung aufgrund der Verschachtelung von verschiedenen Operationen für den Menschen nicht mehr einfach nachzuvollziehen ist und auch die Interpretation derselben bisher zu keinem sinnvollen Ergebnis führt. (Wilson 2016; Szegedy et al. 2013)

Dennoch ist es möglich, mithilfe des t-SNE Algorithmus die Dimensionen auf zwei- oder dreidimensionalen Raum zu reduzieren. Da dieser Prozess sehr komplex ist und in dieser Arbeit nur zur Orientierung gedient hat, wird an dieser Stelle an die entsprechenden Publikationen verwiesen. (van der Maaten und Hinton 2008; van der Maaten 2014)

Dies ist Softwaremäßig beispielsweise mit dem python Natural Language Toolkit und dem pyData Stack möglich. Eine sogenannte stand-alone Lösung bietet Google mit dem tensorflow® Projector an. Mit diesem kann ein n-dimensionales Vektormodell mit den entsprechenden Tags bzw. Beschriftungen visualisiert werden. Dies kann zur visuellen Exploration von nahe beiei-

nander liegenden Wortvektoren genutzt werden und dem Sichtbarmachen von Clustern dienen. Entsprechend der Parameter der t-SNE Algorithmus können Größen wie Perplexität, Dimensionalität und Lernrate, abhängig von der Größe des Modells eingestellt werden. Aus der Anleitung im Programm geht heraus, dass die Lernrate analog zur Größe des Modells eingestellt wird und die Perplexität Werte zwischen 5 und 50, ebenfalls abhängig von der Dichte der Vektoren, also Größe des Modells. Es wird iterativ vorgegangen, bis sich brauchbare Ergebnisse wörtlich „herauskristallisieren“. Desweiteren besteht die Möglichkeit, ähnlich wie mit der *most_similar()* Methode, einen bestimmten Wortvektor anzupeilen, und eine bestimmte Anzahl, meist die nächsten 100 Nachbarvektoren hervorzuheben.

Ein großes Problem dabei ist, dass die Ergebnisse nicht immer wiederholbar sind, da die Dimensionsreduzierung immer auch einen Zufallsanteil in sich trägt und Missinterpretationen möglich sind. (Wattenberg et al. 2016)

### 3.6.3 Messung der Relevanz und Besonderheit von Wörtern

Im Bereich der Informationsgewinnung (engl. Information Retrieval) in Datenbanken, wird zur Beurteilung von Relevanz von Wörtern u. a. das sogenannte Term Frequency - Inverse Document Frequency – Modell (TF - IDF) genutzt. Das Modell besteht aus zwei Komponenten, die miteinander multipliziert werden. (Fuhr 2010)

Die Term Frequency bzw. relative Vorkommenshäufigkeit beschreibt die Häufigkeit des abgefragten Wortes im Verhältnis zur Gesamtanzahl von Wörtern im aktuellen, „lokalen“ Dokument. Der Wert für die Inverse Document Frequency beschreibt die logarithmierte Häufigkeit, die ein Wort in allen Dokumenten eines Corpus auftritt: (Leopold und Kindermann 2002)

$$idf_k = \log \frac{N}{df_k} \tag{3}$$

Mit der Gesamtzahl an Dokumenten (N) und der Anzahl der Dokumente, die das Wort k enthalten ($df_k$).

In der Literatur wird auch von einer Messung der „Besonderheit“ (engl. specificity) bzw. Detailliertheit, in welchem ein Konzept von einem Wort repräsentiert wird, gesprochen. (Spärck Jones 2004) Es gilt außerdem als Möglichkeit der “globalen” Gewichtung eines Terms. (Robertson 2004) Dabei gilt die Annahme, dass ein Wort, das in vielen Dokumenten vorkommt, weniger Gewicht hat, als ein Wort, das nur in wenigen Dokumenten Erwähnung findet. Diese Vorgehensweise erlaubt nicht, Dokumente ins Verhältnis untereinander zu setzen oder auf diese Weise zusammenzufassen. (Garcia 2006)

Mit der Kombination der lokalen und globalen Gewichtung kann allerdings die Relevanz oder auch Besonderheit eines Wortes im Gesamtzusammenhang des Corpus gemessen werden. Eine weitergehende Methodenkritik erfolgt im Diskussionskapitel 6.2.2.

### 3.6.4 Überblick über weitere Tools und Verfahren im Bereich Data und Text Mining

Neben den bisher genannten auf Python basierten Tools (pyData Stack, NLTK etc.) existieren auch andere Data Science Tools, die auch mit den Programmiersprachen Java oder R genutzt werden können, um statistische Modelle und Mining Verfahren anzuwenden. Zu erwähnen sind hier Hadoop, Torch, Spark, Tensorflow, H2O oder Rapidminer. (Statworx 2018)

Die Entscheidung für Python und der Anaconda Navigator Umgebung basiert auf der großen Paketevielfalt und -tiefe und der sehr aktiven Online Community, die bei Fragen in der Entwicklungsphase sehr helfen kann.

# 4 Methodische Vorgehensweise

Die in Abbildung 10 dargestellte Vorgehensweise zeigt die auf Algorithmen und Tools des Text Mining basierende, im Rahmen dieser Arbeit, entwickelte und angewandte Methodik. Parallel zu diesem Vorgehen werden außerdem geringdimensionale Daten wie Produktpreis, Bewertung, Anzahl „hilfreich“, Rezensionslänge, u. a. ausgewertet.

**Abbildung 10: Modellhafte Darstellung der entwickelten Wasserfall - Methodik der vorliegenden Arbeit (eigene Darstellung)**

Auf Basis von Ökodesignrichtlinien für Smartphones und Elektrogeräte allgemein (Fachsprache: *ausfall, defekt*) und als sehr hilfreich gekennzeichneten Kundenrezensionen (Umgangssprache: *kaputt, schrott*) wird eine Liste aus Keywords erstellt, welche als Eingangsgröße für die Anwendung von Text Mining Methoden genutzt wird.

Zunächst wird die Existenz der selektierten Wörter und anschließend die Relevanz (TF - IDF) der Wörter im Gesamtkontext des generierten Textkorpus geprüft. Diese Vorgehensweise ermöglicht eine Sortierung und Priorisierung für die weitergehende Analyse. Anschließend können die relevantesten Keywords in einem auf dem Corpus basierenden Word2Vec Modell abgefragt werden, um Informationen z. B. zu den semantisch ähnlichsten Wörtern zu generieren.

Zuletzt werden die Ergebnisse mit den genannten Tools aufgearbeitet und visualisiert.

## 4.1 Extraktion von Keywords aus Ökodesigndirektiven

Um die Rezensionen auf Nachhaltigkeitsaspekte zu untersuchen, werden die Anforderungen an Komponenten und Produkte und deren Design sowie deren Formulierungen der in der Praxis genutzten Ökodesignrichtlinien durchgearbeitet. Häufig sinngemäß sich wiederholende Formulierungen und relevante Stichwörter, hier auch Keywords genannt, werden zunächst in eine Liste extrahiert und gesammelt.

Unter anderem wurden dabei untersucht:

| Ursprungsland: | Richtlinie: |
|---|---|
| • USA: | Institute of Electrical and Electronics Engineers' Ecodesign Toolkit (Jean et al. 1999) |
| • USA: | UL 110 Sustainability for Mobile Phones (Underwriters Laboratories Inc. 2018) |
| • Schweden: | Tjänstemännens Centralorganisation (TCO) Label (TCO Certified 2018) |
| • USA: | Electronic Product Environmental Assessment Tools (EPEAT) (Gillis und Brody-Heine) |
| • Deutschland: | Der Blaue Engel (RAL gGmbH 2013) |

Die Liste der Wörter wird darüber hinaus um weitere Schlüsselwörter der mit großem Abstand häufigsten, von Kunden als am hilfreichsten gewählte Rezensionen (sowohl sehr schlechte als auch sehr gute Rezensionen), ergänzt. Anschließend wird mittels einfacher Schleife *if <wort> in <liste> das* Vorhandensein und die Anzahl der Nennungen ermittelt und anschließend das TF - IDF Verfahren angewandt, um die Relevanz jedes Wortes im Kontext der ein und fünf

Sterne Rezensionen abzufragen. Diese zusätzliche Ergänzung um die umgangssprachlichen Ausdrücke ist notwendig, da die fachsprachlichen Keywords aus den Direktiven nach einem ersten Abfragen nur in geringen Maße wortwörtlich im Corpus aufzufinden sind.

## 4.2 Data & Text Mining: Datenbeschaffung, -verarbeitung & -visualisierung

Um den in Kapitel 3.4 eingeführten KDD-Prozess zu durchlaufen, soll dieser nun mit den im Rahmen dieser Arbeit benötigten Methoden und Tools bestückt und diese an entsprechender Stelle erklärt werden. Die Auswahl der Tools und Pakete basiert auf weitergegebenen Praxiserfahrung auf Entwicklerplattformen wie *stackoverflow.com* und *github.com*, auf denen Entwickler ihre Software open source zur Verfügung stellen und sich über Probleme austauschen und Hilfe angeboten bekommen.

**Benötigte bzw. in dieser Arbeit verwendete Software:**

| | |
|---|---|
| Basissoftware: | Anaconda Navigator, Sublime Text Editor |
| Selektion: | Scrapy Framework |
| Datenvorverarbeitung: (Preprocessing) | Jupyter Notebooks (Programmierumgebung): Pydata Stack (pandas, numpy, ntlk, regular expressions) |
| Transformation: | pandas, numpy, basic python |
| Data Mining: | gensim, word2vec, scikit learn |
| Interpretation / Evaluation: | Jupyter Notebook, Tensorflow Projector, Tableau, Microsoft Excel |

Anaconda Navigator bietet einen Grundstock an Softwarepaketen, die auf der Programmiersprache Python basieren und mit der Programme entwickelt werden können. Das kostenlose Programm bietet einen vereinfachten und intuitiven Einstieg. Außerdem können benötigte Pythonbibliotheken und Pakete (Scrapy, Pydata Stack) einfach installiert werden. Zum Erstellen und zur Bearbeitung von in Python geschriebenen Dateien bietet sich außerdem der generische Sublime Text Editor an, der eine benutzerfreundliche farbliche Gestaltung des Codes je nach Programmiersprache (Java, Python, C++, etc.) zur einfacheren Handhabung und Vorbeugung von Syntaxfehlern ermöglicht. (Whang und Maurer 2018)

Die Pakete zur Datenvorverarbeitung, Transformation und Data Mining und deren Implementierung werden in den folgenden Unterkapiteln vorgestellt. Zur Visualisierung der extrahierten Daten und Modelle können ebenfalls die Jupyter - Notebooks verwendet werden, sowie der

für den word2vec Algorithmus von Google entwickelte Tensorflow® Projector. Mit dieser Browserlösung werden die in dieser Arbeit generierten Modelle herunterdimensioniert und visualisiert. Außerdem werden die Daten über Microsoft Excel auch in herkömmliche (z. B. zwei- oder dreidimensionale) Diagramme eingepflegt.

### 4.2.1 Datengrundlage schaffen: Datenextraktion

Mithilfe des in Kapitel 3.5 vorgestellten Scrapy Frameworks wird ein individueller sogenannter Web Crawler entwickelt (Code siehe Anhang), der Inhalte vom Online - Händler Amazon.de und dessen ausgeschriebenen Top - Seller im Bereich „Handys & Smartphones“ konzentriert.

Anschließend wird die Datenbank in ein sogenanntes Jupyter Notebook eingelesen und die erforderlichen Schritte für

- die Auswertung geringdimensionaler, den Skalierungstypen der Statistik entsprechender Daten (z. B. durchschnittlicher Verkaufspreis der Produkte, Preisspannweite, Zusammenhänge zwischen Daten wie z. B. Länge der Rezension und wie häufig wird eine Rezension als hilfreich markiert) (Kapitel 5.2) und
- das Pre - Processing mittels Methoden des Natural Language Processings (NLP) angewandt und anschließender Modellbildung mittels „Word Embeddings“ (Kapitel 5.6)

Eine Kombination von Merkmalsausprägungen wie für eine explorative Datenanalyse üblich, soll anhand der Produktrezensionen im Rahmen dieser Arbeit geschehen. Ggfs. soll versucht werden, sehr schlechte oder sehr gute Rezensionen mit bestimmten anderen Merkmalen, z. B. Preis oder durchschnittliche Bewertung in Verbindung zu bringen. Die Ergebnisse können dann z. B. in Verbindung mit dem Verkaufsstart und dem Rezensionsdatum bzw. Produktalter gebracht werden, um herauszufinden, ob bestimmte Merkmale (z. B. defekt, Displaybruch, Akkuversagen) vermehrt nach einer bestimmten Zeit auftreten (z. B. nach Ausfällen nach der gesetzlichen Gewährleistungspflicht).

### 4.2.2 Aufbereiten der Rezensionen mittels Natural Language Processing Methoden

Der Schwerpunkt dieser Arbeit liegt auf der Analyse und dem Erkennen von Zusammenhängen in den Rezensionstexten, welche im Folgenden in einen im computerlinguistischen Sprachgebrauch genannten Corpus umgewandelt wird. Die einzelnen Rezensionstexte liegen als sogenannte Documents im Datentyp String vor. Um diese in sogenannte Tokens umzuwandeln, und damit maschinenlesbar zu machen, werden die Rezensionen mithilfe der Libraries *pandas, numpy* und *regular expressions (re)* mit beispielsweise folgender Befehlsfolge umformatiert:

```
1.  import pandas as pd
2.  train  =  pd.read_json("reviews_cleaned.json",  encoding = "utf-8")
3.  #Lade die Datei reviews_cleaned.json mit dem Codierungsstandard utf-8 (korrekte
4.  #Darstellung von Umlauten)
5.  print(train[(train.stars == 1 )].review_body)
6.  #Zeige Spalte mit allen Reviews mit der Bewertung 1 Stern an und weise sie der
7.  #Variable onestar_reviews zu.
8.  onestar_reviews = train[(train.stars == 1 )].review_body
9.  #Zeige die Länge, in diesem Fall der Liste und damit die Anzahl der Elemente in der
10. Liste
11. len(onestar_reviews)
12. #1799 1-Stern Rezensionen
```

**Abbildung 11: Einlesen der Datenbank und Zuweisen der Rezensionen in Variable "onestar_reviews" (eigene Darstellung)**

Das weitere Pre-processing geschieht mit folgender Befehlsfolge, basierend auf dem Code-Vorschlag einer Online-Coding-Competition Webseite (Kaggle Inc. 2014). Es werden zwei miteinander verkettete Funktionen (*def)* hinzugefügt, die aus den einzelnen Rezensionen eine verkette Liste bzw. „Liste von Listen“ erzeugt: Zum einen ist dies eine einfache Liste von Wörtern als Elemente in einer großen Liste, und zum anderen eine Liste von Rezensionen, in der die einzelnen Wörter der Rezensionen als Elemente der Listen aufgenommen werden. Bei der zweiten Variante bleibt die syntaktische Satzstruktur erhalten, aber die Wörter innerhalb der Rezensionen werden für den Rechner lesbar und quantitativ auswertbar gemacht. Die Funktion *re.sub(…)* tauscht alle Sonderzeichen, Punkte und Kommata gegen eine leeres Zeichen, *.lower()* wandelt alle Buchstaben in Kleinbuchstaben und *.split()* separiert alle Wörter in einzelne Elemente der Liste (auch: Tokenizing):

```
1.  from nltk.corpus import stopwords
2.  def review_to_wordlist(review, remove_stopwords=False ):
3.
4.      letters_only = re.sub("[^a-zA-ZäöüÄÖÜß]"," ", review)
5.  #Entferne alle nicht (^) Buchstaben und ersetze sie durch ein Leerzeichen (“ “).
6.      words = letters_only.lower().split()
7.  # Tausche alle Groß- durch Kleinbuchstaben (.lower) und erzeuge Token (.split)
8.  # Entferne alle Stop-Wörter, also alle Wörter, die im Paket „Natural Language
    Toolkit“ in der Liste words(„german“) zu finden sind, die keinen wesentlichen Inhalt
    beitragen (diskussionswürdig), je nach späterer Anwendung (engl. „downstream task“)
9.      if remove_stopwords:
10.         stops = set(stopwords.words("german"))
11.         words = [w for w in words if not w in stops]
12. # Zeile 11 steht kurz für, also wenn die Funktion aufgerufen wird, und
    remove_stopwords = True gesetzt wird, wird in der Stopwortliste
    „stopwords.words(“german“) gesucht, ob das aktuelle Wort verfügbar ist. Falls nicht,
    wird es in die Liste words hinzugefügt (.append):
        #for w in words:
        #   if not w in stops
        #   words.append(w)
13. # Speichere die Liste in der Variablen „words“
14.     return words
```

```
1.  # Importiere u.a. den “Punkt Tokenizer”
2.  import nltk.data
3.  #nltk.download() #ggfs. lade die fehlenden Pakete herunter
4.
5.  # Lade den punkt Tokenizer (erkennt beispielsweise . () „“)
6.  tokenizer = nltk.data.load('tokenizers/punkt/german.pickle')
7.
8.  # Ähnlich wie oben wird hier eine Funktion definiert, nur dass hier der Tokenizer
    die einzelnen Wörter und Satzzeichen formatiert.
9.  def review_to_sentences(review, tokenizer, remove_stopwords=False):
10.     raw_sentences = tokenizer.tokenize(review.strip())
11.     sentences = []
12.     for raw_sentence in raw_sentences:
13.         # Falls ein Satz keine Wörter mehr enthält, überspringe diesen:
14.         if len(raw_sentence) > 0:
15.             # Falls nicht, rufe review_to_wordlist(), um diese auf diesen einen Satz
                    anzuwenden.
16.             sentences.append(review_to_wordlist(raw_sentence, remove_stopwords))
17.
18.     return sentences
```

**Abbildung 12: Zwei Funktionen zum Pre - Processing der Rezensionen (eigene Darstellung)**

Folgendes Beispiel visualisiert die beiden Funktionen beispielhaft:

Teil der Rezension = [„Kurz zur Vorgeschichte. Ich hatte diverse Handys in meinem Leben. Ich beschränke mich daher auf die Geräte seit 2012. […]“]

words = [„kurz“, „zur“, „vorgeschichte.“ „ich“, „hatte“, „diverse“, „handys“, „in“, „meinem“, „leben“, „ich“, „beschränke“, „mich“, „daher“, „auf“, „die“, „geräte“, „seit“, „2012“]

sentences = [[„kurz“, „zur“, „vorgeschichte“], [„ich“, „hatte“, „diverse“, „handys“, „in“, „meinem“, „leben“], [„ich“, „beschränke“, „mich“, „daher“, „auf“, „die“, „geräte“, „seit“, „2012“]]

Durch diese Vorgehensweise können aus dem Text numerische Informationen generiert werden wie z. B. Länge bzw. Anzahl der Wörter und Sätze in einer Rezension oder auch wie häufig sich bestimmte Wörter wiederholen. Siehe dazu Kapitel 5.5 (Relevanz von Keywords).

Diese Vorverarbeitung oder Pre - Processing kann für spätere Aufgaben entsprechend angepasst werden, sodass z. B. an dieser Stelle die Möglichkeit besteht, bestimmte Wörter, die keinen Wert für die weiteren Aufgaben haben, z. B. bestimmte Pronomen, Konjunktionen o. ä. aus dem Corpus zu entfernen (sogenannte Stopwords). Wie aus der Definition des word2vec Algorithmus hervorgeht, ist für die Gütequalität des Modells aber notwendig, dass der ursprüngliche Kontext vollständig erhalten bleibt.

Für das Term Frequency – Inversed Document Frequency (TF – IDF) Verfahren sind beispielsweise beide Strukturarten nötig, wie in Kapitel 4.2.3 erläutert wird.

Zuletzt werden die neu generierten Wort- und Satzlisten für die weitere Verwendung in einer Datenbank abgelegt. Dazu eignen sich sowohl nichtstrukturierte wie JavaScript Object Notation (.json) oder strukturierte relationale Datenbankformate wie Comma Separated Values (.csv):

```
1. import json
2. with open('five_star_reviews_tokenized_ohne_stopwords.json', 'w') as fp:
3.   json.dump(five_star_sentences, fp)
```

**Abbildung 13: Ablegen der generierten Listen in Datenbanken (eigene Darstellung)**

Der oben gedruckte Code wird jeweils für die ein und fünf Sterne Rezensionen durchgeführt. Damit können die in vorherigen Schritten generierten Keywords jeweils in dem sehr schlecht und dem sehr gut bewerteten Kontext gegenübergestellt werden, um neue Erkenntnisse zu deren Nutzung im Sprachgebrauch der Rezensenten zu erlangen.

Diese polarisierende ein vs. fünf Sterne Aufteilung wird gewählt, da die „weichere“ ein & zwei vs. vier & fünf Aufteilung, wie sie oft in der Literatur verwendet wird, häufig auch relativierende positive bzw. negative Anteile enthalten kann. Ziel der Arbeit soll aber sein, den Wortschatz der nur sehr schlechten und der sehr guten Rezensionen zu vergleichen, um genau diese polarisierenden Produkterfahrungen zu entdecken, die einen z. B. Produktausfälle implizit enthalten können.

### 4.2.3 Term Frequency – Inversed Document Frequency

Zunächst werden die Wörter (Terms) vorbereitet, dazu wird der obige Programmcode (Abbildung 12) verwendet. Die Rezensionen werden tokenisiert, wobei auch Satzzeichen entfernt und Großbuchstaben zu Kleinbuchstaben umgewandelt werden. Für diese Anwendung werden keine Stoppwörter entfernt, da jedes Wort interessant sein kann und gerade die häufigen

Wörter durch die Multiplikation mit dem Inversed Document Frequency niedriger gewichtet werden. Eine Lemmatisierung oder Verkürzung auf die Wortstämme der Wörter wird ebenfalls nicht vollzogen.

Um ein Ranking zu erreichen, wird allerdings abweichend vom allgemeinen TF - IDF Verfahren, bei dem eine Wertematrix entstehen würde, die Term Frequency aller im Corpus befindlichen Wörter normalisiert, damit die relative Häufigkeit jedes Wortes im Gesamtkorpus errechnet werden kann. Diese wird anschließend mit der logarithmierten Inverse Document Frequency, also dem Kehrwert der Anzahl der Dokumente, in dem das Wort verwendet wird, multipliziert:

```
1.  #Code basiert zu großen Teilen auf: https://janav.wordpress.com/2013/10/27/tf-idf-
    and-cosine-similarity/
2.
3.
4.  import pandas
5.  import math
6.
7.  def termFrequency(term, documents):
8.      normalizeDocument = documents
9.      #print(normalizeDocument)
10.     return (normalizeDocument.count(term.lower()) / float(len(normalizeDocument)))
11.
12. def inverseDocumentFrequency(term, allDocuments):
13.     numDocumentsWithThisTerm = 0
14.     for doc in allDocuments:
15.         if term in doc:
16.             numDocumentsWithThisTerm = numDocumentsWithThisTerm + 1
17.
18.     if numDocumentsWithThisTerm > 0:
19.         return 1.0 + math.log(float(len(allDocuments)) / numDocumentsWithThisTerm)
20.     else:
21.         return 1.0
22.
23.
24. def tfidf(wort, dokumentenliste, wortliste):
25.     return termFrequency(wort, wortliste) * inverseDocumentFrequency(wort, dokumente
    nliste)
```

**Abbildung 14: Angepasster TF - IDF Code für das Relevanzranking der Keywords (Vembunarayanan 2013)**

Auf diese Weise entsteht eine gewichtete Liste mit der jeweiligen Relevanz der abgefragten Wörter. Dies wird zum einen für die extrahierten Keywords, aber auch für alle ca. 1,8 Millionen Wörter im Corpus durchgeführt. Die 50 relevantesten Wörter werden anschließend mit den „Nachhaltigkeitskeywords“ verglichen. Aber auch für die Nachhaltigkeitskeywords selbst kann so eine priorisierte Abfrage in nachfolgenden Aufgaben durchgeführt werden.

### 4.2.4 Wortrepräsentationen mit Word2Vec

Wie in Kapitel 3.6.1 erklärt, werden nun zwei word2vec Modelle aus den erzeugten Corpora erzeugt, eines für die sehr schlechten (1 – Stern) und eines für die sehr guten (5 – Sterne) Rezensionen. Der adaptierte Code entstammt einem Tutorial, welches den in der *gensim* Bibliothek zur Verfügung gestellten Algorithmus sehr strukturiert aufbereitet (Kaggle Inc. 2014). In dem Tutorial geht es um die Analyse von Filmrezensionen. Da eine Analogie zu Produkten (Rezensionslänge, Emotionen, Syntax etc.) durchaus denkbar ist, wurden die Parametereinstellungen in Zeile 6. - 12. übernommen, was dennoch diskussionswürdig ist (Kapitel 6.2.2):

```
1.  # Importiere ein python-integriertes Logging-Modul, um Statusmeldungen zu
    erzeugen
2.  import logging
3.  logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s',\
4.      level=logging.INFO)
5.
6.  # Algorithmusparameter können hier angepasst werden:
7.  num_features = 300    # Anzahl der Dimensionen (Features)
8.  min_word_count = 40   # Mindestvorkommen von Wörtern
9.  num_workers = 4       # Anzahl von parallen Prozessen (rechnerkapazitätsabhängig)
10. context = 10          # Anzahl der Wörter, die in dem Kontext eines Wortes mit
11.                       # berücksichtigt werden sollen
12. downsampling = 1e-3   # Downsample setting for frequent words
13.
14. # Initialisiere und trainiere das Model (nicht überwacht)
15. from gensim.models import word2vec
16. print ("Training model...")
17. model = word2vec.Word2Vec(one_star_sentences, workers=num_workers, \
18.             size=num_features, min_count = min_word_count, \
19.             window = context, sample = downsampling)
20. # Speichere das Modell unter dem Namen:
21. model_name = "270818_1star_reviews"
22. model.save(model_name)
23.
24. print("Done.")
```

**Abbildung 15: Parameterkonfiguration und Modellierung des Corpus (eigene Darstellung)**

In der obigen Abbildung nicht sichtbar, aber entscheidend für die Güte des Modells ist außerdem die Auswahl der Parameter für die Vorgehensweise des Algorithmus. So wird standardmäßig nach dem Continuous Bag of Words Verfahren (CBOW) vorgegangen und Negative Sampling angewandt. Die Auswahl ergibt sich analog zu den Filmrezensionen.

Mithilfe der erzeugten Modelle kann nun untersucht werden, welche die ähnlichsten (Cosinus Ähnlichkeit) der jeweils relevantesten Schlüsselwörter im Kontext von Nachhaltigkeitsaspekten im Produktleben (i.e. Produktdesign, -handhabung, -defekt) sind.

Dazu wird die Methode *most_similar()* mit den relevantesten Schlüsselwörtern jeweils auf das 1 Sterne sowie auf das 5 Sterne Modell angewandt und es werden entsprechend die zehn

ähnlichsten Wörter als verkettete Liste (Ausgabewort und Kosinus Distanz zwischen 0 und 1) ausgegeben:

```
1. wort="garantie"  #Beispielwort
2.
3. print("Ähnliche Wörter in sehr schlechten 1 Stern Rezensionen zum Wort: " + wort, st
   r(one_star_model.wv.most_similar(wort)))  #Setze das Wort aus Zeile 1 in das 1
   Sterne Modell und frage die 10 ähnlichsten Wörter ab.
4. print("Ähnliche Wörter in sehr guten 5 Stern Rezensionen zum Wort: " + wort, str(fiv
   e_star_model.wv.most_similar(wort))) #Setze das Wort aus Zeile 1 in das 5 Sterne
   Modell und frage die 10 ähnlichsten Wörter ab.
```

**Abbildung 16: Gegenüberstellende Abfrage eines Zielwortes und Ausgabe der zehn ähnlichsten Wörter in den beiden Modellen (eigene Darstellung)**

Eine Validierung der Vektoren kann mithilfe von Wortvektor Analogien erreicht werden. Dabei kann mithilfe von Methoden wie *doesnt_match(), similarity()* oder ebenfalls *most_similar()* getestet werden:

```
model.most_similar(positive=['woman', 'king'], negative=['man'], topn=1)
[('queen', 0.50882536)]
model.doesnt_match("breakfast cereal dinner lunch";.split())
'cereal'
model.similarity('woman', 'man')
0.73723527
```

**Abbildung 17: Methoden zur Validierung eines Modells (Řehůřek 2014)**

Analog zur obigen Abbildung 17, in der die Methoden dargestellt sind, kommt für den vorliegenden Rezensionscorpus beispielsweise die Abfrage von *samsung* + *apple* – *galaxy* in Frage, bei der die erwartete Ausgabe *iphone* wäre.

## 4.3 Visualisierung durch Dimensionsreduzierung (t-SNE)

Zur vereinfachten Darstellung der Wortvektoren im zweidimensionalen Raum wird der frei nutzbare tensorflow® Projector in einem Browser gestartet. Das Ein – Stern – Modell wird mit geringer Perplexität, das Fünf – Sterne – Modell mit höherer Perplexität kalkuliert. Es wird so lange iteriert, bis ein nahezu Stillstand der Wortvektoren erreicht wurde (ca. 5.000 Wiederholungen)

Einige Beispiele sind in den Ergebnissen in Kapitel 5.7 zu sehen.

# 5 Explorative Datenanalyse mittels Text Mining Methoden

Der in Kapitel 4 vorgestellten Methodik folgend, werden in diesem Kapitel die Ergebnisse der einzelnen Arbeitsschritte dargelegt. Alle *kursiv* gedruckten Wörter sind aus dem Corpus entnommene Wörter und werden bewusst klein geschrieben.

## 5.1 Extraktion von Produktrezensionen und Vorbereiten der Daten zur Analyse

Von den Amazon Topsellern im Bereich Handys & Smartphones wurden ausgewählte Daten zu den 50 meistverkauften Produkten zum Stichtag 31.07.2018 heruntergeladen. Neben ca. 15.000 unterschiedlichen Produktrezensionen wurden außerdem folgende Daten extrahiert (engl. „feature selection & extraction):

**Tabelle 1: Selektierte und Extrahierte Tags, die als abhängige Spalten in die Datenbank geschrieben werden (eigene Darstellung)**

| **Datenname** | **Datentyp [Informatik]** | **Skalierung [Statistik]** |
|---|---|---|
| Produktname | String | Nominal, diskret |
| Durchschnittliche Bewertung | Float | Intervall, stetig |
| Verkaufspreis | Float | Intervall, stetig |
| Amazon Identification Number (ASIN) | String | Nominal, diskret |
| Verkaufsstart bei Amazon | String | Intervall, diskret |
| URL | String | Nominal, diskret |
| Rezensionsdetails<br>• Autor<br>• Rezensionsdatum<br>• Länge der Rezension<br>• Bewertung des Produkts in Sternen<br>• Anzahl der Personen, die die Rezension hilfreich fanden | <br>• String<br>• String<br>• Integer<br>• Integer<br>• Integer | <br>• Nominal, diskret<br>• Intervall, diskret<br>• Intervall, diskret,<br>• Ordinal, diskret<br>• Intervall, diskret |

Der entsprechende Crawlercode ist im Anhang zu finden. Eine Besonderheit ist die abstrakte Struktur und das Arbeiten mit definierten Regeln, die spezifisch für Crawler eingesetzt werden, die über mehrere Seitenhierarchien bzw. Verweise springen und dort entsprechende Daten sammeln. So wird Beispielsweise der Preis und die ASIN Nummer auf höherer Ebene als die Rezensionsdaten extrahiert (Topseller → Produktseite → „alle Rezensionen anzeigen“).

Abbildung 18 zeigt einen Ausschnitt der entstandenen Datenbank mit insgesamt 14.770 Einträgen zu den 50 Smartphones und Handys. Zu jedem Feature konnten entsprechende Daten extrahiert und die Datentypen in auswertbare Formate umgewandelt werden:

| abs_url | asin | author | europreis | found_help | product | ratir | review_body | review_date | review_ | title | verkaufsstart |
|---|---|---|---|---|---|---|---|---|---|---|---|
| https://ww | B072MKS4RY | MalteS | 194 | 3131 | Samsung Gal | 4 | Leider wird die S | 20/08/2017 | 5117 | qualitati | 08/06/2017 |
| https://ww | B01BTZFQX0 | Benjamin Sch | 324 | 1428 | Samsung Gal | 5 | Ich bin seit 7 Jah | 08/03/2016 | 5182 | Iphone a | 16/02/2016 |
| https://ww | B01BTZFQX0 | chgkk | 324 | 1096 | Samsung Gal | 5 | Meine Rezensio | 15/03/2016 | 19767 | Gelunger | 16/02/2016 |
| https://ww | B01N2Z5DIY | Arvid Z. | 208,34 | 1020 | Samsung Gal | 4 | Das Samsung Ga | 07/07/2017 | 7241 | Sehr gute | 02/01/2017 |
| https://ww | B01BTZFQX0 | Steff | 324 | 827 | Samsung Gal | 5 | Nach ca einem Ja | 11/09/2017 | 8610 | S7 oder d | 16/02/2016 |
| https://ww | B06XJ49G5B | Berki | 433 | 809 | Samsung Gal | 5 | Ich habe mir das | 05/06/2017 | 8257 | Fazit nacl | 29/03/2017 |
| https://ww | B00ZUIEMM4 | anglophiler P | 241 | 765 | Huawei P10 l | 5 | Huawei hat schc | 28/11/2017 | 7251 | Endlich g | 27/02/2017 |
| https://ww | B01BTZFQX0 | Der Roland | 324 | 737 | Samsung Gal | 5 | Es gab einige Pu | 25/04/2016 | 3787 | Umstieg | 16/02/2016 |
| https://ww | B072FPB5VW | anglophiler P | 140 | 729 | Samsung Gal | 5 | Bislang noch etv | 27/09/2017 | 6930 | Für Einst | 23/05/2017 |
| https://ww | B06XSP5CNP | Matthias | 275 | 679 | Apple iPhon | 4 | ►►► Vorgesch | 11/08/2016 | 7316 | Back to th | 22/03/2017 |
| https://ww | B06XFRBM26 | BreKo | 291,95 | 561 | LG G6 Smartp | 4 | Hallo,als langjäh | 19/06/2017 | 10978 | Kantig, so | 06/03/2017 |
| https://ww | B06XSP5CNP | Mr. Zet | 275 | 547 | Apple iPhon | 4 | Ich bin ein Apple | 20/04/2016 | 11301 | die inner | 22/03/2017 |
| https://ww | B06XSP5CNP | Michael H. au | 275 | 494 | Apple iPhon | 4 | Kurz zur Vorgesc | 11/04/2016 | 5803 | Fast perf | 22/03/2017 |
| https://ww | B01L9KXU7O | Thomas Stein | 260 | 400 | Apple iPhon | 1 | Habe das Iphone | 30/12/2016 | 1013 | Vorsicht! | 30/08/2016 |
| https://ww | B06W2H865T | Fam. HappyD | 217 | 384 | Samsung Gal | 4 | Vorerst eine Infc | 26/07/2017 | 4161 | UPDATE: | 10/02/2017 |
| https://ww | B06XJ49G5B | Michael Berg | 433 | 373 | Samsung Gal | 1 | Der Vergleich zw | 19/07/2017 | 531 | nicht mar | 29/03/2017 |
| https://ww | B06W2H865T | Entscheidung | 217 | 367 | Samsung Gal | 4 | Die wichtigsten | 09/07/2017 | 2766 | Die wich | 10/02/2017 |
| https://ww | B06XJ49G5B | Michael H. | 433 | 367 | Samsung Gal | 3 | ---------- kon | 25/07/2017 | 8487 | Der Umst | 29/03/2017 |
| https://ww | B06XSP5CNP | Hanns | 275 | 332 | Apple iPhon | 1 | Bei der Bestellur | 17/11/2017 | 1189 | ACHTUNG | 22/03/2017 |
| https://ww | B07BHDDDHB | Leisetreter | 275 | 303 | Huawei 5109 | 3 | Hier nun mein ar | 14/07/2018 | 8811 | Schönes | 27/03/2018 |

**Abbildung 18: Ausschnitt der entstandenen Rezensionsdatenbank mit einer Rezension je Zeile und insgesamt gesamt ca. 14770 Einträgen (eigene Darstellung)**

Die Rezensionen unter *review_body* werden gemäß der in Kapitel 4.2.2 vorgestellten Pre - Processing Methoden aufbereitet. Es lassen sich zunächst folgende Basisinformationen für diesen Corpus festhalten:

```
1.  Anzahl Bewertungen gesamt: 14770
2.  Anzahl 1-Stern Bewertungen: 1799
3.  Anzahl 5-Sterne Bewertungen: 8842
4.
5.  Anzahl Sätze gesamt: 99916
6.  Anzahl 1-Stern Sätze: 9777
7.  Anzahl 5-Sterne Sätze: 54313
8.
9.  Anzahl Wörter gesamt: 1863220
10. Anzahl 1-Stern Wörter: 167445
11. Anzahl 5-Sterne Wörter: 982319
12.
13. #Durchschnittliche Anzahl Wörter pro Review, gesamt: 126.1489505754909
14. #Durschnittliche Anzahl Wörter pro Review, 1 Stern: 93.07670928293497
15. #Durschnittliche Anzahl Wörter pro Review, 5 Sterne: 111.09692377290206
```

**Abbildung 19: Anzahl Rezensionen, Sätze und Wörter innerhalb des Corpus (eigene Darstellung)**

Aus den Informationen ist zu entnehmen, dass mit der methodisch bedingten Entscheidung, nur die 1 bzw. 5 Sterne Rezensionen zu betrachten, ca. 70 % der Rezensionen abgedeckt werden und schlechte Rezensionen tendenziell kürzer sind als die sehr gut bewerteten.

## 5.2 Analyse der geringdimensionalen Daten mittels Interaktiven Statistischen Grafiken

Durch das Importieren der Datenbank in Excel können erste Auswertungen, beispielsweise mithilfe einer zusammenfassenden Pivot - Tabelle oder aber einfachen Zellenberechnungen durchgeführt werden:

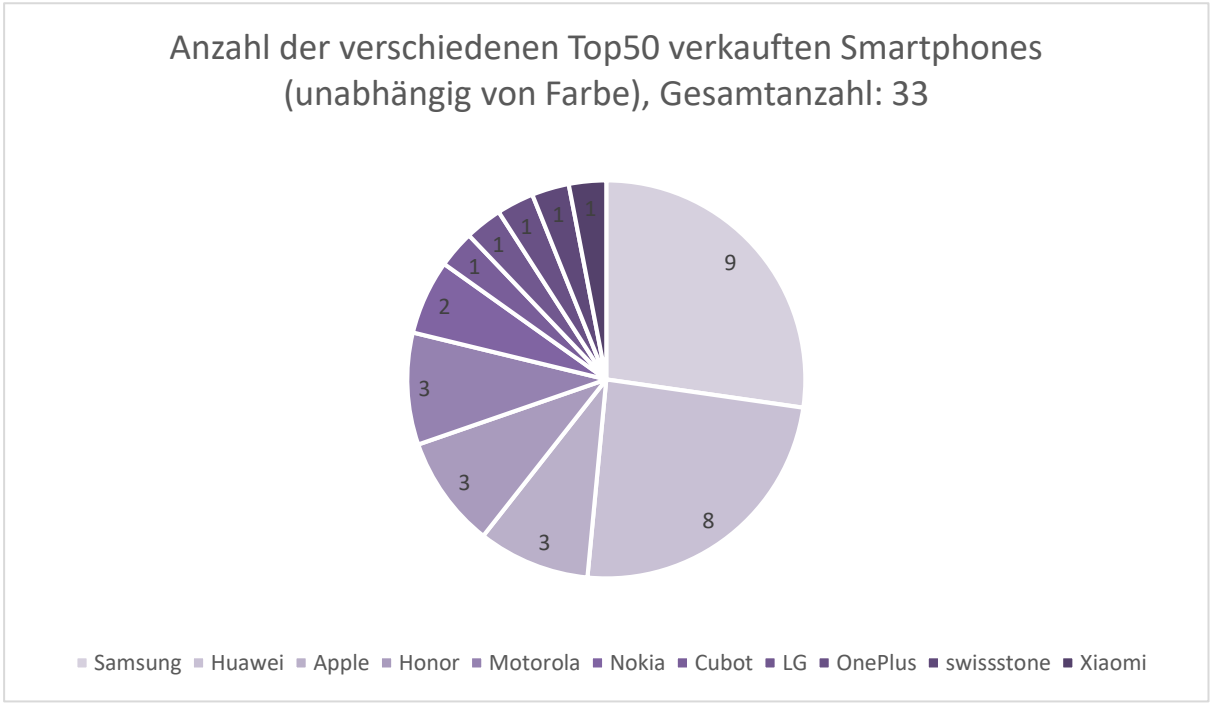


**Abbildung 20: Stichprobenauswertung, 33 verschiedene Modelle (eigene Darstellung)**

Es wird deutlich, dass es sich bei den 50 verschiedenen Datensätzen um 33 verschiedene Geräte handelt. Bei den 17 anderen Datensätzen handelt es sich um Einträge vom selben Hersteller, die aber ggfs. eine andere Modellreihe, eine andere Farbe, Speicherplatz oder sonstige kleine Unterschiede aufweisen. Diese 33 Geräte stammen von elf verschiedenen Herstellern, von denen Samsung und Huawei über die Hälfte der Geräte anbietet.

Es kann und soll sich dabei nicht um eine repräsentative Schnittmenge der allgemein verkauften Geräte handeln, da der Fokus der Arbeit bei der Auswertung der Rezensionen auf Produktgruppenebene liegt. Dennoch soll hier ein kurzer Überblick gegeben werden, welche Daten extrahiert wurden und welche Schlüsse beispielsweise daraus gezogen werden können.

**Tabelle 2: Übersicht Ergebnisse der Datenextraktion (eigene Darstellung)**

| Anzahl Geräte | Durchschnittliche Bewertungen aller Geräte (Sterne) | Durchschnittlicher Verkaufspreis aller Geräte | Anzahl Rezensionen | Durchschnittliche Anzahl von Rezensionen je Gerät |
|---|---|---|---|---|
| **50** **(davon 33 verschiedene Modelle von verschiedenen 11 Herstellern** | 4,03 | 272,96 € | 14.769 | 307 |

Tabelle 2 zeigt außerdem die durchschnittliche Bewertung aller Geräte, deren durchschnittlichen Verkaufspreis und die durchschnittliche Anzahl von Rezensionen je Gerät. Diese Informationen mithilfe einer Interaktiven Statistischen Grafik (ISG) visualisiert ergeben einige Einsichten:

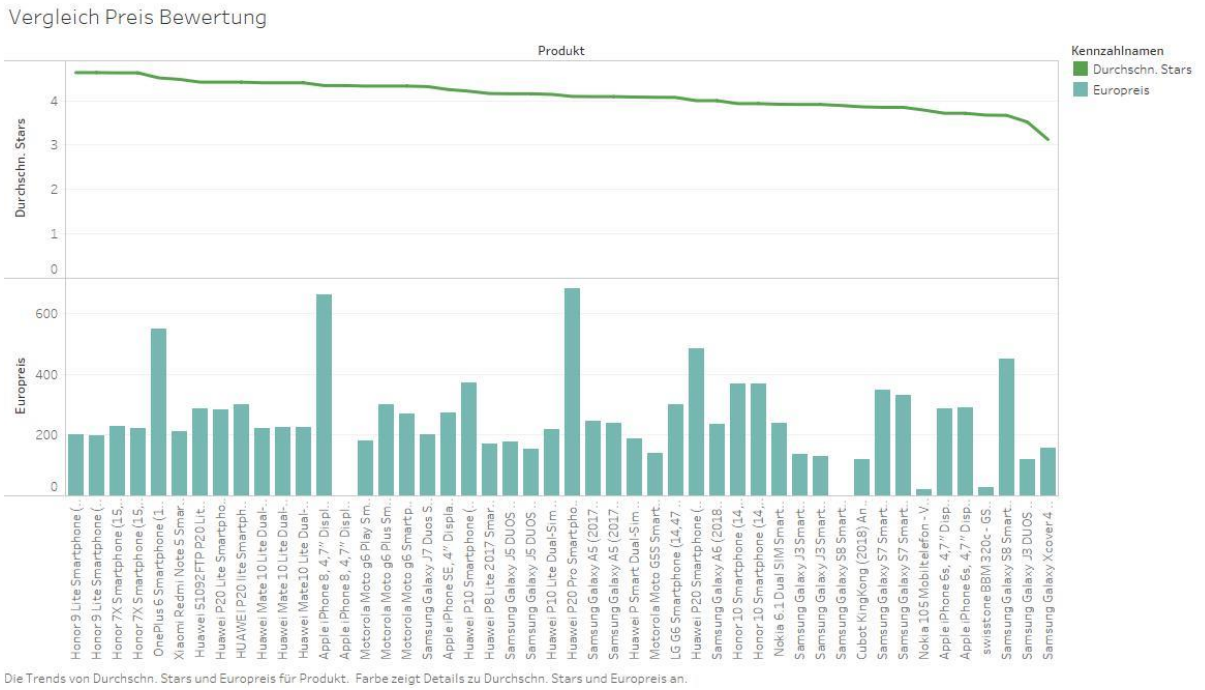


**Abbildung 21: Gegenüberstellung der durchschnittlichen Bewertung und dem Gerätepreis (eigene Darstellung mit Tableau®)**

In obiger Abbildung werden die extremen Preisunterschiede zwischen den durchschnittlich teuren Geräten und den sogenannten „Flagship“ Geräten abgebildet. Diese werden im Mittel nicht wesentlich besser bewertet. Genau das Gegenteil ist in der Stichprobe der Fall: Vier besonders günstige (< 200 €) Geräte werden im Durchschnitt besonders gut bewertet. Dies lässt ein vom Kunden besonders gut bewertetes Preis- / Leistungsverhältnis vermuten.

## 5.3 Keywordextraktion auf Basis von Ökodesignrichtlinien

Wie in Kapitel 4.1 vorgestellt, werden sechs Ökodesignrichtlinien qualitativ auf wiederkehrende Elemente und charakterisierende Wörter untersucht und diese in einer Liste festgehalten.

Die ca. 250 Wörter umfassende Liste ist im Anhang VIII.I zu finden. Von dieser Liste wurden k = 84 Wörter im verhältnismäßig kleinen Teilkorpus der sehr schlechten Rezensionen gefunden, und k = 100 in dem fünf Sterne Corpus, welcher fast ein sechsfaches Wortvolumen hat. Die entsprechenden absoluten und relativen Häufigkeiten sind in den Diagrammen auf der nächsten Seite zu sehen.

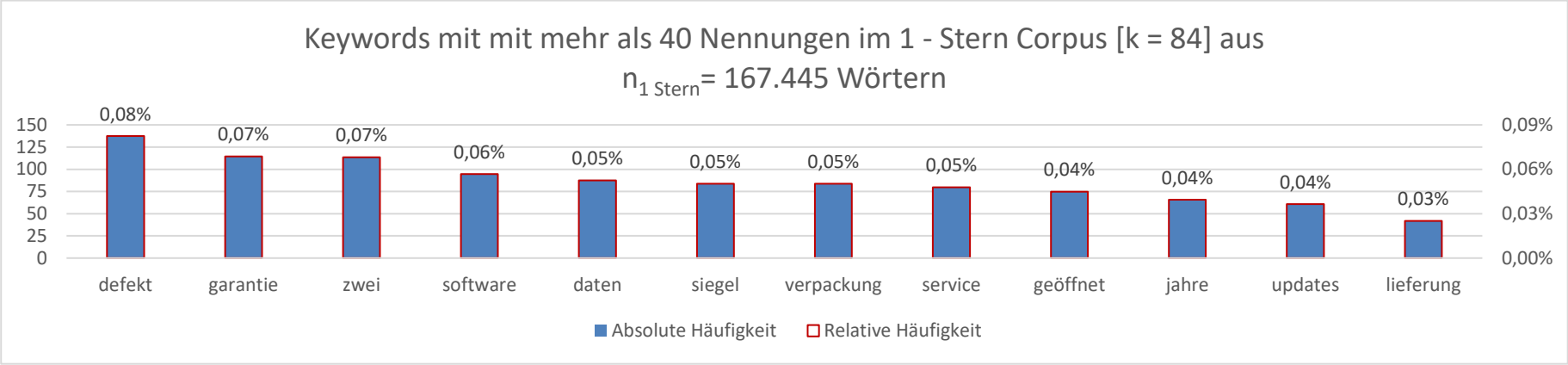


**Abbildung 22: Keywords mit mehr als 40 Nennungen in sehr schlechten / ein Stern Rezensionen (eigene Darstellung)**

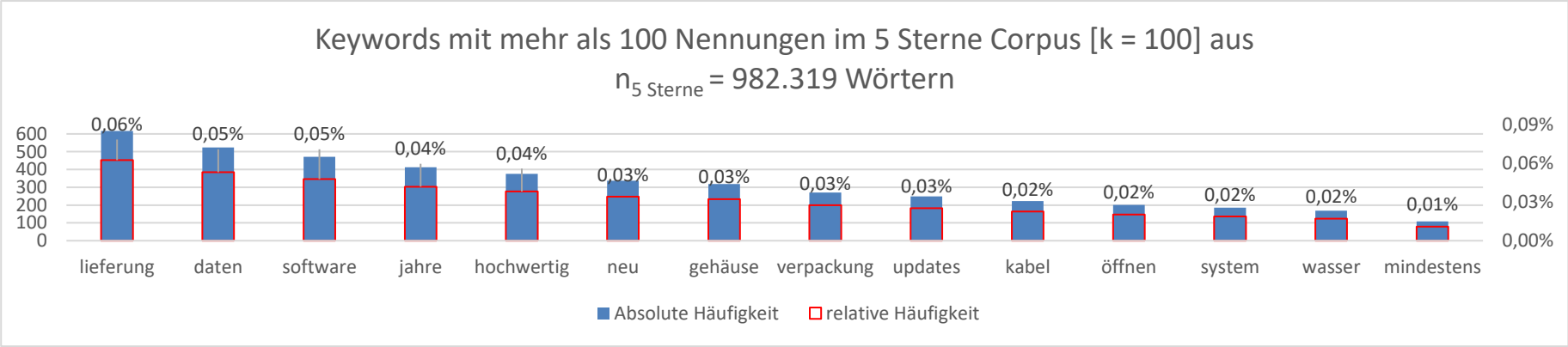


**Abbildung 23: Keywords mit mehr als 100 Nennungen in sehr guten / fünf Sterne Rezensionen (eigene Darstellung)**

Die obigen Abbildungen zeigen die absoluten und relativen Häufigkeiten der Nennungen in den Bewertungsklassen 1 und 5 Sternen. In den sehr schlechten Rezensionen werden n = 84 Keywords in einem Corpus aus 167.445 Wörtern gefunden. Vermehrt werden Wörter wie *defekt*, *zwei*, *verpackung, geöffnet* sowie *bedienungsanleitung* genannt. Ob zwischen diesen ein direkter Zusammenhang besteht, muss im Weiteren durch ein qualitatives Sichten geprüft werden.

Im Teilcorpus der sehr guten Bewertungen (n = 982.319 Wörter) fallen ebenfalls die Wörter *lieferung* und *verpackung* sowie *hochwertig*, *software*, und *updates* ins Gewicht.

In der oberen Abbildung der schlechten Rezensionen fällt beispielsweise auf, dass die beiden Wörter „zwei“ und „Jahre“ enthalten sind. Es ist demnach zu prüfen, ob dies tatsächlich häufig vorkommt und in dem Kontext gegebenenfalls von Defekten und Ausfällen gesprochen wird.

Auf der anderen Seite findet man in sehr guten Bewertungen das Wort *wasser*, dessen Ursprung nicht direkt eindeutig ist und im Weiteren untersucht wird.

Außerdem fällt auf, dass in beiden Corpora die Wörter *lieferung, daten, software, service* und *updates* besonders häufig vorkommen.

Da die reine Häufigkeit an Nennungen von Wörtern noch nicht unbedingt aussagekräftig ist, wird in Kapitel 5.5 mithilfe des TF-IDF- Verfahren die Relevanz der Keywords im Gesamtzusammenhang dargestellt.

Im Folgenden wird durch ein qualitatives Screening der Rezensionstexte versucht, mehrdeutige Nennungen in ihrem Kontext zu analysieren und weitere relevante umgangssprachliche Wörter aus den Rezensionen herauszufiltern und damit einen Versuch, eine umgangssprachliche Liste an Wörtern im Hinblick auf Ökodesign zu generieren, die später weiter verwendet werden kann (siehe Ausblick, Kapitel 7).

## 5.4 Qualitatives Screening des Corpus nach beschreibender äquivalenter Umgangssprache

Es werden die von Amazon-Kunden 15 hilfreichsten Rezensionen (1 und 5 Sterne) nach den in dem Teilcorpus am häufigsten vorkommenden Schlüsselworten durchsucht, deren Bedeutung herausgearbeitet wird und, wenn möglich, Schlüsselwörter extrahiert:

**Tabelle 3: Ergebnisse des Durchsuchens der Corpora nach Kontext und umgangssprachlichen Wörtern (eigene Darstellung)**

| **1-Stern-Wort:** | **Kontext:** | **5-Sterne-Wort:** | **Kontext:** |
|---|---|---|---|
| „Jahre“ | Akkulaufzeit lässt nach, Entwicklungsrückstand zu anderen Herstellern | „Jahre“ | Nutzungsdauer aber auch Altern des vorherigen Gerätes |
| „Software“ | „mehr Softwareprobleme als je zuvor“, | „Software“ | Überschrift bzw. Gliederungspunkt in der Bewertung |
| „daten“ | Unerlaubte Datensammlung von Herstellern | „daten“ | Datensicherung und -übertragung |
| „Verpackung“ | Zumeist wird die Verpackung bemängelt, weil sie beschädigt ist oder im Fall „generalüberholte Geräte“, die als „Neu“ verkauft werden, nicht original | „Verpackung“ | Überschrift bzw. Gliederungspunkt in der Bewertung |
| „siegel“ | nicht neu, nicht original, überklebt | „hochwertig“ | Rahmen, Konstruktion, Fassung Display |
| „defekt“ | nicht austauschbarer Akku | „neu“ | Neugerät |

| | | | |
|---|---|---|---|
| „geöffnet“ | als „neu“ gekennzeichnete bzw. verkaufte Ware, defekt geliefert | „Gehäuse“ | Hochwertige Verarbeitung, aus Metall |
| Bedienungsanleitung | fehlt, falsch geliefert | Bedienungsanleitung | Beschreibung des Lieferumfangs, Gefühl beim Auspacken |
| akku | Nicht wechselbar | „Lieferung“ | Geschwindigkeit, Unversehrtheit |

Beim Screening der häufigsten Ökodesignextrakte konnten weitere für das nachhaltige Produktdesign relevante Wörter aus der Rezensionsjargon bzw. Umgangssprache gesammelt werden.

Dabei kann unterschieden werden in Wörter mit Bezug zum Produktdesign, zur Software oder zur Serviceleistung.

Produktdesign: *sturz, spider-app, billig, marktreif, wasserschaden, ip68, kratzer, risse, grottig, designschnitzer, schutzfolie, hülle, lose, müll, unbrauchbar*

Software: *flüssig, zügig, ruckfrei, langsam, aufhängen, aufpoppen, unübersichtlich, freeware, werbung*

Serviceleistung: *grottig, unterirdisch, gefälscht, rückrufaktion, unverschämt, betrug,*

Auffällig zu beobachten ist, dass die Rezensenten in den sehr guten Rezensionen die hochwertige Verarbeitung der Geräte bewundern, obwohl unterstellt werden kann, dass viele Smartphonebesitzer schon einmal erlebt haben dürften, wie schnell ein Displaybruch oder Kratzer an solch einem Gerät entstehen. Zumindest wird dies auf Webseiten, Social Media Kanälen und Zeitschriften getestet und berichtet. (Orellana 2017; Stiftung Warentest 2018)

Als Ergebnis und Fazit des qualitativen Screenings lässt sich festhalten, dass sich gerade sehr schlechte Bewertungen häufig nicht auf das Gerät oder dessen schlechte Qualität, sondern viel mehr auf den Händler, die Verkaufsabwicklung, oder die vom ausgewiesenen Artikelzustand abweichende Beschaffenheit des Produkts bzw. Smartphones beziehen.

Dennoch können einige neue umgangssprachliche Keywords generiert werden, die in die weitere Analyse mit einfließen. Diese werden in die Keywordliste mit aufgenommen und auf ebenfalls ihre Relevanz geprüft.

## 5.5 Bestimmung der Relevanz der ausgewählten Keywords im Corpus

Tabelle 4 zeigt die das Ergebnis der Anwendung des angepassten TF-IDF Methode auf alle sich jeweils in den sehr schlechten (1 Stern) und sehr guten (5 Sterne Rezensionen) befindlichen Wörter. Es wird untersucht, ob in den 50 relevantesten Wörtern auch einige der in den vorherigen Kapiteln zusammengestellten „Nachhaltigkeitskeywords“ zu finden sind. Anschließend werden gezielt die TF-IDF Werte der in den Kapiteln 5.3 und 5.4 extrahierten Schlüsselwörter analysiert und der Relevanz nach sortiert.

**Tabelle 4: Gegenüberstellung von den jeweils relevantesten Wörtern in sehr schlechten und sehr guten Rezensionen, gemessen am TF-IDF Wert in den Teilcorpora (eigene Darstellung)**

| | 50 relevanteste Wörter in sehr schlechten Rezensionen | 50 relevanteste Wörter in sehr guten Rezensionen |
|---|---|---|
| 1. | ('handy', 0.0518285869478), | ('s', 0.02221436507964175), |
| 2. | ('fur', 0.03929925986722), | ('handy', 0.022122547268484262), |
| 3. | ('gerat', 0.03751135180264), | ('gut', 0.02008939363918561), |
| 4. | ('samsung', 0.037379398720), | ('smartphone', 0.01694404200550), |
| 5. | ('s', 0.03449961598652402), | ('kamera', 0.01521091730539919), |
| 6. | ('dass', 0.033300534656220), | ('display', 0.0141475035713233), |
| 7. | ('mehr', 0.031583215705217), | ('samsung', 0.0133548272330776), |
| 8. | ('wurde', 0.02895412994377), | ('super', 0.01270626203465816), |
| 9. | ('schon', 0.02501653762129), | ('akku', 0.012555362491368273), |
| 10. | ('mal', 0.024285999749896), | ('dass', 0.012363426124637606), |
| 11. | ('leider', 0.0239603748594), | ('gerät', 0.01158528516599481), |
| 12. | ('display', 0.023513509792), | ('preis', 0.011491186879958851), |
| 13. | ('akku', 0.0211999463084), | ('schnell', 0.0114814124646648), |
| 14. | ('smartphone', 0.020482813), | ('mehr', 0.011360607443784934), |
| 15. | ('beim', 0.01884655086179), | ('iphone', 0.0108304842819969), |
| 16. | ('amazon', 0.0185197176098), | ('schon', 0.010613652521479689), |
| 17. | ('immer', 0.0177100108386), | ('huawei', 0.01060288462829020), |
| 18. | ('zuruck', 0.01754178964828466), | ('einfach', 0.0103096995409851), |
| 19. | ('uber', 0.017085850953828963), | ('zufrieden', 0.01012209823759), |
| 20. | ('apple', 0.016516747353877734), | ('beim', 0.009522305174826013), |
| 21. | ('iphone', 0.01610815938350901), | ('immer', 0.009408468283670065), |
| 22. | ('einfach', 0.0154997823286334), | ('wirklich', 0.009386457650899), |
| 23. | ('geht', 0.015070270039768562), | ('gb', 0.009122132179367732), |
| 24. | ('problem', 0.0146568523185128), | ('lite', 0.009093632084789702), |
| 25. | ('a', 0.013543645308447448), | ('android', 0.008104952521134), |
| 26. | ('funktioniert',0.013213567302), | ('macht', 0.008076644346329952), |
| 27. | ('gut', 0.013089661994632492), | ('galaxy', 0.00806135892685209), |
| 28. | ('kamera', 0.0128223220856650), | ('a', 0.007924907367602044), |
| 29. | ('ja', 0.012456986988891302), | ('mal', 0.007889334884791717), |

| | | |
|---|---|---|
| 29. | ('ja', 0.012456986988891302), | ('mal', 0.007889334884791717), |
| 30. | ('telefon', 0.0123196487662202), | ('gute', 0.007735202570181641), |
| 31. | ('nie', 0.012190505625449238), | ('hand', 0.007643761196137977), |
| 32. | ('gekauft', 0.0121529241221378), | ('top', 0.007604818354210954), |
| 33. | ('galaxy', 0.0121358396809404), | ('p', 0.007554649044517813), |
| 34. | ('probleme', 0.012026507791923), | ('bilder', 0.0074211627320161), |
| 35. | ('wochen', 0.01133856685505299), | ('apps', 0.007325194206999603), |
| 36. | ('android', 0.0109117682320194), | ('gibt', 0.007007862920425304), |
| 37. | ('kam', 0.010848777144799619), | ('fotos', 0.006764143115168049), |
| 38. | ('konnte', 0.01083551020396186), | ('x', 0.0067383204107787364), |
| 39. | ('sim', 0.010787279334420365), | ('honor', 0.006729911297066676), |
| 40. | ('bekommen', 0.01077224325326), | ('funktioniert', 0.00659358379), |
| 41. | ('apps', 0.010639058943221652), | ('g', 0.006539348975253431), |
| 42. | ('gar', 0.010626238126444754), | ('liegt', 0.006377080984893477), |
| 43. | ('g', 0.010416605712504563), | ('hält', 0.0062605676846450 36), |

**Gemeinsamkeiten der Corpora (grün):**

Erwartungsgemäß hohe Relevanz scheinen die Komponenten und damit die Eigenschaften des Akkus, Speichers, der Kamera und des Displays zu haben, was sich in Beschwerden als auch Zufriedenheit zeigt.

Interessant ist, dass von den Herstellern nur *Samsung* in beiden Corpora häufig zu finden ist. Andere genannte Hersteller wiederum finden sich entweder in den sehr schlechten oder sehr guten. Gleiches gilt für das Betriebssystem *Android*.

Das „*s“* bei den fünf Sterne Bewertungen das relevanteste Wort ist, kann ggfs. durch die Produktbenennungen „Galaxy S<Generation>“ “ bzw. „iPhone <Generation>s“ begründet werden. Da bei der Vorbereitung der Daten alle Zahlen gelöscht werden bleibt ein „*s*“ als eigener Token stehen.

Auffällig ist außerdem, dass das Wort *Handy* das relevanteste Wort in beiden Corpora ist, und sich der „Rufname“ des Geräts während der Entwicklung der letzten Jahre vom quasimonofunktionalen Tastentelefon hin zum multifunktionalen Smartphone scheinbar nicht verändert hat.

**1 Sterne Beobachtungen (gelb, links):**

In den sehr schlechten Rezensionen fällt zunächst auf, dass – wahrscheinlich stichprobenbedingt – die Hersteller *Apple* und *Samsung* und ihre „Flagship“ Modelle *iPhone* und *Galaxy* eine hohe Relevanz haben. Dies kann an den großen Markt- & Stichprobenanteil der beiden führenden Hersteller liegen, zeigt aber auch, dass die Rezensenten sich über diese beiden Hersteller häufig negativ äußern. Auch kann interpretiert werden, dass sehr schlecht bewertete Geräte zu Amazon zurückgeschickt werden.

**5 Sterne Beobachtungen (gelb, rechts):**

In den sehr gut bewertenden Rezensionen weisen zusätzlich zu dem genannten Hersteller *Samsung* außerdem *Huawei*, *Honor* eine hohe Relevanz auf.

Positiv werden außerdem der *speicher*, *fotos*, *apps* und auch der *preis* hervorgehoben.

**Direktes Antreffen von Nachhaltigkeitskeywords:**

Die am häufigsten vorkommenden zuvor extrahierten Nachhaltigkeitskeywords wie *defekt*, *software*, *lieferung*, etc. finden sich nicht unter den 50 relevantesten Wörtern der beiden Corpora. Auch lassen sich keine Schlüsse aus obiger Tabelle ziehen, die implizit oder explizit auf Nachhaltigkeitsaspekte hinweisen.

Diesen Ergebnissen folgend, werden diese um die neu generierten umgangssprachlichen Keywords wie *Müll*, *unbrauchbar*, *billig*, etc. ergänzt und auf ihre Relevanz (Erklärung siehe 4.2.3: Term Frequency – Inversed Document Frequency) im Corpus geprüft:

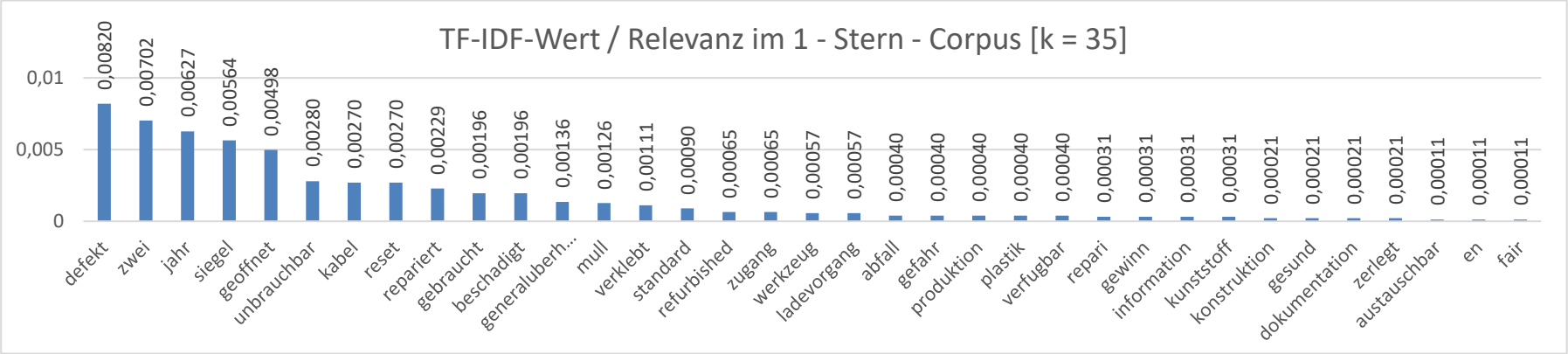


**Abbildung 24: TF - IDF - Werte für die relevantesten Keywords im Teilcorpus (1 Stern) (eigene Darstellung)**

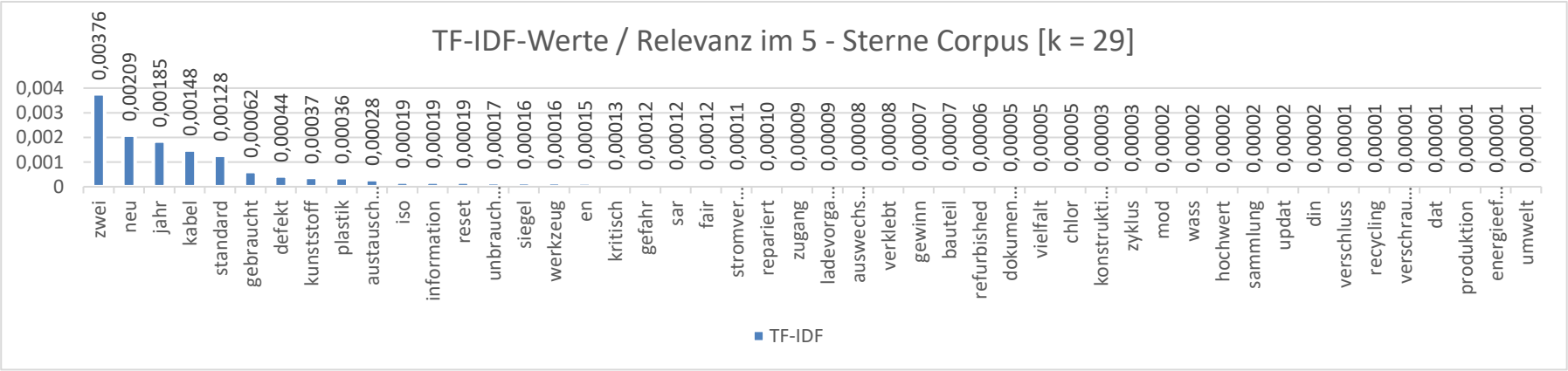


**Abbildung 25: TF - IDF - Werte für die relevantesten Keywords im Teilcorpus (5 Sterne) (eigene Darstellung)**

Zum einen wird sichtbar, dass die errechneten Werte generell unter den der 50 relevantesten Wörter in den Corpora (Tabelle 4) liegen. Dennoch konnte eine Rangfolge nach Relevanz erstellt werden, um den Aufwand der weiteren Vorgehensweise zu reduzieren, indem nun nur die relevantesten „Nachhaltigkeitskeywords“ analysiert werden können.

## 5.6 Untersuchung der relevantesten Wörter mithilfe von Word Embeddings

Im Folgenden werden die Wörter nach und nach abgefragt und gegenübergestellt und Auffälligkeiten entsprechend unter der jeweiligen Zelle festgehalten. Oben in blau das Keyword, unten die nach dem Skalarprodukt errechneten Wörter mit der Kosinus-Ähnlichkeit zwischen 0 und 1. Aus den Ergebnissen werden in Kapitel 6.2.2 weitergehende Forschungsfragen und Hypothesen generiert:

**Tabelle 5: Keywords und ihre Kosinus Ähnlichkeiten in den Corpora (eigene Zusammenstellung)**

| | |
|---|---|
| 1. Ähnliche Wörter in sehr schlechten (1 Stern) Rezensionen zum Wort: garantie<br>2. [('bzw', 0.9995433688163757),<br>3. ('verkäufer', 0.9995420575141907),<br>4. ('updates', 0.9995350241661072),<br>5. ('welche', 0.999524712562561),<br>6. ('hin', 0.9994893670082092),<br>7. ('genau', 0.9994864463806152),<br>8. ('funktion', 0.9994790554046631),<br>9. ('extrem', 0.9994789361953735),<br>10. ('dadurch', 0.9994734525680542),<br>11. ('klar', 0.999462366104126)] | 12. Ähnliche Wörter in sehr guten (5 Sterne) Rezensionen zum Wort: garantie<br>13. [('monate', 0.7276219129562378),<br>14. ('dementsprechend',0.67377156019210)<br>15. ('verkauft', 0.6718186140060425),<br>16. ('eur', 0.6628209948539734),<br>17. ('testberichte', 0.6585674285888672),<br>18. ('china', 0.6579269170761108),<br>19. ('schließlich', 0.6579117774963379),<br>20. ('microsoft', 0.6507819890975952),<br>21. ('defekt', 0.6438179016113281),<br>22. ('eins', 0.6380600929260254)] |
| Das wort garantie hat zu allen zehn ähnlichsten Wörtern eine sehr hohe Kosinus Ähnlichkeit, was generell für eine hohe Relevanz des Wortes im Corpus sprechen kann. | Bei den sehr gut bewerteten Geräten spielt Garantie scheinbar keine besonders große Rolle, da die Kosinus Ähnlichkeiten relativ gering sind. |
| 23. Ähnliche Wörter in sehr schlechten (1 Stern) Rezensionen zum Wort: ständig<br>24. [('ladegerät', 0.9997557997703552),<br>25. ('ganze', 0.9997520446777344),<br>26. ('bleibt', 0.9997513294219971),<br>27. ('neustart', 0.9997304677963257),<br>28. ('starten', 0.9997267723083496),<br>29. ('gab', 0.999707043170929),<br>30. ('ins', 0.9997059106826782),<br>31. ('heiß', 0.9996994733810425),<br>32. ('stunden', 0.999692440032959),<br>33. ('folie', 0.9996813535690308)] | 34. Ähnliche Wörter in sehr guten (5 Sterne)<br>35. Rezensionen zum Wort: ständig<br>36. [('deinstallieren', 0.7443109154701)<br>37. ('mehrere', 0.732969343662262),<br>38. ('sekunden', 0.732553243637085),<br>39. ('löschen', 0.7303502559661865),<br>40. ('anrufe', 0.717505693435669),<br>41. ('deaktiviert', 0.7090358138084412),<br>42. ('gehen', 0.690481424331665),<br>43. ('benutzt', 0.6813015937805176),<br>44. ('tauschen', 0.6763468384742737),<br>45. ('bleiben', 0.6757492423057556)] |
| Auffällig: Hinweise auf Softwareprobleme und starkes Erhitzen der Geräte. | Scheinbar keine sinnvollen Zusammenhänge. Das Keyword wird ggfs. nur im negativen Sinne genutzt. |

```
Ähnliche Wörter in sehr schlechten (1 Stern) Rezensionen zum Wort: kratzer
[('gefallen', 0.9996333122253418),
('meinung', 0.9996203184127808),
('touch', 0.9996122121810913),
('cm', 0.9995765686035156),
('gerätes', 0.9995704293251038),
('akkulaufzeit', 0.999510884284973),
('laut', 0.9994934797286987),
('einwandfrei', 0.999491810798645),
('gebrauch', 0.9994782209396362),
('allerdings', 0.9994766712188721)]
```

```
Ähnliche Wörter in sehr guten (5 Sterne) Rezensionen zum Wort: kratzer
[('macken', 0.6136834621429443),
('rand', 0.5903582572937012),
('gebrauchsspuren', 0.5866429209707),
('hänger', 0.5853250622749329),
('schutz', 0.5580377578735352),
('sturz', 0.5565347075462341),
('weiteres', 0.5518763065338135),
('tippen', 0.5468664169311523),
('einzigen', 0.5358551740646362),
('verzögerungen', 0.5196951627731323]
```

Interessant und gleichzeitig irreführend ist, dass „Kratzer“ im Zusammenhang mit „gefallen“ genannt wird. Die anderen Wörter hängen nicht logisch miteinander zusammen.

Auffällig: In dem Zusammenhang von Kratzern (optischen Mängeln) wird von Schäden nach einem Sturz und auch von Softwareproblemen gesprochen, auch in dem Kontext von sehr gut bewerteten Geräten.

```
Ähnliche Wörter in sehr schlechten (1 Stern) Rezensionen zum Wort: jahr
[('erste', 0.9994968175888062),
('ersten', 0.9994473457336426),
('bereits', 0.9994246959686279),
('update', 0.999419629573822),
('zweite', 0.9993118643760681),
('knapp', 0.9993036389350891),
('woche', 0.9992203116416931),
('ladekabel', 0.9991980195045471),
('vor', 0.9991299510002136),
('gleiche', 0.9991104602813721)]
```

```
Ähnliche Wörter in sehr guten (5 Sterne) Rezensionen zum Wort: jahr
[('halben', 0.7863253355026245),
('monat', 0.7693888545036316),
('vertrag', 0.7359796762466431),
('jahren', 0.7147518992424011),
('nachtrag', 0.6970269680023193),
('damals', 0.6843738555908203),
('jahre', 0.6754512190818787),
('geist', 0.6749156713485718),
('aufgegeben', 0.6700228452682495),
('monate', 0.662226676940918)]
```

Auffällig: Das Wort „bereits“ könnte für ein Nichterfüllen der Eigenschaften der Geräte sprechen.

„Geist aufgegeben“ könnte als Synonym für den Prozess des Kaputtgehens verstanden werden.

```
Ähnliche Wörter in
sehr schlechten (1 Stern)
Rezensionen zum Wort: akku
[('aus', 0.9979011416435242),
('einfach', 0.9975090026855469),
('lässt', 0.9975042343139648),
('funktioniert', 0.99648487567901),
('gar', 0.9958599209785461),
('auch', 0.9957276582717896),
('wird', 0.9953492879867554),
('nur', 0.9952620267868042)
('auf', 0.9949772357940674),
('ist', 0.994901180267334)]
```

```
Ähnliche Wörter in
sehr guten (5 Sterne)
Rezensionen zum Wort: akku
[('lange', 0.6439356207847595,
('batterie', 0.634077250957489,
('akkuverbrauch', 0.5955811738678),
('locker', 0.5937812328338623,
('er', 0.5858451128005981),
('verspricht', 0.58572572469713),
('akkuleistung', 0.5730884075147),
('länger', 0.5689495205879211),
('normaler', 0.566897928714752),
('tage', 0.5619903802871704)]
```

Nicht eindeutig, aber es kann eine eher willkürliche Funktionsweise herausgelesen werden.

Bei Neugeräten scheint die Akkuleistung ausreichend und positiv bewertet zu sein.

```
Ähnliche Wörter in
sehr schlechten (1 Stern)
Rezensionen zum Wort: display
[('bildschirm', 0.997706174850463),
('einer', 0.9976996183395386),
('ließ', 0.9973305463790894),
```

```
Ähnliche Wörter in
sehr guten (5 Sterne)
Rezensionen zum Wort: display
[('bildschirm', 0.720758974551545),
('amoled', 0.684346079826355),
('design', 0.6404651999473572),
```

| | |
|---|---|
| 122. ('aus', 0.9972907304763794),<br>123. ('defekt', 0.9971866607666016),<br>124. ('sturz', 0.9971614480018616),<br>125. ('glas', 0.9971200227737427),<br>126. ('auspacken', 0.9971162676811218),<br>127. ('hosentasche', 0.997039139270782),<br>128. ('leer', 0.9970061779022217)] | 135. ('ips', 0.60289466381073),<br>136. ('full', 0.5884303450584412),<br>137. ('panel', 0.5871849060058594),<br>138. ('große', 0.5866104364395142),<br>139. ('bild', 0.582200288772583),<br>140. ('hell', 0.5731487274169922),<br>141. ('fhd', 0.5724231004714966)] |
| Im negativen Kontext "Display" weist der Großteil der Wörter auf eine Anfälligkeit gegenüber Stürzen hin. | Bei den positiven Rezensionen fällt der Bezug zu den Produktfeatures auf, die scheinbar gelobt werden. |
| | 142. Ähnliche Wörter in<br>sehr guten (5 Sterne)<br>154.Rezensionen zum Wort: austauschbar<br>143. [('austauschbare', 0.89152264595031)<br>144. ('verbaute', 0.8703145980834961),<br>145. ('wechselbar', 0.8562930822372437),<br>146. ('eingebaut', 0.8511594533920288),<br>147. ('akkuder', 0.8286426067352295),<br>148. ('fest', 0.8151286840438843),<br>149. ('ausdauernder', 0.8042827844619751),<br>150. ('akkukapazität', 0.75897490978240),<br>151. ('mah', 0.7551845908164978),<br>152. ('beachtlich', 0.7522680759429932)] |
| Keine ausreichende Datenmenge im Modell verfügbar. | Es wird sichtbar, dass ein austauschbarer Akku von den Kunden sehr gelobt wird. |
| | 155.Ähnliche Wörter in<br>sehr guten (5 Sterne)<br>Rezensionen zum Wort: fair<br>156.[('stimmt', 0.7738270163536072),<br>157.('geboten', 0.7562133073806763),<br>158.('highend', 0.7458354234695435),<br>159.('gerechtfertigt', 0.732097744941711),<br>160.('leistungs', 0.7281908988952637),<br>161.('angemessen', 0.7225494384765625),<br>162.('fairen', 0.7223147749900818),<br>163.('günstigen', 0.7174841165542603),<br>164.('unschlagbar', 0.7147505283355713),<br>165.('leistungssieger', 0.70849090814590)] |
| Keine ausreichende Datenmenge im Modell verfügbar. | Im Zusammenhang mit dem Wort „fair" stehen Wörter im Verhältnis zur Preis / Leistung |

Dieses Vorgehen impliziert eine Art Cluster der ähnlichsten zehn Wörter um ein Zielwort herum. Es findet allerdings keine direkte Zuordnung der Wortvektoren zu einem bestimmten Cluster statt, ohne dass dieses Wort nicht ähnlich zu einem anderen, sich in der Nähe befindlichen Wortes befindet. Bei anderen Verfahren, z. B. das k - means Clustering wird die Datenmenge tatsächlich in k verschiedene Gruppen / Cluster eingeteilt, in dem kein Wort bzw. Vektor doppelt vorkommt.

Ein angewandter Vergleich der beiden Methoden hat jedoch gezeigt, dass bei einem iterativen Vorgehen zum Finden einer optimalen Anzahl k Cluster, keine trennscharfen Ergebnisse im Sinne der Nachhaltigkeitsaspekte im Zusammenhang mit den Produkteigenschaften erzeugt werden konnten.

## 5.7 Ergebnisse Dimensionsreduzierung mit Tensorflow Projector

Für den 1 Sterne Corpus können aufgrund der wenigen Datenpunkte mit dem t-SNE Algorithmus keine aussagekräftigen Darstellungen erzeugt werden. Mit dem größeren 5 Sterne Corpus hingegen ergeben sich eindeutig zuordbare Wortanhäufungen bzw. Cluster unter anderem zu den Worten *batterie, display* und *rutschig.* Die Ergebnisse dienen allerdings maximal der Anschaulichkeit. Abstände zwischen den Wortvektoren können zwar beispielsweise als nah oder fern bezeichnet werden, eine Messung des Abstandes selbst führt zu keiner validen Aussage:

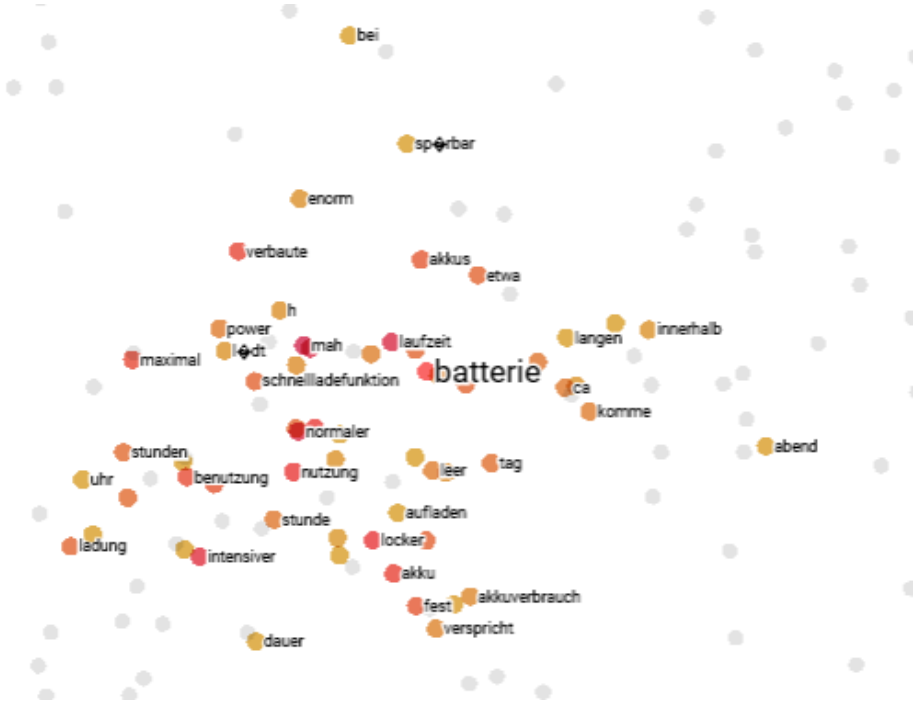


**Abbildung 26: Wörter im Umfeld des Wortes *batterie im 5 Sterne Corpus* (eigene Darstellung mit Tensorflow® Projector)**

In Abbildung 26 ist zu erkennen, dass beispielsweise Wörter, die in dem Zusammenhang mit *batterie* in den Rezensionen genannt werden, auch in dieser Abbildung zu finden sind. Erwartungsgemäß findet man Wörter im Bereich der Akkulaufzeit (*locker, langen, ca, etwa, innerhalb)*, des Nutzungsverhaltens (*intensiver, normaler)* und des Ladeverhaltens (*schnellladefunktion).* In Bezug zu den Nachhaltigkeitsaspekten ist zu sehen, dass in dem Kontext der sehr guten Rezensionen auch Wörter zu finden sind, die auf die Demontagefähigkeit des Akkus hinweisen können (*verbaute, fest).*

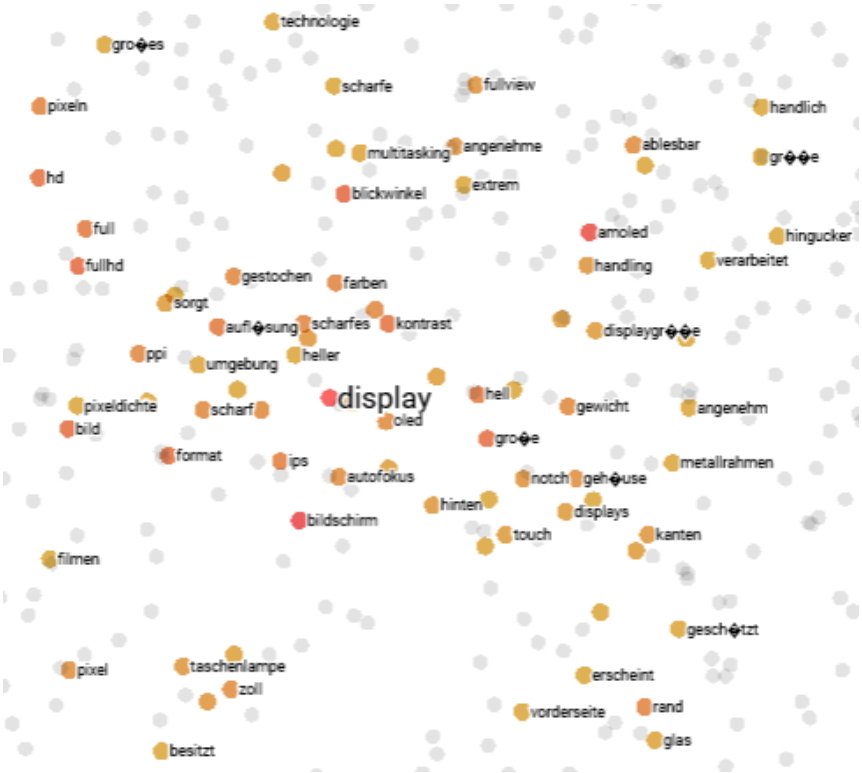


**Abbildung 27:Wörter im Umfeld des Wortes *display* im 5 Sterne Corpus (eigene Darstellung mit Tensorflow® Projector)**

Im Umfeld des Wortes *display* fallen – wenig überraschend – zunächst viele technische Eigenschaften ins Auge (*touch, gestochen, pixeldichte, fullhd, amoled, ips*[1]*, kontrast, notch*[2]*)* aber auch Überschneidungen mit den Eigenschaften der Kamera und der allgemeinen Verarbeitung sind zu finden (*filmen, autofokus, taschenlampe)*. Für das nachhaltige Produktdesign- & Entwicklung interessante Wörter können sein: *handling, angenehm, kanten, rand, glas, geschützt.* Da es sich hierbei um die sehr guten Bewertungen handelt, ist diese Anordnung und Wortauswahl zu erwarten.

[1] In Plane Switching (IPS) – Technik, die den optimalen Betrachtungswinkel vergrößert und die Farbdarstellung optimiert.

[2] Oberer Teil des Displays, an dem z.B. Hörlautsprecher, Sensoren und Frontkamera verbaut sind. Das Display wird in den neuesten Smartphone Generationen häufig um die Notch herumdesignt, um die Bildschirmdiagonale zu vergrößern.

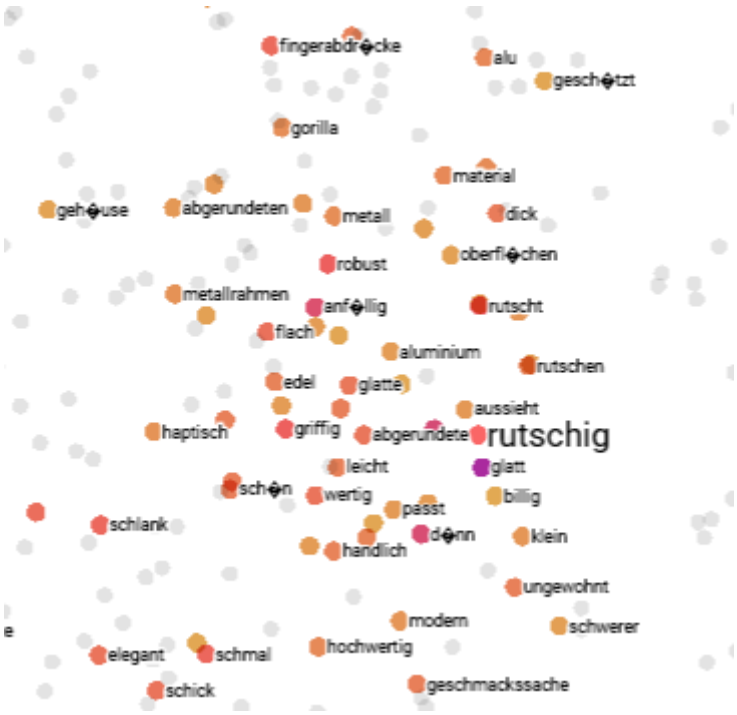


**Abbildung 28: Wörter im Umfeld des Wortes *rutschig* im 5 Sterne Corpus (eigene Darstellung mit Tensorflow® Projector)**

Einem anderen Ansatz folgend konnte das Wort *rutschig* identifiziert werden. Man kann vermuten, dass damit das Gerät an sich beschrieben wird und eine Folge einer *rutschigen Oberfläche* impliziert, dass der Rezensent die Gefahr hat, dass das Gerät leicht herunterfallen kann. In dem Umfeld des Wortes finden sich außerdem beschreibende Wörter wie *glatt, rutschen, rutscht, anfällig* aber genauso positiv annotierte Wörter *edel, abgerundet, schick, elegant, hochwertig,* was zum anderen darauf hinweist, dass sich die Rezensionen verständlicherweise positiv über das Design, Haptik und Erscheinung des Gerätes äußern.

# 6 Evaluierung und Diskussion der Ergebnisse

## 6.1 Ergebnisse der Untersuchung mit Bezug zur Fragestellung

In Hinblick auf die Fragestellung, Schlüsselwörter für die nachhaltige Produktentwicklung aus Produktrezensionstexten herauszuarbeiten und deren Zusammenhang abzubilden, konnte die entwickelte Methodik interessante Ergebnisse liefern. Durch das strukturierte Vorgehen konnten zunächst Ökodesignrichtlinien auf inhaltliche Unterschiede und Parallelen untersucht werden und deren Schlüsselwörter extrahiert werden. Bei der Bearbeitung wurde relativ zügig deutlich, dass sich das Fachvokabular nur zur ca. 20 % in dem Rezensionscorpus wiederfinden ließ. Dies hatte eine Erweiterung der Methodik um die Umgangssprache aus dem Corpus zur Folge, welche eine Art „mixed-Methods“ Ansatz verfolgte und zu einer erweiterten Liste führte. Auf diese Art und Weise konnten Nachhaltigkeitskriterien gesammelt werden. Die Liste kann als Anfang und nur auf Smartphones bezogene Liste gesehen werden. Die allgemeine Vorgehensweise – angefangen bei der Extraktion der „Topseller“ – Daten – ist universell anwendbar und für andere Produkte denkbar. Durch die polarisierende Auswertung der Rezensionen entstanden zwei möglichst kontrastreiche Modelle, mit denen die Unterschiede bei der Wortwahl bei unzufriedenen und zufriedenen Kunden zu bestimmten Produkteigenschaften herausgearbeitet werden konnten. Diese Modelle sind umso brauchbarer, je größer die extrahierte Menge an Rezensionstexten ist. In dieser Arbeit wurde nur die Produktgruppenebene betrachtet, welches nur unzureichend Rückschlüsse auf einzelne Hersteller oder gar Produkte zulässt. Eine größere Datenmenge gekoppelt mit einer entsprechenden Adaption der Methodik durch einen sensibleren Algorithmus (beispielsweise FastText), könnte hier Abhilfe schaffen.

## 6.2 Methodik

Es werden die Einschränkungen und Grenzen der Vorgehensweise allgemein und je Zwischenschritt zu diskutiert. Zunächst generell ist das Wasserfall - Modell als ein lineares Modell zu verstehen, dennoch ist es während der Bearbeitung nötig gewesen, gedanklich zwischen den Stufen zu springen und Anpassungen im Arbeitsprozess vorzunehmen. Auch innerhalb der Stufen fanden ständig, auch durch Diskussionen mit dem Betreuer, Optimierungen statt.

### 6.2.1 Datenextraktion & Preprocessing

Es wurde versucht, durch das Durcharbeiten der sechs Ökodesignrichtlinien möglichst nur die Wörter zu extrahieren, die tatsächlich sinngemäß in dem Großteil der Richtlinien so auch zu

finden waren. Es konnte dabei weder sicher gestellt werden, dass wirklich alle relevanten Wörter extrahiert wurden, noch Wörter extrahiert wurden, die dem Ökodesign nicht direkt zuzuordnen sind. Dies gilt auch für das spätere Screening der Rezensionen, bei dem kontextbasiert die Wörter ausgesucht wurden.

Weiterhin kann nicht davon ausgegangen werden, dass die Stichprobe der Smartphones repräsentativ für z. B: die in Deutschland verkauften Smartphones stehen kann. Es handelt sich um eine Fallstudie über die Rezensionen der 50 meistverkauften Geräte auf Amazon, deren Inhalt quantitativ hinsichtlich möglicher Nachhaltigkeitskriterien ausgewertet wurde. Dabei muss auch die Frage beachtet werden, wer mögliche Rezensenten sind und warum die Menschen überhaupt die Produktbewertungen schreiben. Man kann unterstellen, dass nur wenige wirklich umweltbewusste Menschen dort Produkte auf ihr ökologisch ausgerichtetes Design bewerten. Dafür sollten ggfs. andere Datenquellen, wie utopia.de, fairphone Händler o.ä., mit integriert werden, was in dem Design dieser Arbeit aber nicht geplant war, was aber für eine zukünftige Optimierung der Methodik kein Hindernis darstellen sollte.

Außerdem können mit der aktuellen Methode aufgrund der sehr ungleich verteilten Daten- bzw. Rezensionsmenge je Produkt zum aktuellen Zeitpunkt nur auf Produktgruppenebene solche positiv und negativ Modelle erstellt werden. Der beste Fall, dies bis auf die Produkt- bzw. Herstellerebene herunterzubrechen, und dies ggfs. Kunden sichtbar zu machen, kann zukünftig mit der Kombination verschiedener Datenquellen geschehen.

Ein weiteres Hindernis bei der Datenextraktion sind zurzeit nicht eindeutig geklärte urheberrechtliche Fragen auf europäischer Ebene. Demnach wurde im Herbst 2017 von der Europäischen Kommission vorgeschlagen, das Urheberrecht in Europa entsprechend anzugleichen. Dabei soll es nur unter expliziter Erlaubnis der jeweiligen Urheber möglich sein, über TDM ermittelte Daten für wissenschaftliche Zwecke zu nutzen bzw. darf deren Veröffentlichung auf keinem Weg kommerziellen Nutzen haben. Ein Zusammenschluss aus Unternehmen „European Alliance for Research Excellence“ (EARE), die von den Ergebnissen der Forschungsinstitutionen im Bereich TDM abhängig sind, haben sich in offenen Briefen an die Europäische Kommission gewandt, um öffentliche und private Forschung weiterhin wettbewerbsfähig zu halten und nicht „genau“ an dem Zeitpunkt, an dem andere Länder der Welt die Barrieren für Datenexploration herunterfahren. (Coulet 2018)

Aktuell gilt dazu in Deutschland *§60d Text und Data Mining* im Urheberrechtsgesetz (UrhG), dessen Text allerdings von verschiedenen Seiten interpretiert werden kann. Demnach ist es:

> „(1)[…] zulässig, das Ursprungsmaterial auch automatisiert und systematisch zu vervielfältigen, […]“ sowie

> „2. Das Korpus einem bestimmt abgegrenzten Kreis von Personen für die gemeinsame wissenschaftliche Forschung sowie einzelnen Dritten zur Überprüfung der Qualität wissenschaftlicher Forschung öffentlich zugänglich zu machen. Der Nutzer darf hierbei nur nicht kommerzielle Zwecke verfolgen.“ und
>
> „(3) Das Korpus und die Vervielfältigungen des Ursprungsmaterials sind nach Abschluss der Forschungsarbeiten zu löschen; die öffentliche Zugänglichmachung ist zu beenden. Zulässig ist es jedoch, das Korpus und die Vervielfältigungen des Ursprungsmaterials den in den §§ 60e und 60f genannten Institutionen zur dauer-haften Aufbewahrung zu übermitteln.“

In wie weit die §§60e, f (Bibliotheken und öffentliche Archive) auch Onlinebibliotheken sein können und welche Auswirkungen das hat, bleibt im Gesetzestext offen. (Pachali 2018)

### 6.2.2 Datenanalyse

Die Ergebnisse der Datenanalyse zeigen generell, dass es mittels Text Mining Methoden möglich ist, das Feedback in Form von Rezensionen auf Produktgruppenebene durch Kontextwörter bestimmter Schlüsselwörter quantitativ zu erfassen und entsprechend der Kriterien des nachhaltigen Produktdesigns auszuwerten. Dazu wurden zunächst Schlüsselwörter extrahiert und Modelle erzeugt, um den Kontext der Schlüsselwörter darin abzufragen.

**Häufigkeiten und TF-IDF**

Um einen ersten Überblick zu erhalten, ob und wie häufig die jeweiligen extrahierten Schlüsselwörter in den sehr schlechten und sehr guten Rezensionen vorkamen, wurden deren relative und absolute Häufigkeit zunächst ermittelt, um anschließend eine vereinfachte Relevanzmessung nach dem Verfahren TF - IDF durchzuführen. Diese wurde außerdem für alle Wörter des Corpus durchgeführt, wodurch ermittelt werden konnte, dass die extrahierten Wörter beispielsweise nicht unter den 50 relevantesten Wörtern im Corpus liegen. Dies war nicht der Fall, was auf eine untergeordnete Relevanz der extrahierten Wörter hindeutet. Im Folgenden wurden die Keywords analysiert. Interessant zu beobachten ist, dass die Wörter *zwei* und *jahr* sowohl in sehr guten als auch sehr schlechten Corpus ähnlich häufig genannt und als unter den Keywords sehr relevant eingestuft wurde. Es müsste folglich untersucht werden, ob diese Wörter tatsächlich oft zusammen genannt werden und welche Unterschiede sich im schlechten und guten Kontext formen. Zu vermuten ist zum einen die Vertragslaufzeit von zwei Jahren, die häufig eine Neuanschaffung eines Gerätes auslöst und auf der anderen Seite ein mögliches Versagen des Gerätes kurz nach der zweijährigen gesetzlichen Gewährleistung.

Dennoch muss es auch keinen zwangsläufigen Zusammenhang zwischen den Wörtern geben, da diese Wörter auch in einer anderen Absicht genannt wurden (Polysemie). Die Methode TF - IDF betrachtet jedes Wort für sich und demnach weder Reihenfolge noch den Kontext, in dem das Wort genannt wurde. So können auch synonyme Wörter nicht erfasst werden, was ggfs. durch eine kontinuierliche Erweiterung der Liste erledigt werden könnte. (Garcia 2006)

**Word2Vec**

Die in Kapitel 5.6 in der Tabelle dargestellten Ergebnisse zeigen, dass es möglich ist, aufgrund der entwickelten Methodik die ähnlichsten Wörter von verschiedenen Schlüsselwörtern im Bereich des nachhaltigen Produktdesigns gegenüberzustellen.

Aus den Beobachtungen können nun Thesen und Fragen für weitere empirische Untersuchungen entwickelt werden:

1. Was bedeutet eine hohe Kosinus Ähnlichkeit zu einem Wort in einem Corpus, während in dem anderen Corpus dieselbe wesentlich geringer ausfällt (Zeilen: 1 - 22)? Hat diese Beobachtung überhaupt eine Bedeutung?
2. These: Regelmäßige Probleme mit der Software und einem ungewöhnlichen Heißwerden der Geräte führt zu schlechten Bewertungen (Zeile 23 – 45).
3. *Kratzer* vs. *Macken*, *Rand*, *Gebrauchsspuren* vs. *Hänger* & *Verzögerungen*: Äußerliche Makel werden im Zusammenhang mit der Performance des Geräts gebracht. Kann dafür auf technischer Ebene eine Erklärung gefunden werden oder ist dies eine psychologisch wirkende Folge einer „Abnutzung“ (Zeile 46 – 67)?
4. Wie hoch ist die Lebenserwartung an die Geräte durch den Kunden (Zeile 68 – 89)? Erste Antworten im deutschen Raum liefern beispielsweise (Jaeger-Erben und Hipp 2017)
5. Ab welchem (Zeit-) Punkt bzw. welche Erwartungshaltung ist ausschlaggebend, dass die Akkuleistung nicht (mehr) als zufriedenstellend beschrieben wird (Zeile 90 – 115)?
6. Eine hohe Relevanz und Anfälligkeit für Schäden nach Stürzen wird in den schlechten Bewertungen deutlich, während, solange das Display (-glas) intakt ist, die Features gelobt werden. Ein defektes Display könnte demnach zum Ersetzen des Gerätes führen, wie bereits untersucht wurde (Venkitachalam et al. 2015). Hier muss das Feedback in den Produktentwicklungsprozess einfließen, um Geräte robuster zu machen.
7. Ein austauschbarer Akku (Z. 142 – 153) wird von den Geräten sehr gelobt, was dafürspricht, dass die Rezensenten in diesem Fall implizit einen Nachhaltigkeitsfaktor beschreiben und deren positive Eigenschaften hervorheben.
8. Im Kontext *fair* wird das Preis/Leistungsverhältnis des Produktes beschrieben. Der ursprünglich in den Ökodesignrichtlinien intendierte Bezug zu bspw. einer *fairen Produktionsweise* scheint hier noch keine Rolle zu spielen.

Der Forschungsstand im Bereich der Wortrepräsentationen durch Vektoren in Vektorraummodellen hat sich seit 2013 erheblich entwickelt (Wu et al. 2017). Dabei geht es aber meistens um die zeitlich optimierte Anwendung des Algorithmus und die Abbildung eines besonders großen Datensatzes in einem einzigen Modell, mit dem Ähnlichkeitsuntersuchungen, wie Übersetzungen, Objekterkennung, etc. angestellt werden können. In dieser Arbeit wurden zwei Modelle parallel entwickelt, um die unterschiedliche Bedeutung und Wahrnehmung von Wörtern darzustellen, was mit Einschränkungen funktioniert hat.

Dennoch ist allgemein zu sagen, dass es scheinbar keinen exakten Weg gibt, die Word2Vec – Modelle zu evaluieren. Die Evaluation beginnt in dem Moment, an dem der Benutzer etwas Sinnvolles aus den Ergebnissen für die individuelle Anwendung findet. Es gibt entsprechende Test - Datensätze, mit denen die Leistung der Algorithmen verglichen werden kann, die Auswertung der Ergebnisse der Kosinus Distanz bzw. analog der Ähnlichkeit wird aber immer individuell entschieden. (Řehůřek 2014)

Zwei Möglichkeiten dazu funktionierten auch in dem Corpus: die *doesnt_match()* Methode konnte beispielsweise herausfinden, dass von den vier Wörtern *apple, samsung, huawei, p20* das „Wort“ p20 in diesem Fall nicht logisch in die Reihe gehört. Ähnlich bei dem Vergleich *apple + samsung – iphone = galaxy.* Auch diese „Aufgabe“ konnten die Modelle lösen, sind aber die für den in die in der Arbeit entwickelten Methodik nicht relevant.

**Datenqualität & Gütekriterien**

Nach aktuellem Stand der Forschung ist es nicht möglich nicht - überwacht gelernte Modelle objektiv, z. B. über quantitative Methoden zu evaluieren. Es gibt verschiedene Ansätze, bei denen z. B. Fehlerfunktionen optimiert werden, aber die meisten Methoden erfordern eine subjektive Evaluierung. (Lyra 2017)

Um die Qualität der Daten beurteilen zu können, sind dennoch verschiedene Szenarien denkbar, so wäre beispielsweise denkbar, die gleichen extrahierten Keywords auf andere Datensätze einzusetzen (Reliabilität), um zu prüfen, ob die gleichen oder ähnliche Ergebnisse entstehen. Da bei Amazon zurzeit die wahrscheinlich größte Bandbreite an Produktrezensionen gesammelt ist, wäre maximal eine Erweiterung der Modelle mithilfe anderer Quellen vorstellbar. (Hellrich und Hahn 2017)

Durch die statistische Herangehensweise, nur die relevantesten Keywörter zu prüfen, kann man davon ausgehen, dass die Analyse der Daten so objektiv, wie mit dem Verfahren möglich, geschehen ist.

Mit der Auswahl der Tools und der Entwicklung der Methodik wurde versucht, verfügbare, geeignete Verfahren so miteinander zu kombinieren, dass möglichst valide Aussagen getroffen

werden konnten. Diese können in weiterführenden empirischen Untersuchungen mit den üblichen z. B. deskriptiven Methoden entsprechend der Fragestellungen evaluiert werden.

Dennoch ist festzuhalten, dass es wünschenswert gewesen wäre, eine größere und differenziertere Auswahl an schlechten Rezensionen zu erhalten, denn wie die Ergebnisse zeigen, liegen in diesen die vermuteten Schwachstellen der Produkte verborgen, die mithilfe der Methodik umso besser herausgearbeitet werden kann, je größer die Datenbasis und daraus folgende Modellgröße ist. In den schlechten Rezensionen äußern sich die Kunden außerdem nicht unbedingt über das Produkt selbst, sondern über die Verpackung, Abwicklung oder den Versand, was die Auswahl an geeigneten und relevanten Daten zusätzlich verringert und eine Differenzierung gleichzeitig erschwert.

# 7 Fazit und Ausblick

Im Rahmen dieser Arbeit konnte eine Methodik zur Identifikation und Auswertung von relevanten Schlüsselwörtern im Bereich des nachhaltigen Produktdesigns entwickelt und angewandt werden. Die Methodik wurde auf einen Textcorpus angewandt, der ca. 15.000 Rezensionen aus der Produktgruppe „Handys & Smartphones“ im deutschsprachigen Amazon - Vertriebsnetz (amazon.de) umfasste. Es konnte gezeigt werden, dass sich ausgewählte Nachhaltigkeitsaspekte quantitativ erfassen lassen, indem Wortnachbarn eines Schlüsselwortes aus zwei Modellen abgefragt wurden. Ein Modell beinhaltete dabei positive, fünf Sterne Rezensionen und das andere enthielt die negativen, ein Stern Rezensionen. Durch die Modellierung mittels Vektorrepräsentation von Wörtern konnten zu einem Schlüsselwort die semantisch nächsten oder auch „ähnlichsten“ negativen und positiven Wörter identifiziert werden. Auf dieser Basis konnten, entsprechend der Zielvorstellung einer explorativen Datenanalyse, nach einer subjektiven Evaluierung entsprechende Thesen und weitergehende Forschungsfragen formuliert werden, auf deren Bearbeitung - soweit möglich - in der Literatur verwiesen wurde, und die durch weitere empirische Versuche widerlegt oder bestätigt werden sollten.

Die Ergebnisse zeigen, dass es grundsätzlich möglich ist, aus frei verfügbaren, ambienten Daten quantitativ Informationen zu Produkteigenschaften herauszuarbeiten und interessante Ergebnisse erzielt werden können, die über einen Feedbackprozess im Sinne eines Review & Rating Managementprozesses in die Produktentwicklung einfließen können. Eine wichtige Voraussetzung für die Modellqualität ist, dass mindestens eine ausreichend große, noch zu definierende Menge an Textdaten in Form von Rezensionen zur Verfügung stehen muss, damit es zu brauchbaren Ergebnissen kommen kann. Die Evaluation der Ergebnisse obliegt dabei immer dem Betrachter, der entsprechende Schlüsse zieht.

Die in der Untersuchung entstandene Liste an Schlüsselwörtern im Rahmen von Nachhaltigkeitsaspekten für Smartphones auf Basis von Ökodesignrichtlinien und geschriebener Umgangssprache ist für viele Felder anwendbar. So können diese Wörter in erster Linie als Feedback für die Produktentwicklung dienen, aber auch in der Search Engine Optimization (SEO) sowie im Online - Marketing in Form von #Hashtags oder ähnlichem verwendet werden.

Außerdem ist die Methodik generisch auch auf andere Produktgruppen (Waschmaschinen, Kaffeevollautomaten, Notebooks, etc.) nach ähnlichem Schema anwendbar. Die verfügbare Datenmenge auf Produktgruppenebene bei Onlinehändlern wie Amazon.de ist ergiebig und es sind sicher interessante Verbindungen bis auf die Produktebene aufzudecken.

Eine ursprünglich geplante Untersuchung, einen möglichen Trend zu entdecken, der mit einem Produktalter von mehr als 730 Tagen (= 24 Monate) und damit abgelaufener gesetzlicher Gewährleistungspflicht die Bewertung (gemessen in Sternen) des Produktes abnimmt, konnte

nicht ohne weiteres durchgeführt werden. Grund dafür waren die Differenzen zwischen realer Markteinführung und dem Verkaufsstart bei Amazon, dessen Datum scheinbar häufiger wechselt, ggfs. wenn das entsprechende Gerät wieder zum Kauf verfügbar ist. Dabei sollte angenommen werden, dass das Rezensionsdatum mehr oder weniger dem Kaufdatum gleicht und die Differenz zwischen Verkaufsstart und Rezensionsdatum schließlich auf das Produktalter rückschließen lässt.

Dies war aber auf die gedachte Art und Weise nicht untersuchbar: Bei der Datenerhebung wurden nur Neugeräte bzw. als neu beworbene betrachtet und wirkliche Langzeiterfahrungen konnten mit dieser Datenbasis und Methodik nicht erfasst werden. Eine Alternative, weitergehende Verknüpfung der Datenbasis mit dem offiziellen Verkaufsstart und Informationen von einem Wiederverkaufsportal wie ebay.de mit anschließender Auswertung wären hier denkbar.

Ein solcher Trend würde dennoch z. B. die These hinterfragen, ob nach Ende einer minimalen Gewährleistung eventuelle Softwareupdates nicht mehr angeboten werden bzw. das jeweils aktuelle Betriebssystem nicht mehr ordnungsgemäß und sicher arbeitet oder sonstige Ausfälle beklagt werden. Bei Gebrauchtgeräten, also Geräte mit einer gewissen geleisteten Betriebsdauer, könnte ein solcher Vergleich einen möglichen Verschleiß aufdecken und Gegenstand weiterer Untersuchungen sein.

# VIII Anhang

## VIII.I Formatierte Keywordliste

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 14001 | 60950 | abfall | addi-tive | altmo-disch | anfor-derun-gen | anlei-tung | anti-drip-ping | antidis-krimi-nie-rung | ar-beits-bedin-gun-gen |
| aus-beu-tung | ausbildung | aus-flüss e | aus-tausc hbar | aus-tausch-barkeit | aus-wech-selbar-keit | autho-risiert | batteriela-dung | bauteil | bedie-nungs anlei-tung |
| be-schaf-fung | beschädigt | be-schäf ti-gung | biozid | brom | cad-mium | chlor | co2klimane utral | compli-ance | corpo-rate |
| daten | datenlöschu ng | de-fekt | de-mon-tage | din | diversi-tät | doku-men-tation | ecf | eg-ver-ord-nung | ein-zel-schick sal |
| elektrg | emissionen | ems | en | en62133 | en626 84 | ener-gieeffi-zienz | energiema-nagement-system | entla-dezeit | ent-nehm-bar-keit |
| erbgut-verän-dernd | erden | er-satz-teil | eti | euthro-phierung | fair | flamm-schutz mittel | fluororga-nisch | fort-pflan-zungs-gefähr-dend | fsc |
| garan-tie | gebraucht | ge-fähr-lich | ge-häuse | general-überholt | ge-sund-heit | ge-werk-schaft | gewinnung | ge-währ-leis-tung | geöff-net |

| giftig | gri | gut | halogene | herausforderung | herstellung | hochwertig | illegal | ilo | informationen |
|---|---|---|---|---|---|---|---|---|---|
| informationsbereitstellung | initiative | iso | jahre | kabel | kernarbeitsnorm | kinderarbeit | klemme | konfliktfrei | konfliktgebiete |
| konfliktrohstoffe | konstruktion | korruption | krebserregend | krebserzeugend | kritisch | kunststoff | kunststoffteile | ladestandsanzeige | ladevorgang |
| langlebigkeit | lebensdauer | lebensendemanagement | lebenszyklusanalyse | leiterplatte | lieferkette | lieferung | löschen | materialanforderungen | materialeffizienz |
| materialkodierung | menschenrechte | menschenwürdig | mindestens | missstände | mode | müll | nachhaltig | nachhaltigkeit | nachhaltigkeitsbericht |
| netztrennung | neu | nickel | nutzungsdauer | nutzungsdauerverlängerung | ozon | ozonabbau | ozonloch | pefc | peripheriegeräte |
| personenbezogen | pfandsystem | phillips | plastik | polymere | problemstofffrei | produkteinstellung | produktion | produktionsabfall | produktionsbedingungen |
| produktionsende | produktkoma | produkttod | prozessbedingt | quecksilber | reach | reagenzien | rebound | recycling | recyclinganteil |

| recyc-lingge-recht | refurbished | refur-bishe d | refur-bish-ment | reparier en | reparie rt | reset | ressource | rohma-terial | roh-stoffe |
|---|---|---|---|---|---|---|---|---|---|
| rück-nahme | rücknahme-möglichkeit | rück-nah-me-sys-tem | rück-sen-deak-tion | sachge-recht | samm-lung | sar | sar-wert | schnitt-stelle | schra uben |
| schrau-bendre-her | schrauben-zieher | sel-tene | ser-vice | sicher-heit | sicher-heits | sicher-heit-sup-dates | siegel | sie-geld-schun-gel | silber |
| sklave | smog | so | soft-ware | soft-ware-updates | stan-dardi-siert | stan-dards | standard-werkzeug | stecker | strom-ver-brauc h |
| sustain ability | system | sys-tem-frage | torx | toxizität | trans-parenz | treib-haus-gaspo-tenzial | umwelt | un-brauch bar | up-dates |
| verbes-serung | verfügbar | ver-füg-bar-keit | ver-klebt | verpa-ckung | ver-pflich-tung | ver-samm-lungs-freiheit | verschluss | ver-schrau bt | ver-skla-vung |
| ver-spre-chen | versäuerung | ver-wer-tung | viel-fältig-keit | vorsorge | wasser | was-ser-ver-brauch | weichma-cher | werks-einstel-lungen | werk-zeug |
| wie | wiederauf-ladbar | wie-der-ver-wend bar-keit | wie-der-ver-wer-tung | zerlegen | zerlegt | zu-gang | zurückge-ben | zwang sarbeit | zwei |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| zweit-nut-zung | zyklus | öf-fent-lich | öf-fent-lich-keit | öffnen | | | | | |

## VIII.II Codebausteine Webcrawler

topseller.py

```
1.  # -*- coding: utf-8 -*-
2.  import scrapy
3.  from scrapy.http import Request
4.  from scrapy.spiders import CrawlSpider, Rule
5.  from scrapy.linkextractors import LinkExtractor
6.  from scrapy.selector import HtmlXPathSelector
7.
8.
9.  from amazontopseller.items import AmazontopsellerItem
10.
11.
12. class TopsellerSpider(CrawlSpider):
13.     name = 'topseller'
14.     allowed_domains = ['amazon.de']
15.     start_urls = ['https://www.amazon.de/gp/bestsellers/ce-de/3468301']
16.    #https://www.amazon.de/gp/bestsellers/appliances/1399929031 z.B. Andere Bestseller-
    liste einfügen
17.
18. #Falls der crawler den u.g. xpath-Link (/@href wird nicht benötigt) findet, rufe Funk-
    tion parse_all_reviews auf. siehe Rule() LinkExtractor in Scrapy Dokumentation
19.     rules = (
20.
21.             Rule(
22.                 LinkExtractor(
23.                     restrict_xpaths='//li[contains(@class, "zg-item-immer-
    sion")]/span/div/span/a[contains(@class, "a-link-normal")]'
24.                                 ),callback='parse_all_reviews'
25.                 ),
26.             )
27.
28.     def parse_all_reviews(self,response):
29.         zeige_alle_link = response.xpath('//a[contains(@data-hook,"see-all-reviews-
    link-foot")]/@href').extract_first()
30.
31.         # Sammle Items asin und verkaufsstart auf Hauptseite
32.         asin = response.xpath('//li[contains(.,"ASIN: ")]/text()').extract()
33.         # slice das erste Zeichen ( = Leerzeichen ) des wiedergegebenen Strings
34.         verkaufsstart = response.xpath('//li[contains(.,"Im Angebot von Ama-
    zon.de seit:")]
    /text()').extract_first()[1:]
35.         # Finde und folge dem "Alle Rezensionen zeigen"-Button, rufe Funk-
    tion "parse_reviews" auf und übergebe die "Meta" Items asin und verkaufsstart
36.         yield response.follow(zeige_alle_link,callback=self.parse_re-
    views, meta={'asin': asin, 'verkaufsstart': verkaufsstart})
37.
38.     def parse_reviews(self, response):
39.
40.
```

```
41.         for sel in response.xpath('//div[@data-hook="review"]'):
42. # Beliebige Elemente von der Seite in Items gespeichert werden. Dabei helfen xpath Sel-
    ectors, die den HTML-Code nach den entsprechenden Bausteinen durchsucht.
43. # Der selector ermöglicht beliebige Elemente aus den Bausteinen zu extrahie-
    ren. z.B. /text (reiner Text), /@href (Links) oder auch ganze Grafiken.
44.
45. # Da mehrere Items (bei Amazon 10 Stk) in dem Bereich "review" zu fin-
    den sind, wird hier mit einer for-Schleife gearbeitet.
46. # Diese gibt je Item eine Liste aus 10 Bausteinen wieder und muss entspre-
    chend durch eine weitere for-Schleife in einzelne Strings "zerschlagen" werden:
47.             item = AmazontopsellerItem()
48.             for author in sel.xpath('.//a[contains(@data-hook, "review-au-
    thor")]/text()').extract():
49.                 item ['author'] = author
50.             for title in sel.xpath('.//a[contains(@data-hook, "review-
    title")]/text()').extract():
51.                 item ['title'] = title
52.             for rating in sel.xpath('.//i[contains(@data-hook, "review-star-ra-
    ting")]/span[contains(@class, "a-icon-alt")]'):
53.                 #alternative: //div[contains(@data-hook, "review")]//i[contains(@data-
    hook, "review-star-rating")]
54.                 sterne = rating.xpath('text()').extract()
55.                 sterne = [item.replace(",0 von 5 Sternen","") for item in sterne] #re-
    move constant part of the string, für englische Seiten anpassen und slicen [:1]?
56.                 item ['rating'] = sterne
57.             for review_date in sel.xpath('.//span[contains(@data-hook, "review-
    date")]'):
58.                 datum =  review_date.xpath('text()').extract()
59.                 datum = [item.replace("am ","") for item in datum] #remove con-
    stant part of the string, für englische Seiten anpassen und slicen [:1]?
60.                 item ['review_date'] = datum
61.             for found_helpful in sel.xpath('.//span[contains(@data-hook, "helpful-vote-
    statement")]'):
62.                 helpful = found_helpful.xpath('text()').extract()
63.                 helpful = [item.replace(" Personen fanden diese Informationen hilf-
    reich","").replace("Eine Person fand diese Informationen hilfreich","1") for i-
    tem in helpful] #remove constant part of the string, für englische Seiten anpas-
    sen und slicen [:1]?
64.                 item ['found_helpful'] = helpful
65.             for review_body in sel.xpath('string(.//span[contains(@data-hook, "review-
    body")])').extract():
66.                                     # Der gesamte Reviewbereich wird nur durch die Um-
    wandlung in einen String erreicht. Das reine Extrahieren nimmt auch html-
    tags wie <br> oder <n> mit.
67.                 item ['review_body'] = review_body
68.                 item ['review_length'] = len(review_body) # Anzahl der Zei-
    chen im String "review_body" = Länge der Rezension in Zeichen
69.
70. # Da wir uns auf derselben Seite befinden, wird der Produktname und der Link nur ein-
    mal abgerufen.
71.
72.             item ['product'] = response.xpath('normalize-space(//h1/a)').extract()
73.
74. # Die relative URL des Produkts wird mit der Haupt-URL zusammengefügt:
75.
76.             abs_url = response.urljoin(response.xpath('//*[@id="cm_cr-pro-
    duct_info"]//a[contains(@data-hook, "product-link")]/@href').extract_first())
77.             item ['abs_url'] = abs_url
78.
79.             item ['europreis'] = response.xpath('normalize-space(//span[con-
    tains(@class,"arp-price")])').extract_first().re-
    place("\xa0€","") # nur das die reine Zahl extrahieren, um damit einfacher arbei-
    ten zu können
80.
```

```
81.             asin = response.meta['asin'] #Aufrufen der zwischengespeicherten "Meta"-
    Items aus der übergeordneten Funktion "parse_all_reviews"
82.             item ['asin'] = asin
83.
84.             verkaufsstart = response.meta['verkaufsstart'] #siehe asin.
85.             item ['verkaufsstart'] = verkaufsstart
86.
87.             #yield ist eine besondere Form der Wiedergabe in py-
    thon, die in der for schleife bleibt (stichwort: generator = iterable). ein re-
    turn würde die funktion nach einmaligem durchlauf beenden.
88.             yield item
89.
90.         #suche und finde den "nächste Seite" Button, folge ihm solange es mög-
    lich ist. Bemerke dass dieser Teil außerhalb der großen for-schleife befin-
    det. also erst nachdem alle items gecrawlt wurden, wird nach dem butten gesucht.
91.         next_page_url = response.xpath('//li[@class="a-
    last"]/a/@href').extract_first()
92.         if next_page_url:
93.         #next_page_url = response.urljoin(next_page_url)
94.             yield response.follow(next_page_url, callback=self.parse_re-
    views, meta={'asin': asin , 'verkaufsstart': verkaufsstart})
95.
96. # Siehe settings.py und items.py für die übergeordneten Einstellungen und Defini-
    tion der items!
```

items.py

```
1.  import scrapy
2.
3.
4.  class AmazontopsellerItem(scrapy.Item):
5.    # define the fields for your item here like:
6.    # name = scrapy.Field()
7.    product = scrapy.Field()
8.    abs_url = scrapy.Field()
9.    author = scrapy.Field()
10.   title = scrapy.Field()
11.   rating = scrapy.Field()
12.   review_date = scrapy.Field()
13.   found_helpful = scrapy.Field()
14.   review_body = scrapy.Field()
15.   asin = scrapy.Field()
16.   verkaufsstart = scrapy.Field()
17.   europreis = scrapy.Field()
18.   review_length = scrapy.Field()
19.   #pass
```

settings.py

```
1.  # -*- coding: utf-8 -*-
2.
3.  # Scrapy settings for amazontopseller project
4.  #
5.  # For simplicity, this file contains only settings considered important or
6.  # commonly used. You can find more settings consulting the documentation:
7.  #
8.  #     http://doc.scrapy.org/en/latest/topics/settings.html
9.  #     http://scrapy.readthedocs.org/en/latest/topics/downloader-middleware.html
10. #     http://scrapy.readthedocs.org/en/latest/topics/spider-middleware.html
11.
12. BOT_NAME = 'amazontopseller'
13.
14. SPIDER_MODULES = ['amazontopseller.spiders']
15. NEWSPIDER_MODULE = 'amazontopseller.spiders'
16.
```

```
#DUPEFILTER_DEBUG = True
#UTF-8 in Json Format aktivieren. Achtung: Die normale Darstellung ist korrekt und wird von jeder Datenbank verstanden.
#Aus Kosmetikgründen kann man hier den utf-8 Standard speziell für json aktivieren.
#https://stackoverflow.com/questions/39582409/how-can-i-get-an-output-in-utf-8-encoded-unicode-from-scrapy

#FEED_EXPORT_ENCODING = 'utf-8'

# Crawl responsibly by identifying yourself (and your website) on the user-agent
USER_AGENT = "Mozilla/5.0 (Windows NT 6.2; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/27.0.1453.93 Safari/537.36"

# Obey robots.txt rules
ROBOTSTXT_OBEY = True

# Configure maximum concurrent requests performed by Scrapy (default: 16)
#CONCURRENT_REQUESTS = 32

# Configure a delay for requests for the same website (default: 0)
# See http://scrapy.readthedocs.org/en/latest/topics/settings.html#download-delay
# See also autothrottle settings and docs
 DOWNLOAD_DELAY = 1
# The download delay setting will honor only one of:
#CONCURRENT_REQUESTS_PER_DOMAIN = 16
#CONCURRENT_REQUESTS_PER_IP = 16

# Disable cookies (enabled by default)
#COOKIES_ENABLED = False

# Disable Telnet Console (enabled by default)
#TELNETCONSOLE_ENABLED = False

# Override the default request headers:
#DEFAULT_REQUEST_HEADERS = {
#   'Accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8',
#   'Accept-Language': 'en',
#}

# Enable or disable spider middlewares
# See http://scrapy.readthedocs.org/en/latest/topics/spider-middleware.html
#SPIDER_MIDDLEWARES = {
#    'amazon_de_reviews.middlewares.AmazonDeReviewsSpiderMiddleware': 543,
#}

# Enable or disable downloader middlewares
# See http://scrapy.readthedocs.org/en/latest/topics/downloader-middleware.html
#DOWNLOADER_MIDDLEWARES = {
#    'amazon_de_reviews.middlewares.MyCustomDownloaderMiddleware': 543,
#}

# Enable or disable extensions
# See http://scrapy.readthedocs.org/en/latest/topics/extensions.html
#EXTENSIONS = {
#    'scrapy.extensions.telnet.TelnetConsole': None,
#}

# Configure item pipelines
# See http://scrapy.readthedocs.org/en/latest/topics/item-pipeline.html
#ITEM_PIPELINES = {
#    'amazon_de_reviews.pipelines.AmazonDeReviewsPipeline': 300,
#}

# Enable and configure the AutoThrottle extension (disabled by default)
# See http://doc.scrapy.org/en/latest/topics/autothrottle.html
```

```
79. AUTOTHROTTLE_ENABLED = True
80. # The initial download delay
81. #AUTOTHROTTLE_START_DELAY = 5
82. # The maximum download delay to be set in case of high latencies
83. AUTOTHROTTLE_MAX_DELAY = 60
84. # The average number of requests Scrapy should be sending in parallel to
85. # each remote server
86. AUTOTHROTTLE_TARGET_CONCURRENCY = 1.0
87. # Enable showing throttling stats for every response received:
88. #AUTOTHROTTLE_DEBUG = False
89.
90. # Enable and configure HTTP caching (disabled by default)
91. # See http://scrapy.readthedocs.org/en/latest/topics/downloader-middleware.html#http-
    cache-middleware-settings
92. #HTTPCACHE_ENABLED = True
93. #HTTPCACHE_EXPIRATION_SECS = 0
94. #HTTPCACHE_DIR = 'httpcache'
95. #HTTPCACHE_IGNORE_HTTP_CODES = []
96. #HTTPCACHE_STORAGE = 'scrapy.extensions.httpcache.FilesystemCacheStorage'
```

pipelines.py, middlewares.py, __init__.py unverändert!

## VIII.III DVD / CD

Auf der beigelegten DVD / CD befinden sich:

- Digitale Version der Arbeit im PDF - Format
- Top 50 Artikel Handy & Smartphone von Amazon.de (Datenbank im Excelformat)
- Tableau Projektmappe
- Diverse Programmcodes
    - Scrapy Crawler
    - Pre Processing und Anwendung word2vec
    - TF - IDF

# 8 Literaturverzeichnis

Allahyari, Mehdi; Pouriyeh, Seyedamin; Assefi, Mehdi; Safaei, Saied; Trippe, Elizabeth D.; Gutierrez, Juan B.; Kochut, Krys (2017): A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. Online verfügbar unter http://arxiv.org/pdf/1707.02919v2.

Apple Inc. (2017): iPhone X Environmental Report. Online verfügbar unter https://www.apple.com/environment/pdf/products/iphone/iPhone_X_PER_sept2017.pdf, zuletzt geprüft am 23.10.2018.

Bakker, Conrad (2014): Products that last. Product design for circular business models. Delft: TU Delft.

Beyreuther, Tabea; Duske, Katrin; Eismann, Christian; Hornung, Sabine; Kleemann, Frank (2012): consumerswork. Zum neuen Verhältnis von Unternehmen und Usern im Web 2.0. 1. Aufl. Frankfurt am Main: Campus Verlag (Sozialwissenschaften 2012). Online verfügbar unter http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=832225.

Bojanowski, Piotr; Grave, Edouard; Joulin, Armand; Mikolov, Tomas (2016): Enriching Word Vectors with Subword Information. Online verfügbar unter https://arxiv.org/pdf/1607.04606.pdf, zuletzt geprüft am 04.11.2018.

Coulet, Amelie (2018): The European Parliament must improve the Text and Data Mining (TDM) exception to benefit European research and innovation. Hg. v. European Alliance for Research Excellence. Online verfügbar unter http://www.eifl.net/sites/default/files/resources/openletter-to-juri-committee-on-tdm_26march2018.pdf, zuletzt geprüft am 30.10.2018.

Dale, Kyran (2016): Data visualization with Python and JavaScript. Scrape, clean, explore & transform your data. Sebastopol, CA: O'Reilly Media. Online verfügbar unter http://proquest.tech.safaribooksonline.de/9781491920565.

Dhar, Vasant (2013): Data science and prediction. In: *Commun. ACM* 56 (12), S. 64–73. DOI: 10.1145/2500499.

Diehl, Joerg M.; Arbinger, Roland (2001): Einführung in die Inferenzstatistik. 3., korrigierte Aufl. Eschborn bei Frankfurt am Main: Klotz.

Drees, Bastian (2016): Text und Data Mining: Herausforderungen und Möglichkeiten für Bibliotheken. Perspektive Bibliothek, Bd. 5, Nr. 1 (2016). DOI: 10.11588/PB.2016.1.33691.

Dubin, David (2004): The Most Influential Paper Gerard Salton Never Wrote. In: *Library Trends 52(4)* 52 (4), S. 748–764. Online verfügbar unter http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.184.910&rep=rep1&type=pdf, zuletzt geprüft am 29.10.2018.

Fahrmeir, Ludwig; Heumann, Christian; Künstler, Rita; Pigeot, Iris; Tutz, Gerhard (2016): Statistik. Der Weg zur Datenanalyse. 8., überarbeitete und ergänzte Auflage. Berlin, Heidelberg: Springer Spektrum (Springer-Lehrbuch). Online verfügbar unter http://dx.doi.org/10.1007/978-3-662-50372-0.

Fahrner, Ulrich (2011): Die Explorative Datenanalyse als Lern- und Erkenntniswerkzeug. Online verfügbar unter https://www.researchgate.net/profile/Ulrich_Fahrner/publication/278383090_Die_Explorative_Datenanalyse_als_Lern-_und_Erkenntniswerkzeug/links/5a6a18d8a6fdccf88497c547/Die-Explorative-Datenanalyse-als-Lern-und-Erkenntniswerkzeug.pdf?origin=publication_detail, zuletzt geprüft am 18.09.2018.

Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Pdhraic (1996): From Data Mining to Knowledge Discovery in Databases. In: *AI Magazine* 17, S. 37–54.

Fuhr, Norbert (2010): Einführung in Information Retrieval. Skriptum zur Vorlesung im SS 10. Universität Duisburg. Duisburg. Online verfügbar unter http://www.is.informatik.uni-duisburg.de/courses/ir_ss10/folien/skript_1-5.pdf, zuletzt geprüft am 05.11.2018.

Gabler (2018a): Algorithmus. Online verfügbar unter https://wirtschaftslexikon.gabler.de/definition/algorithmus-27106/version-250769, zuletzt aktualisiert am 19.02.2018, zuletzt geprüft am 10.11.2018.

Gabler (2018b): Open Source. Online verfügbar unter https://wirtschaftslexikon.gabler.de/definition/open-source-43032/version-266368, zuletzt aktualisiert am 10.11.2018.

Gaikwad, Sonali Vijay; Chaugule, Archana; Patil, Pramod (2014): Text Mining Methods and Techniques. In: *International Journal of Computer Applications* 85 (17), S. 42–45. Online verfügbar unter https://pdfs.semanticscholar.org/9fde/fcb2bd8f4781afcf8ceac02864d6bb76f357.pdf, zuletzt geprüft am 30.10.2018.

Garcia, E. (2006): The Classic TF-IDF Vector Space Model. Online verfügbar unter http://www.minerazzi.com/tutorials/term-vector-3.pdf, zuletzt aktualisiert am 19.03.2016, zuletzt geprüft am 05.11.2018.

Gartner (2018): Global market share held by smartphone operating systems 2009-2018 by quarter. statista.com. Online verfügbar unter https://www.statista.com/statistics/266136/global-market-share-held-by-smartphone-operating-systems/, zuletzt geprüft am 04.10.2018.

Gillis, Nancy; Brody-Heine, Pamela: Green Electronics Council's Dynamic Standards Development Process. Unter Mitarbeit von Green Electronics Council. Online verfügbar unter https://greenelectronicscouncil.org/wp-content/uploads/2018/10/DSDP-Overview-September-17-2018.pdf, zuletzt geprüft am 04.10.2018.

Goldberg, Yoav; Levy, Omer (2014): word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. Online verfügbar unter http://arxiv.org/pdf/1402.3722v1.

Goodfellow, Ian; Bengio Yoshua; Courville, Aaron (2016): Deep Learning. Hg. v. MIT Press. Online verfügbar unter https://www.deeplearningbook.org/contents/linear_algebra.html, zuletzt aktualisiert am 26.10.2018, zuletzt geprüft am 30.10.2018.

Grossman, David A.; Frieder, Ophir (2004): Information retrieval. [algorithms and heuristics]. 2nd ed., [repr.]. Himayatnagar, Hyderabad: Universities Press (India) (The Kluwer international series on information retrieval).

Gupta, Saurabh; Khare, Vineet (2018): Amazon SageMaker BlazingText: Parallelizing Word2Vec on Multiple CPUs or GPUs | Amazon Web Services. Amazon Web Services. Online verfügbar unter https://aws.amazon.com/de/blogs/machine-learning/amazon-sagemaker-blazingtext-parallelizing-word2vec-on-multiple-cpus-or-gpus/, zuletzt aktualisiert am 14.09.2018, zuletzt geprüft am 04.11.2018.

Han, Jiawei; Kamber, Micheline (2010): Data mining. Concepts and techniques. 2. ed., [Nachdr.]. Amsterdam: Elsevier/Morgan Kaufmann (The Morgan Kaufmann series in data management systems).

Hardeniya, Nitin (2015): NLTK essentials. Build cool NLP and machine learning applications using NLTK and other Python libraries. Birmingham, UK: Packt Publishing (Community experience distilled). Online verfügbar unter http://proquest.tech.safaribooksonline.de/9781784396909.

Harris, Zellig S. (1954): Distributional Structure. In: *WORD* 10 (2-3), S. 146–162. DOI: 10.1080/00437956.1954.11659520.

Hellrich, Johannes; Hahn, Udo (2017): Don't Get Fooled by Word Embeddings. Better Watch their Neighborhood. Online verfügbar unter https://dh2017.adho.org/abstracts/487/487.pdf, zuletzt geprüft am 08.11.2018.

IBM (2014): Performance and Capacity Implications for Big Data. In: *Modern Philology* 112 (2), Front Cover-Front Cover. DOI: 10.1086/679571.

IDC (2018a): Global smartphone shipments 2010-2022 (in million units). Hg. v. statista.com. Online verfügbar unter https://www.statista.com/statistics/263441/global-smartphone-shipments-forecast/, zuletzt geprüft am 13.11.2018.

IDC (2018b): Global smartphone shipments from 2009 to 2018 (in million units). Statista - The Statistics Portal. Online verfügbar unter https://www.statista.com/statistics/271491/worldwide-shipments-of-smartphones-since-2009/., zuletzt geprüft am 23.10.2018.

J.R. Firth (1957): A synopsis of linguistic theory. Oxford: Blackwell (Studies in linguistic analysis).

Jaeger-Erben, Melanie; Hipp, Tamina (2017): Letzter Schrei oder langer Atem? - Erwartungen und Erfahrungen im Kontext von Langelebigkeit bei Elektronikgeräten. Deskriptive Auswertung einer repräsentativen Online-Befragung in Deutschland. Hg. v. Nachwuchsgruppe Obsoleszenz.

BMBF (1/2017). Online verfügbar unter https://challengeobsolescence.info/aktuelles/letzter-schrei-oderlanger-atem, zuletzt geprüft am 23.10.2018.

Jansen, Stefan (2018): Word and Phrase Translation with word2vec. Online verfügbar unter http://arxiv.org/pdf/1705.03127v4, zuletzt geprüft am 22.10.2018.

Jardim, Elizabeth (2017): Report: 10 Jahre Smartphone | Greenpeace. Hg. v. Greenpeace Inc. Washington, D.C. Online verfügbar unter https://www.greenpeace.de/sites/www.greenpeace.de/files/publications/s01981_greenpeace_report_10_jahre_smartphone.pdf, zuletzt geprüft am 24.10.2018.

Jean, P.; Coulon, R.; Timmons, D. (1999): Building an EcoDesign toolkit for the electronics industry. In: Hiroyuki Yoshikawa (Hg.): Proceedings. First International Symposium on Environmentally Conscious Design and Inverse Manufacturing : February 1-3, 1999, Tokyo, Japan. Proceedings First International Symposium on Environmentally Conscious Design and Inverse Manufacturing. Tokyo, Japan, 2/3/1999 - 2/3/1999. International Symposium on Environmentally Conscious Design and Inverse Manufacturing; Inverse Manufacturing Forum, Japan; IEEE Computer Society; Japan Environment Management Association for Industry. Los Alamitos, Calif: IEEE Computer Society, S. 701–706.

Kaggle Inc. (2014): Bag of Words Meets Bags of Popcorn. Online verfügbar unter https://www.kaggle.com/c/word2vec-nlp-tutorial#part-1-for-beginners-bag-of-words, zuletzt geprüft am 08.10.2018.

Khasanah, Annisa Uswatun; Erlangga, Deliana Ardhitama; Jamil, Ahmad Mustopa (2018): An application of data mining techniques in designing catalogue for a laundry service. In: *MATEC Web Conf.* 154 (3), S. 1099. DOI: 10.1051/matecconf/201815401099.

Kouzis-Loukas, Dimitrios (2016): Learning Scrapy. Learn the art of efficient web scraping and crawling with Python. Birmingham, UK: Packt Publishing (Community experience distilled). Online verfügbar unter http://proquest.tech.safaribooksonline.de/9781784399788.

Kreutzer, Ralf T. (2017): Review- und Rating-Management – ein (noch) unterschätztes Aufgabenfeld. In: Dialogmarketing Perspektiven 2016/2017. Tagungsband 11. wissenschaftlicher interdisziplinärer Kongress für Dialogmarketing. Wiesbaden: Springer Gabler (Research), S. 109–126.

Le, Quoe; Mikolov, Tomas (2014): Distributed Representations of Sentences and Documents. Google.com (Proceedings of Machine Learning Research, 32). Online verfügbar unter https://cs.stanford.edu/~quocle/paragraph_vector.pdf, zuletzt aktualisiert am 16.07.2018, zuletzt geprüft am 04.11.2018.

Lekovec (2013): SNAP: Web data: Amazon reviews. Stanford University. Online verfügbar unter https://snap.stanford.edu/data/web-Amazon.html, zuletzt geprüft am 24.10.2018.

Leopold, Edda; Kindermann, Jörg (2002): Text Categorization with Support Vector Machines. How to Represent Texts in Input Space? In: *Machine Learning* 46 (1/3), S. 423–444. DOI: 10.1023/A:1012491419635.

Liu, Yang; Liu, Zhiyuan; Chua, Tat-Seng; Sun, Maosong (Hg.) (2015): Topical Word Embeddings. 29th AAAI Conference on Artificial Intelligence. Austin, 25/01/2015.

Lyra, Matti (2017): Evaluating Unsupervised Models. Hg. v. GitHub.com. Online verfügbar unter https://github.com/mattilyra/pydataberlin-2017/blob/master/notebook/EvaluatingUnsupervisedModels.ipynb, zuletzt geprüft am 30.10.2018.

Maimon, Oded Z.; Rokach, Lior (2010): Data mining and knowledge discovery handbook. 2. ed. New York: Springer. Online verfügbar unter http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10412834.

Manning, Christopher D.; Raghavan, Pabhakar; Schütze, Hinrich (2008): Introduction to Information Retrieval. An example information retrieval problem. Hg. v. Cambridge University Press. Stanford University. Online verfügbar unter https://nlp.stanford.edu/IR-book/html/htmledition/an-example-information-retrieval-problem-1.html#928, zuletzt aktualisiert am 08.04.2009, zuletzt geprüft am 12.11.2018.

Marwede, Max; Jokinen, Tapani (2018): Learning Factory Ecodesign Sprint for Circular Economy. Hg. v. Umweltbundesamt. Learning Factory. Dessau. Online verfügbar unter https://www.umweltbundesamt.de/sites/default/files/medien/376/dokumente/marwede_et_al_lernfabrik-workshop_4.pdf, zuletzt geprüft am 23.10.2018.

MathWorks.com (2018): Machine Learning: Drei Dinge, die Sie wissen sollten. Online verfügbar unter https://de.mathworks.com/discovery/machine-learning.html, zuletzt aktualisiert am 02.11.2018, zuletzt geprüft am 02.11.2018.

McAuley, Julian (2014): Amazon review data. UC San Diego. San Diego. Online verfügbar unter http://jmcauley.ucsd.edu/data/amazon/, zuletzt aktualisiert am 04.09.2018, zuletzt geprüft am 24.10.2018.

McCabe, Chris (2018): Amazon’s Fake Review Problem is Worse Than Ever. Here’s Why. WebRetailer.com. Online verfügbar unter https://www.webretailer.com/lean-commerce/amazon-fake-reviews/#/, zuletzt geprüft am 24.10.2018.

Mikolov, Tomas; Chen, Kai; Corrado, Greg; Dean, Jeffrey (2013): Efficient Estimation of Word Representations in Vector Space. Online verfügbar unter http://arxiv.org/pdf/1301.3781v3.

Miner, Gary; Nisbet, Robert; Fast, Andrew; Waner, Angela; Delen, Dursun; Thompson, Jennifer et al. (2012): Practical text mining and statistical analysis for non-structured text data applications. Waltham, MA: Academic Press. Online verfügbar unter http://proquest.tech.safaribooksonline.de/9780123869791.

Mitra, Bashkar (2016): Using Text Embeddings for Information Retrieval. Glasgow. Online verfügbar unter https://www.slideshare.net/BhaskarMitra3/using-text-embeddings-for-information-retrieval?next_slideshow=1, zuletzt geprüft am 17.10.2018.

Mitra, Bhaskar; Craswell, Nick (2017): Neural Models for Information Retrieval. Online verfügbar unter https://arxiv.org/pdf/1705.01509.pdf, zuletzt geprüft am 04.11.2018.

Mohanty, Hrushikesha (Hg.) (2015): Big data. A primer. New Delhi: Springer (Studies in big data, 11).

Morik, Katharina; Bhaduri, Kanishka; Kargupta, Hillol (2012): Introduction to data mining for sustainability. In: *Data Min Knowl Disc* 24 (2), S. 311–324. DOI: 10.1007/s10618-011-0239-5.

Morley, Janine; Widdicks, Kelly; Hazas, Mike (2018): Digitalisation, energy and data demand: The impact of Internet traffic on overall and peak electricity consumption. In: *Energy Research & Social Science* 38, S. 128–137. DOI: 10.1016/j.erss.2018.01.018.

Orellana, Vanessa Hand (2017): iPhone X drop test: It cracked on the first drop. https://www.facebook.com/cnet. Online verfügbar unter https://www.cnet.com/news/apple-iphone-x-drop-test/, zuletzt geprüft am 12.11.2018.

Pachali, David (2018): Benjamin Raue: Text- und Data-Mining nicht weiter beschränken. Hg. v. iRights.info. Online verfügbar unter https://irights.info/artikel/benjamin-raue-text-und-data-mining-nicht-weiter-beschraenken/29014, zuletzt geprüft am 30.10.2018.

Pennington, Jeffrey; Socher, Richard; D.Manning, Christopher (2014): GloVe: Global Vectors for Word Representation. Computer Science Department, Stanford University. Stanford. Online verfügbar unter https://www.aclweb.org/anthology/D14-1162, zuletzt geprüft am 04.11.2018.

Press, Gil (2013): A Very Short History Of Big Data. Forbes Magazine. Online verfügbar unter https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#247ca25065a1, zuletzt aktualisiert am 05.09.2013, zuletzt geprüft am 24.10.2018.

Pustejovsky, J.; Stubbs, Amber (2012): Natural Language Annotation for Machine Learning. Cambridge, LaVergne: O'Reilly Media, Incorporated; Ingram Publisher Services [distributor].

RAL gGmbH (2013): Blauer Engel - Das Umweltzeichen. Mobiltelefone - DE-UZ 106. Vergabekriterien. Bundesministerium für Umwelt, Naturschutz und nukleare Sicherheit. Bonn. Online verfügbar unter https://produktinfo.blauer-engel.de/uploads/criteriafile/de/DE-UZ%20106-201302-de%20Kriterien.pdf, zuletzt geprüft am 04.10.2018.

Řehůřek, Radim (2014): Word2vec Tutorial. RARE Technologies. Online verfügbar unter https://rare-technologies.com/word2vec-tutorial/, zuletzt aktualisiert am 02.02.2014, zuletzt geprüft am 05.11.2018.

Rensink, J. M. (2013): What motivates people to write online reviews and which role does personality play? A study providing insights in the influence of seven motivations on the involvement to write positive and negative online reviews and how five personality traits play a role.

University of Twente, Twente. Marketing Communication. Online verfügbar unter https://essay.utwente.nl/63536/1/Rensink_Maarten_-s_0176486_scriptie.pdf, zuletzt geprüft am 24.10.2018.

Review Honesty Inc. (2018): How Accurate is ReviewMeta.com. Online verfügbar unter https://reviewmeta.com/blog/how-accurate-is-reviewmeta-com/, zuletzt aktualisiert am 16.05.2018, zuletzt geprüft am 02.11.2018.

Robertson, Stephen (2004): Understanding inverse document frequency: on theoretical arguments for IDF. In: *Journal of Documentation* 60 (5), S. 503–520. DOI: 10.1108/00220410410560582.

Rogall, Holger (2009): Nachhaltige Ökonomie. Ökonomische Theorie und Praxis einer Nachhaltigen Entwicklung. Marburg: Metropolis (Grundlagen der Wirtschaftswissenschaft, 15), zuletzt geprüft am 23.10.2018.

Rong, Xin (2016): word2vec Parameter Learning Explained. Online verfügbar unter http://arxiv.org/pdf/1411.2738v4, zuletzt geprüft am 22.10.2018.

scrapinghub.com (2014): Scrapy | A Fast and Powerful Scraping and Web Crawling Framework. Online verfügbar unter https://scrapy.org/, zuletzt aktualisiert am 03.09.2018, zuletzt geprüft am 31.10.2018.

Sens, Irina; Katerbow, Matthias; Schöch, Christof; Mittermaier, Bernhard (2015): Zusammenfassung Workshop Und Umfrageergebnisse "Bedarf Und Anforderungen An Ressourcen Für Text Und Data Mining". Hg. v. Zenodo. Online verfügbar unter http://doi.org/10.5281/zenodo.32584, zuletzt geprüft am 30.10.2018.

Sievers, Uwe (2017): Sensoren im Smartphone geben Daten preis. Hg. v. iRights.info. Online verfügbar unter https://mobilsicher.de/hintergrund/sensoren-im-smartphone-geben-daten-preis, zuletzt aktualisiert am 09.10.2018, zuletzt geprüft am 29.10.2018.

Spärck Jones, Karen (2004): A statistical interpretation of term specificity and its application in retrieval. In: *Journal of Documentation* 60 (5), S. 493–502. DOI: 10.1108/00220410410560573.

Statworx (2018): Data Science Tools. http://www.facebook.com/statworx. Online verfügbar unter https://www.statworx.com/de/data-science/data-science-tools/, zuletzt aktualisiert am 13.11.2018, zuletzt geprüft am 13.11.2018.

Stiftung Warentest (2018): Sieger und Sensible. In: *test*, 2018 (11). Online verfügbar unter https://www.test.de/Apple-iPhone-XS-und-XS-Max-Luxushandys-versagen-bei-Falltest-5384370-0/, zuletzt geprüft am 10.11.2018.

Szegedy, Christian; Zaremba, Wojciech; Sutskever, Ilya; Bruna, Joan; Erhan, Dumitru; Goodfellow, Ian; Fergus, Rob (2013): Intriguing properties of neural networks. Online verfügbar unter http://arxiv.org/pdf/1312.6199v4.

tagesschau.de (2018): ARD-Themenwoche: So ungerecht ist Ihr Smartphone. tagesschau.de. Online verfügbar unter https://www.tagesschau.de/inland/themenwoche-gerechtigkeit-smartphone-101.html, zuletzt aktualisiert am 12.11.2018, zuletzt geprüft am 12.11.2018.

TCO Certified (2018): A more sustainable IT product choice. Hg. v. tcocertified.com. Online verfügbar unter https://tcocertified.com/files/materials/fact-sheet-tco-certified.pdf, zuletzt aktualisiert am 04.10.2018, zuletzt geprüft am 04.10.2018.

TensorFlow.org (2018): Vector Representations of Words. Online verfügbar unter https://www.tensorflow.org/tutorials/representation/word2vec, zuletzt aktualisiert am 10.10.2018, zuletzt geprüft am 17.10.2018.

Thede, Jan Tewes (2014): Ambient Data zur Trendforschung. Unter Mitarbeit von Technische Universität Berlin und Volker Trommsdorff. Berlin (Inst. Betriebswirtschaftlehre). Online verfügbar unter https://depositonce.tu-berlin.de/handle/11303/4551.

Tokic, Michel (2013): Reinforcement Learning mit adaptiver Steuerung von Exploration und Exploitation. Unter Mitarbeit von Universität Ulm. Ulm. Online verfügbar unter https://oparu.uni-ulm.de/xmlui/handle/123456789/2544, zuletzt geprüft am 02.11.2018.

Trim, Craig (2013): The Art of Tokenization. Hg. v. IBM. Online verfügbar unter https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en, zuletzt geprüft am 12.11.2018.

Tukey, John W. (1977): Exploratory data analysis (Addison-Wesley series in behavioral science : Quantitative methods).

Underwriters Laboratories Inc. (2018): UL Environment 110 Ed. 2 March 24, 2017. Standard for Sustainability for Mobile Phones. UL. Online verfügbar unter https://standardscatalog.ul.com/standards/en/standard_110, zuletzt aktualisiert am 28.09.2018, zuletzt geprüft am 04.10.2018.

van der Maaten, Laurens (2014): Accelerating t-SNE using Tree-Based Algorithms. In: *Journal of Machine Learning Research* (15), S. 1–21. Online verfügbar unter https://lvdmaaten.github.io/publications/papers/JMLR_2014.pdf, zuletzt geprüft am 05.11.2018.

van der Maaten, Laurens; Hinton, Geoffrey (2008): Visualizing Data using t-SNE. In: *Journal of Machine Learning Research* (9), S. 2579–2605. Online verfügbar unter https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf, zuletzt geprüft am 05.11.2018.

Vembunarayanan, Jana (2013): Tf-Idf and Cosine similarity. https://www.facebook.com/WordPresscom. Online verfügbar unter https://janav.wordpress.com/2013/10/27/tf-idf-and-cosine-similarity/, zuletzt aktualisiert am 05.11.2018, zuletzt geprüft am 06.11.2018.

Venkitachalam, Vijay S.; Namboodiri, Vinod; Joseph, Siny; Dee, Emily; Burdsal, Charles A. (2015): What, Why, and How: Surveying what consumers want in new mobile phones. In: *IEEE Consumer Electron. Mag.* 4 (2), S. 54–59. DOI: 10.1109/MCE.2015.2390651.

W3C (2014): XPath Tutorial. Online verfügbar unter https://www.w3schools.com/xml/xpath_intro.asp, zuletzt aktualisiert am 08.04.2014, zuletzt geprüft am 30.10.2018.

W3C (2018): XML Tutorial. Online verfügbar unter https://www.w3schools.com/xml/default.asp, zuletzt aktualisiert am 30.10.2018, zuletzt geprüft am 30.10.2018.

Wattenberg, Martin; Viégas, Fernanda; Johnson, Ian (2016): How to Use t-SNE Effectively. In: *Distill* 1 (10). DOI: 10.23915/distill.00002.

Weber, Mathias (2014): Big-Data-Technologien - Wissen für Entscheider - Leitfaden. Hg. v. Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e.V. (BITKOM). Online verfügbar unter https://www.bitkom.org/noindex/Publikationen/2014/Leitfaden/Big-Data-Technologien-Wissen-fuer-Entscheider/140228-Big-Data-Technologien-Wissen-fuer-Entscheider.pdf, zuletzt geprüft am 01.11.2018.

Whang, Jay; Maurer, Zach (2018): Python Review. CS224N. Hg. v. Stanford University. Online verfügbar unter https://web.stanford.edu/class/cs224n/lectures/python-review.pdf, zuletzt geprüft am 01.11.2018.

Wilson, Benjamin (2016): Don’t interpret linear hidden units, they don’t exist. Building Babylon. Online verfügbar unter http://building-babylon.net/2016/10/19/dont-interpret-linear-hidden-units-they-dont-exist/, zuletzt geprüft am 05.11.2018.

Wu, Ledell; Fisch, Adam; Chopra, Sumit; Adams, Keith; Bordes, Antoine; Weston, Jason (2017): StarSpace: Embed All The Things! Facebook AI. Online verfügbar unter http://arxiv.org/pdf/1709.03856v5.

Zaki, Mohammed J.; Yu, Jeffrey Xu; Ravindran, B.; Pudi, Vikram (2010): Advances in knowledge discovery and data mining. 14th Pacific-Asia conference, PAKDD 2010, Hyderabad, India, June 21 - 24, 2010 ; proceedings, part I. Berlin: Springer (Lecture notes in computer science Lecture notes in artificial intelligence, 6118). Online verfügbar unter http://dx.doi.org/10.1007/978-3-642-13657-3.