FOM Hochschule für Oekonomie & Management

Hochschulzentrum Düsseldorf

Projektarbeit

im Studiengang Big Data & Business Analytics

im Modul

Analyse semi- & unstrukturierter Daten

über das Thema

Evaluation der Praxistauglichkeit von Feature-based Opinion Mining

von

Philipp Lukasewycz

Erstgutachterin:	Jasmin Schmank
Matrikelnummer:	560194
Abgabedatum:	31.08.2021

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungszeichnis

CRISP-DM..... Cross Industry Standard Process for Data Mining

CSS.....Cascading Style Sheets

HTML.....Hypertext Markup Language

MVP.....Minimum Viable Product

POS.....Part Of Speech

STSS.....Stuttgart-Tübingen-Tagset

VADER.....Valence Aware Dictionary and sEntiment Reasoner

1. Einführung in die Arbeit

1.1 Problemstellung

Verkürzte Produktlebenszyklen, eine steigende Anzahl disruptiver Geschäftsmodelle infolge des technologischen Fortschritts sowie ein stetig wachsender Wettbewerbsmarkt infolge zunehmender Globalisierung – expansive Marktdynamiken fordern von Unternehmen eine kontinuierliche Anpassung an die Bedürfnisse des Marktes. Hierbei stellt mit Blick auf den E-Commerce die Loyalität der Kunden einen Schlüsselfaktor zur Steigerung der Unternehmensprofitabilität dar (Reichheld & Schefter, 2000, S. 106ff.). Um diese zu erreichen sind die Bedürfnisse der Kunden hinsichtlich Produkt- & Serviceangebot umfassend zu erfüllen: Sofern der Kunde eine positive Customer Experience erfährt, ist er mit gegebener Wahrscheinlichkeit eher bereit, die Produkte & Dienstleistungen des Anbieters auch bei zukünftigem Bedarf in Anspruch zu nehmen (Frow & Payne, 2007, S. 98). Kann der Kunde jedoch keine persönlichen Erfahrungen mit dem Anbieter zu Rate ziehen, bedarf er den Erfahrungen seiner Mitmenschen. Vor den Zeiten des Internets geschah dies über Word of mouth (File, Cermak & Price, 1994, S. 311). Heutzutage basieren viele Kaufentscheidungen auf Meinungsäußerungen aus dem World Wide Web, wie etwa Produktrezensionen auf Shopseiten, Social Media oder diversen Internet-Foren (Cui, Lui & Guo, 2012, S. 39). Die öffentliche Transparenz des Kundenfeedbacks kann somit die Kaufentscheidung neuer User positiv wie negativ beeinflussen. Für die Anbieter ist es folglich unabdingbar, die Kundenbedürfnisse bestmöglich zu erfüllen: Neben der Bindung von Bestandskunden hängt auch die Gewinnung von Neukunden folglich explizit von geteilten Erfahrungen der Bestandskunden ab. Aus diesem Grund gilt die sogenannte Customer Centricity, sprich eine konsistente, unternehmensweite Ausrichtung an die Bedürfnisse und Anforderungen der Kunden entlang aller Geschäftsprozesse, bereits seit längerer Zeit als wesentlicher Schlüsselfaktor für nachhaltigen Unternehmenserfolg (Shah, Rust, Parasuraman, Staelin & Day, 2006, S. 121). In diesem Kontext bieten sich angesprochene frei verfügbare Kundenrezensionen für die Anbieter als optimale Datengrundlage zur Ermittlung des Marktbedürfnisses an. Entgegen klassischer Marktforschungsinstrumenten wie Erhebungen und Befragungen kommt nutzenstiftend hinzu, dass Feedbacks in Freitextform eine neutrale, ungeleitete Sicht auf die Meinung des Kunden bieten, da hierbei keine vorgegebenen Skalen und Beurteilungskriterien die Bewertungsinhalte effektuieren, so dass zusätzliche Erkenntnisgewinne erzielt werden

können. Als Folge ist es für Unternehmen von Interesse, diese Daten entsprechend zu extrahieren, analysieren und in das eigene Produkt- und Serviceangebot einfließen zu lassen. Dieser Prozess stellt sich in der praktischen Umsetzung jedoch als problematisch dar: Eine manuelle Analyse der Daten ist aufgrund ihrer Vielzahl und den hieraus resultierenden hohen Aufwänden nicht praxistauglich, so dass automatisierte Analyseprozesse einen effizienten praktischen Einsatz bedingen. Als Herausforderung erweist sich hierbei die automatisierte Extraktion der kundenindividuellen Bewertungskriterien sowie deren subjektiven Beurteilungen aus den entsprechenden Kundentexten. In der Literatur werden in dem Forschungsgebiet des Opinion Mining – wahlweise auch Sentiment Analyse genannt – entsprechende Ansätze zur Lösung dieses Problems diskutiert, die im Rahmen dieser Arbeit auf ihre Praxistauglichkeit evaluiert werden.

1.2 Zielsetzung

Das Ziel dieser Arbeit liegt in der Konzeption und Bewertung eines Minimum Viable Product (MVP)-Opinion Mining-Ansatzes, um mit möglichst geringem Aufwand eine valide automatisierte Auswertung von Kundenmeinungen für die Praxis aufzuzeigen. Nachfolgende Abbildung 1 visualisiert hierzu die vorliegend angewandten Bewertungskriterien einer potenziellen praktischen Implementierung:

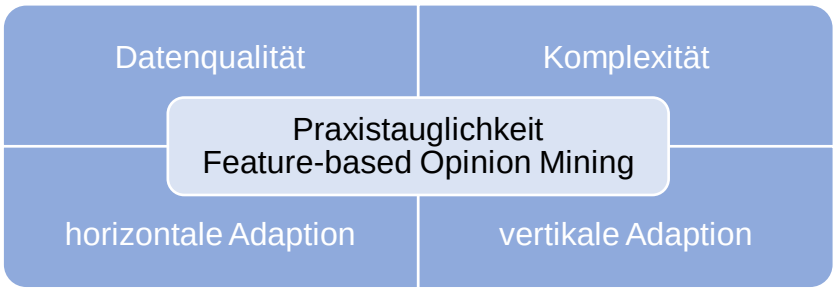

Abbildung 1: Zielkriterien der Untersuchung

Als Methodik des Opinion Mining liegt dieser Arbeit demnach das sogenannte Feature-based Mining zugrunde. Nach der Diskussion des gegenwärtigen Forschungsstands wird dieser Ansatz mittels eines konkreten Fallbeispiels anhand von vier Bewertungskriterien, welche in Teilen unmittelbar mit potenziellen Endanwendern aus der Praxis eruiert wurden, auf einen potenziellen praktischen Einsatz als MVP-Lösung evaluiert. Hierzu zählen neben der Beurteilung des inhaltlichen Erkenntnisgewinns und deren Validität sowie der Komplexität der technischen Umsetzung auch eine Einschätzung über die Generalisierbarkeit der angewandten Methodik, um den MVP-

Ansatz in der Lösungsfindung zu berücksichtigen. Dieses Kriterium umfasst sowohl die vertikale als auch horizontale Adaptionsfähigkeit: Bei der vertikalen Adaptionsfähigkeit liegt die zentrale Anforderung darin, die entwickelte Methode universell für verschiedene Inhalte – bspw. für unterschiedliche Produktgruppen oder -typen – ohne großen zusätzlichen Aufwand nutzbar zu machen. Hinsichtlich der horizontalen Adaption ist zu prüfen, ob heterogene Datenquellen, wie etwa Produktrezensionen aus unterschiedlichen Shops & Foren simultan ohne Mehraufwand im Opinion Mining-Modell berücksichtigt werden können. Als Ergebnis hat diese Arbeit folglich eine konkrete Aussage über die Praxistauglichkeit des Feature-based Mining als MVP-Methode zum Ziel und bietet dementsprechend zugleich Anknüpfungspunkte für eine mögliche praktische Umsetzung, aber auch Ansätze für vertiefte Forschungsarbeiten.

1.3 Aufbau der Arbeit

Zur Untersuchung der zugrundeliegenden Forschungsfrage besteht die Arbeit - wie Abbildung 2 dargestellt - aus vier aufeinander aufbauende Bestandteile.

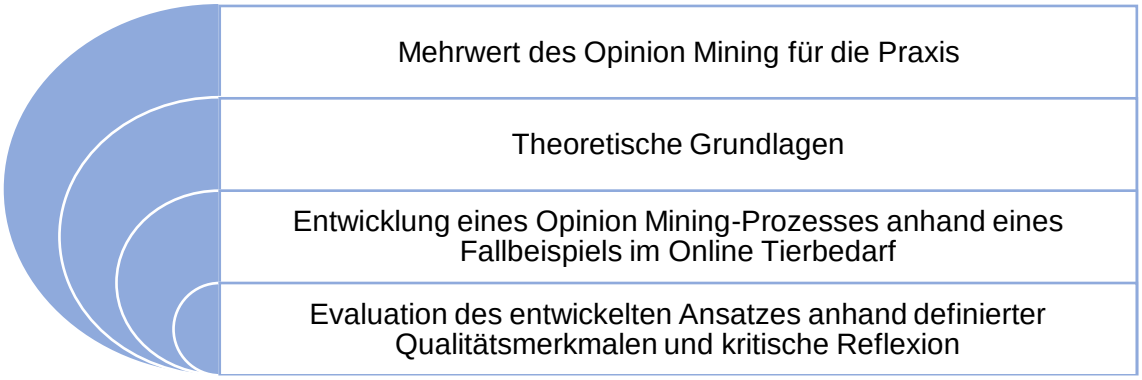

Abbildung 2: Aufbau der Arbeit

Dargelegte Problemstellung und Zielsetzung sowie entsprechende inhaltliche Vertiefung durch die praktische Anwendung im Rahmen einer Fallstudie begründen die Motivation sich mit der Umsetzung von Opinion Mining-Methoden in der Praxis auseinanderzusetzen. Im Grundlagenteil findet einführend zunächst eine kontextuelle Einordnung von Text- und Opinion Mining statt, bevor der gegenwärtige Stand aus Forschung & Literatur hinsichtlich möglicher Umsetzungsmöglichkeiten und -methoden aufbereitet wird. Der ausgewählte Ansatz des Feature-based Mining wird im darauffolgenden Praxisteil anhand eines konkreten Fallbeispiels nach dem Cross Industry Standard Process for Data Mining (CRISP-DM) durchgeführt und im Anschluss kritisch reflektiert, indem Limitationen und Optimierungspotenziale der dargestellten

Methodik skizziert werden. Im Rahmen des Fazits werden die erarbeiteten Erkenntnisse anhand dezidierter Bewertungskriterien evaluiert und abschließend abgewogen, inwieweit der entwickelte Opinion-Mining-Ansatz Potenzial für eine universelle Nutzung in der Praxis aufweist.

1.4 Literaturrecherche

Die vorliegende Arbeit zieht zur Bearbeitung der Untersuchungsfrage unterschiedliche Konzepte und Erkenntnisse aus der wissenschaftlichen Literatur heran. Hierfür wurden im Rahmen der Literaturrecherche ausschließlich wissenschaftliche Datenbanken genutzt, welche im Anhang mit den genutzten Literaturquellen zusammengefasst sind. In dieser Auflistung sind zusätzlich die entsprechenden Suchtaxonomien aufgeführt, welche für eine zielgerichtete Identifikation zweckmäßiger Forschungsarbeiten genutzt wurden. Zur Gewährleistung der wissenschaftlichen Akzeptanz wurde für die abschließende Auswahl der Quellen neben der inhaltlichen Kohärenz ein angemessener Hirsch-Index vorausgesetzt, dessen Ermittlung für Journals gemäß Scimago und für Konferenzen auf Basis von Google Scholar vorgenommen wird.

2. Theoretischer Bezugsrahmen: Opinion Mining als Teilgebiet des Text Mining

2.1 Einordnung in Data Analytics

Opinion Mining ist als eine Teildisziplin von Data Analytics zu klassifizieren. Dieses beschreibt grundsätzlich alle Methoden und Ansätze zur Erkennung und Interpretation von Mustern in Daten, um hierdurch Informationen und somit Schlussfolgerungen für inhaltliche Fragestellungen ableiten zu können. Abbildung 3 gibt einen schematischen Überblick über die wesentlichen Methoden von Data Analytics:

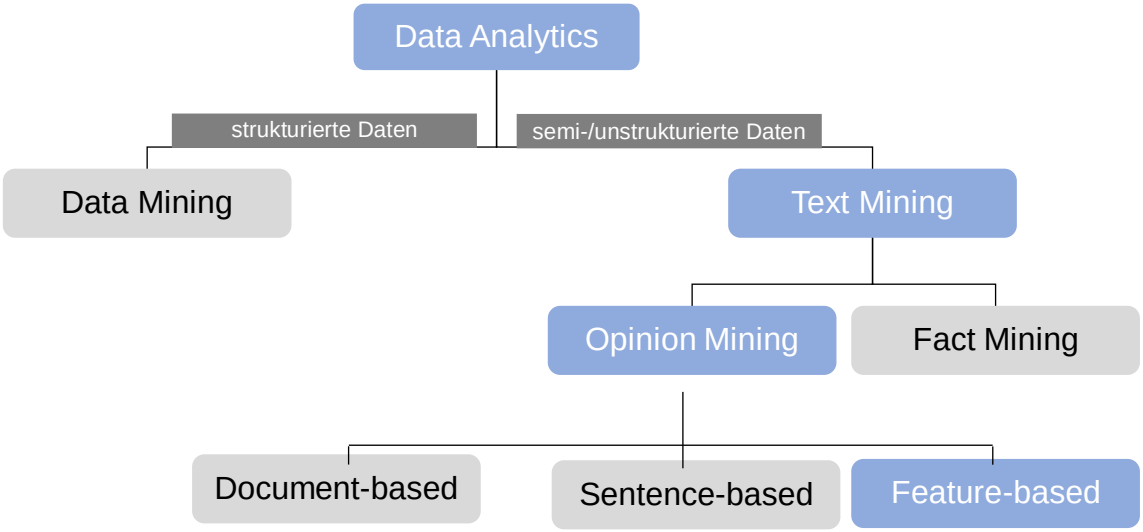

Abbildung 3: Einordnung von Opinion Mining in Data Analytics

Hierbei lassen sich die unterschiedlichen Teilbereiche zunächst nach der Art der zugrundeliegenden Daten trennen. Liegen die zu analysierenden Daten in strukturierter Form vor, sind per definitionem demnach formatiert und lassen sich in einer einheitlich strukturierten Tabelle darstellen, ist der Teilbereich des Data Mining angesprochen. Analyse von semi-/unstrukturierten Daten hingegen bedürfen fortgeschrittene Methoden und zählen zum Teilgebeit des Text Mining, worunter auch die Analyse von Fließtexten zählt (Klass, 2019, S. 268). Prinzipiell lässt sich innerhalb des Text Mining wiederum zwischen Fact Mining und Opinion Mining trennen – das Fact Mining umfasst hierbei Methoden zur Analyse von faktenbasierten Daten, wie etwa einer neutralen Berichterstattung von Nachrichten, bei denen lediglich eine syntaktische Analyse von Texten notwendig ist. Das Opinion Mining hingegen hat die zusätzliche Herausforderung einer semantischen Analyse, um es hierdurch zu ermöglichen, auch die subjektive Note des zu analysierenden Textes valide auszuwerten. Zusammengefasst lässt sich demnach das Opinion Mining als Teildisziplin des Text Mining einordnen mit der Intention, die Meinung und Beurteilung eines Bewertungssubjekts über ein Bewertungsobjekt sowie Teilaspekten dieses Objekts zu untersuchen (Rajeev & Rekha, 2015, S. 2). Gemäß gegenwärtiger Forschungen ist das Opinion Mining wiederum in drei verschiedene Ansätze zu separieren, welche jeweils unterschiedliche Analyseebenen bedienen (Ganeshbhai & Shah, 2015, S. 920). Das Document Level-based Sentiment Classification beschäftigt sich hierbei mit der Analyse auf Dokumentenebene und stellt somit die höchstmögliche Betrachtungsperspektive dar – und ist aus diesem Grund für eine umfassende Berücksichtigung von Kundenbedürfnissen für die zugrundeliegende Problemstellung als zu grob einzuschätzen. Als weiteres Analyselevel wird in der

Literatur das Sentence-based Opinion Mining diskutiert – hierbei erfolgt die Analyse des Sentiments auf Satzebene. Da diese Arbeit jedoch auf die Identifikation und Erfassung der Kundenmeinungen über einzelne (Produkt-)Features abzielt, ist ein möglichst granularer Ansatz auf Wortebene notwendig: Aus diesem Grund wird nachfolgend das Opinion Mining auf Basis des sogenannten Feature-based Mining beurteilt, welches die Sentimentanalyse auf eben dieser kleinstmöglichen Ebene ermöglicht.

2.2 Prozessablauf des Opinion Mining

2.2.1 Prozess des Text Mining

Da das Feature-based Mining als Methodik des Opinion Mining eine Teildisziplin des Text Mining darstellt, sind für eine praktische Umsetzung die notwendigen vorgelagerten Prozessschritte des Text Mining ganzheitlich zu betrachten. Abbildung 4 gibt hierzu einen entsprechenden Überblick (Hippner & Rentzmann, 2006, S. 288):

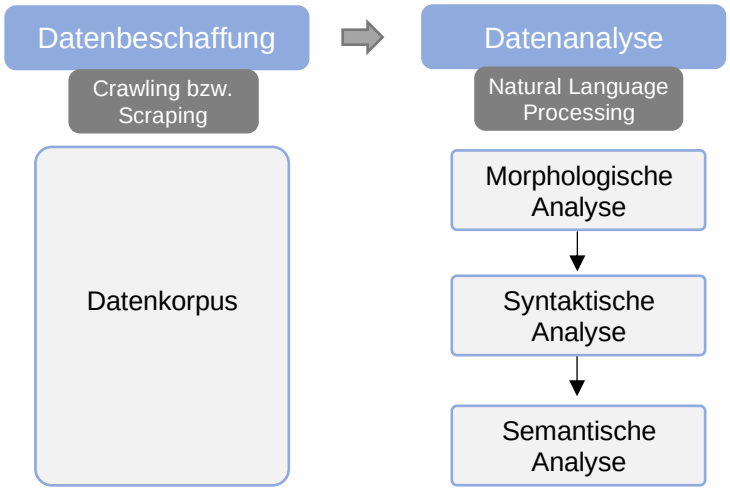

Abbildung 4: Prozess des Text Mining

Zunächst ist die Datenbeschaffung durchzuführen. Sofern hierfür das World Wide Web nach bestimmten Daten zunächst durchsucht werden muss, kommt hierzu das Crawling zur Anwendung – hierbei durchsucht ein Programm, der Crawler, das World Wide Web nach definierten Inhalten, liest die entsprechenden Daten aus und führt in der Folge zusätzlich das Scraping durch, um die Daten den folgenden Analyseprozesse zur Verfügung zu stellen. Wenn sich die Datenbeschaffung auf eine bestimmte Webseite konzentriert, ist der Crawling-Prozess nicht notwendig, so dass die Texte direkt via Scraping von der Webseite ausgelesen werden. Neben der reinen Datenbeschaffung kann der erste Schritte des Text Mining zusätzlich auch eine Datenaufbereitung

umfassen, sofern die ausgelesenen Daten hinsichtlich erforderlicher Formatierungskriterien bearbeitet werden müssen (Singh, Sachdeva, Mahajan, Pande & Sharma, 2014, S. 330f.). Im Ergebnis steht der Datenkorpus, der die Gesamtheit der zu analysierenden Daten in Form von Rohdaten für die folgenden Analyseschritte bereithält. Im abschließenden Analyseprozess des Natural Language Processing werden die Textdaten aus dem Datenkorpus auf Basis linguistischer Techniken für die maschinelle Verarbeitung und Interpretation natürlicher Sprache vorbereitet, bevor im Rahmen der semantischen Analyse eine inhaltliche Analyse der Texte ermöglicht wird. Eine Vertiefung, inklusive der Einordnung des Opinion Mining, folgt.

2.2.2 Natural Language Processing

Das Natural Language Processing fungiert als Datenanalyseprozess und umfasst linguistische Methoden, die gemäß Abbildung 5 in drei unterschiedliche Prozessstufen mit differierenden Verfahrenszielen eingeteilt werden können (Chowdhury, 2005, S. 56).

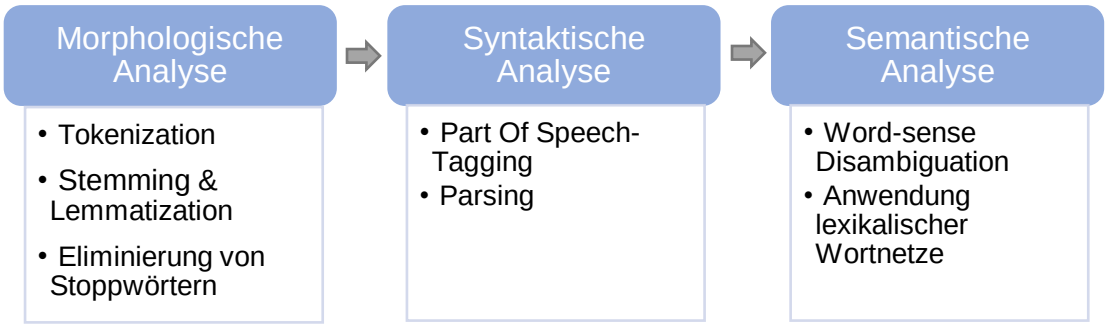

Abbildung 5: Prozess des Natural Language Processing

Beginnend mit den morphologischen Analysen, steht zunächst die Komplexitätsreduktion der zugrundeliegenden Texte im Fokus, um eine valide maschinelle Datenverarbeitung in den folgenden Stufen zu ermöglichen. Eine klassische Methode ist hierbei die Tokenisierung: Dabei werden die Texte in separate Teilstücke, sogenannte Tokens, aufgespaltet und Satzzeichen entfernt. Mittels Stemming und Lemmatization werden im Anschluss die Wörter in ihre Wortstämme respektive Grundform überführt. Zusätzlich werden Stoppwörter, die keine inhaltliche Aussagekraft und somit für den Analyseprozess nicht von Bedeutung sind, eliminiert. Im Rahmen der folgenden syntaktischen Analyse steht die Strukturierung der Daten im Vordergrund, um die spätere semantische Analyse zu unterstützen. Zentrale Methoden der syntaktischen Analyse stellen das Part Of Speech (POS)-Tagging sowie Parsing dar. Hierbei werden den Tokens ihre grammatikalischen Wortarten sowie Funkionen zugeordnet. Über die

abschließende semantische Analyse wird das Textverständnis im Modell eingebettet. Durch die Anwendung lexikalischer Wortnetze wird der Maschine Kenntnis über die Wortbedeutungen verliehen, so dass eine inhaltliche Interpretation hinsichtlich des Sentiments ermöglicht wird. Mittels einer word-sense Disambiguation ist es zusätzlich möglich, kontextuelles Verständnis in die Verarbeitung zu integrieren.

Da vorliegende das Opinion Mining im Rahmen des Text Mining behandelt wird, erfolgt im weiteren Verlauf die Einbettung des Feature-based Opinion Mining-Ansatzes als spezifizierte Datenanalysemethode in den vorgestellten generalistischen Natural Language Prozess.

2.2.3 Features-based Opinion Mining

Aus den vorgestellten Opinion-Mining Ansätzen in 2.2.1 baut diese Arbeit auf das sogenannte Features-based Mining auf. Dieses lässt sich gemäß Abbildung 6 in drei wesentliche Prozessschritte nach Popescu & Etzioni (2005, S. 339) untergliedern:

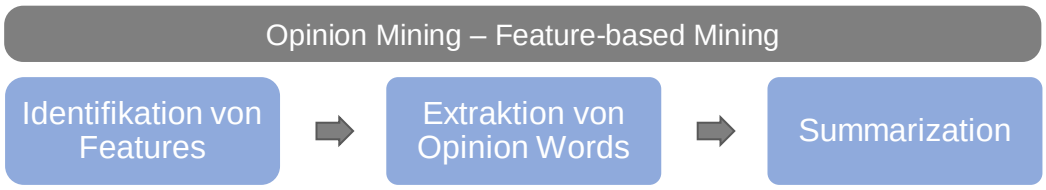

Abbildung 6: Prozess des Feature-based Opinion Mining

Zunächst ist eine Identifikation der Features erforderlich. Diese umfassen jene Merkmale, über welche in den Texten eine Meinung geäußert wird – im Rahmen einer Produktrezension sind dies beispielsweise die Kriterien, welche die Rezensenten bei ihrer Meinungsäußerung bewerten. Die Identifikation kann hierbei im einfachsten Fall über die absoluten Häufigkeiten der benutzten Nomen geschehen, da Features gemeinhin in Form von Substantiven auftreten (Zhang & Narayanan, 2005, S. 3f.).

Der folgende Prozessschritt beinhaltet die Identifikation und Extraktion der Opinion Words, also jenen Wörtern, welche die Meinung über einen entsprechenden Sachverhalt – in diesem Kontext Feature – abbilden. Weil hierfür zumeist Adjektive genutzt werden, liegt der Fokus bei der Extraktion gemeinhin auf Adjektiven, welche in Verbindung mit Features auftreten (Hu & Liu, 2004, S. 172). Hieran setzt eine semantische Analyse durch die Anwendung lexikalischer Wortnetze, um ein Verständnis über das Sentiment im Datenmodell zu implementieren, indem die Opinion Words unterschiedlichen Clustern, wie etwa „positive Äußerung“ und „negative Äußerung“, zugeordnet werden. In

der abschließenden Summarization werden die Features und Opinion Words-Paare statistisch insofern aufbereitet, dass transparent und valide konkrete Schlüsse aus den untersuchten Meinungen gezogen werden können.

Mit der Identifikation entsprechender Bewertungskriterien in Form von Produktfeatures und der gleichzeitigen Beurteilung dieser durch die Analyse der dazugehörigen Opinion Words eignet sich der Feature-based Mining-Ansatz sehr gut für die zugrundeliegende Problemstellung: Aus einer Vielzahl von Kundenrezensionen die individuellen Kundenanforderungen und -kriterien, nachfolgend als Features bezeichnet, aus den einzelnen Texten zu identifizieren und die dazugehörige subjektive Einschätzung zu analysieren entspricht exakt den Anforderungen, welche dieser Arbeit zugrundeliegen. Da das Feature-based Mining als Opinion Mining-Ansatz die semantische Analyse im allgemeinen Text-Mining-Prozess für diese Arbeit spezifiziert, lassen sich somit folgende Modifikationen am Ursprungsmodell gemäß Abbildung 7 festhalten:

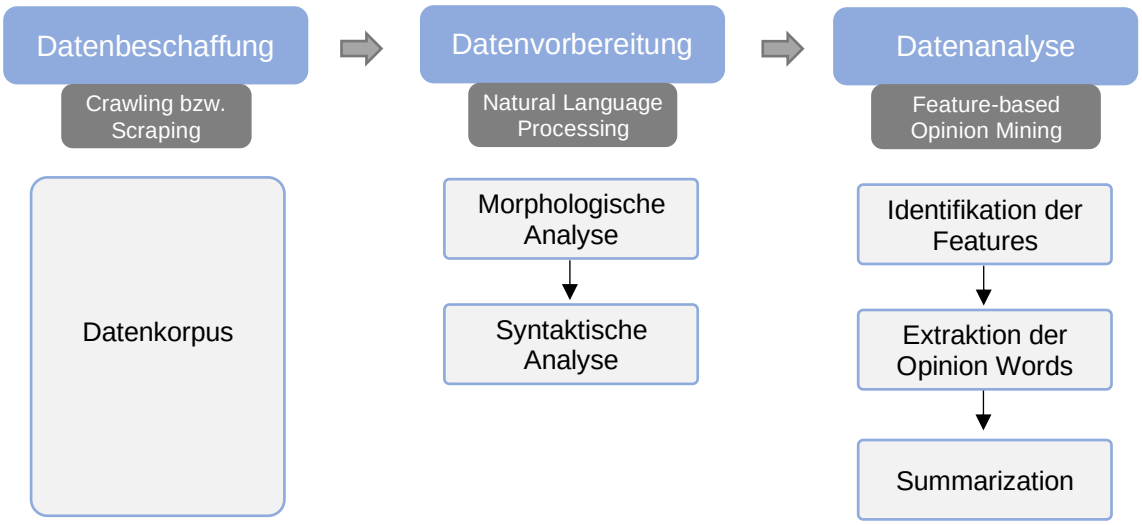

Abbildung 7: Angewandter Opinion Mining-Prozess

Dieses Modell stellt somit den MVP-Opinion Mining Ansatz dar, der im folgenden Praxisteil anhand eines Fallbeispiels umgesetzt und darauf aufbauend hinreichend auf Praxistauglichkeit diskutiert wird.

3. Analyse von Produktrezensionen mittels Opinion Mining

3.1 Einführung in das Fallbeispiel

Im Folgenden wird der vorgestellte MVP-Opinion Mining Ansatz anhand eines praktischen Anwendungsbeispiels aus dem Online-Tierbedarfshandel evaluiert. Hierzu fand ein Austausch mit dem Category Management von zooroyal.de, dem drittgrößten nationalen Onlinehändler im Tierbedarf, statt, um die Fallstudie praxisnah zu gestalten und die kritische Reflexion der gewählten Methodik eng an die Kriterien einer potenziellen Adaption in der Praxis zu knüpfen. Die Ausgestaltung und Diskussion des Fallbeispiels richtet sich nach dem CRISP-DM-Modell, welches ein anerkanntes Standardprozessmodell für die Durchführung von Data Mining-Projekten darstellt. Abbildung 8 fasst hierzu die inhaltlichen Bearbeitungsstufen der Fallstudie im Rahmen der einzelnen Phasen des CRISP-Modells zusammen und dient somit gleichzeitig als Ablaufplan für folgenden Praxisteil.

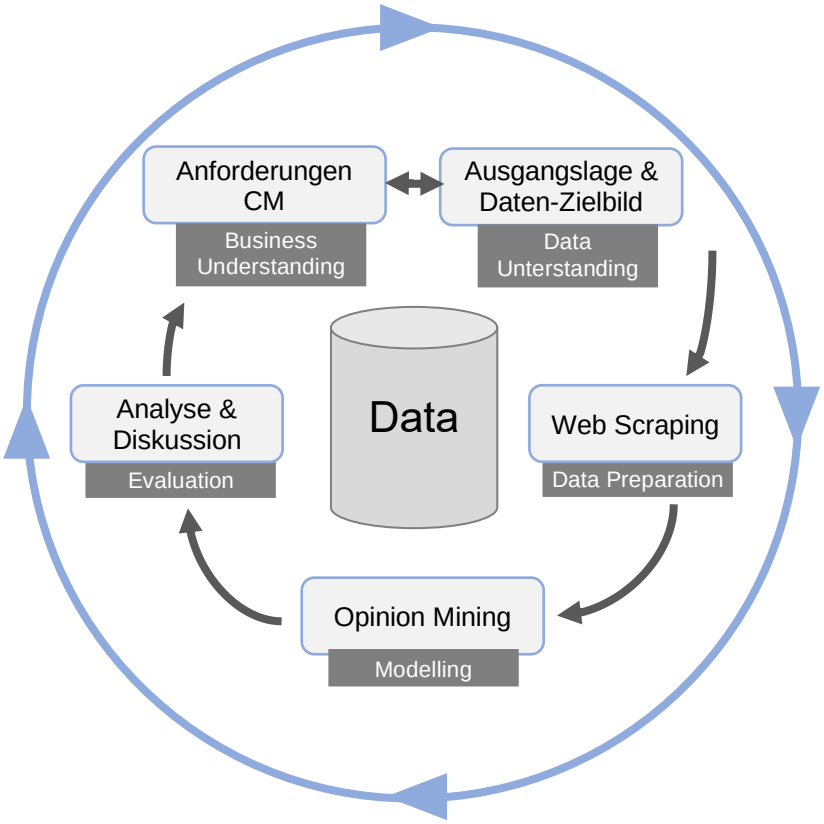

Abbildung 8: Studienverlauf anhand des CRISP-DM-Modells

Demnach wird im Rahmen des Business Understanding zunächst der inhaltliche Austausch mit dem Category Management von zooroyal.de thematisiert, auf dessen Basis die Konzeption der Fallstudie fußt. Hierdurch werden diskutierte Anforderungen und Bedürfnisse potenzieller Anwender von Opinion Mining identifiziert und adäquat in der Studie berücksichtigt. Hieran anknüpfend erfolgt eine Auseinandersetzung mit der zugrundeliegenden Datenlandschaft. Ziel hierbei ist es, sowohl über die zu analysierenden Daten, welche in einem Datenkorpus nachfolgend aggregiert werden müssen, als auch über das nachgefragte Zielbild unter Berücksichtigung des Business Kontexts ein ausgeprägtes Verständnis zu entwickeln. Im Zuge der Data Preparation wird die Datenaufbereitung analog des in 2.2.3 vorgestellten Opinion Mining-Prozesses via Scraping technisch umgesetzt, um im folgenden Modelling den diskutierten Feature-based Mining-Ansatz anzuwenden.

In der abschließenden Evaluation erfolgt die Analyse & Diskussion des praktizierten MVP-Ansatzes hinsichtlich der Praxistauglichkeit unter Berücksichtigung definierter Bewertungskriterien aus dem Austausch mit dem Category Management von ZooRoyal.

3.2 Business Understanding: Praxisrelevanz des Studienziels

Um eine praxisnahe und somit valide Aussage über die Praxistauglichkeit des entwickelten Opinion Mining-Ansatzes sicherzustellen, basiert die inhaltliche Konzeption der Fallstudie auf einen Austausch mit einem wesentlichen Konsumenten von Markt- und Kundenfeedbacks: dem Category Management. Abbildung 9 veranschaulicht hierzu die Funktion und den Gegenstand des Category Managements:

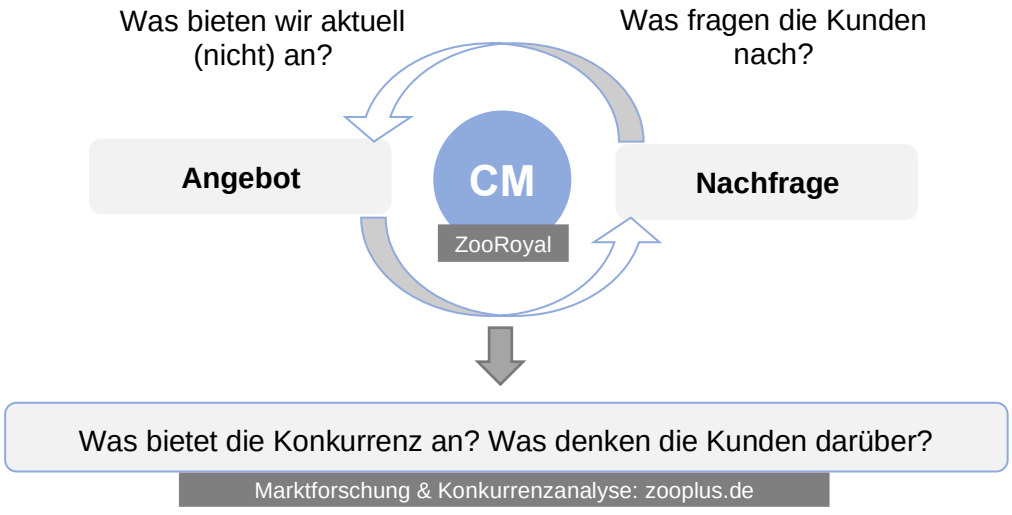

Abbildung 9: Funktionen das Category Managements

Die Kernaufgabe liegt hierbei in der optimalen Ausbalancierung von Leistungsangebot und Marktnachfrage. Als verantwortliche Stelle für das angebotene Produktsortiment ist es die zentrale Herausforderung für das Category Management, dieses mit den Bedürfnissen und Anforderungen des Marktes zu vereinen. Zur Erreichung eines kundenzentrierten Angebots ist demnach das Wissen über die Kundenbedürfnisse unabdingbar – aus diesem Grund berücksichtigt diese Arbeit die Fragestellungen und praktischen Erfahrungen aus Category Managementsicht. Als Resultat der Diskussion ist festzuhalten, dass insbesondere die Kundenrezensionen außerhalb des eigenen Online-Shops von Interesse sind, da zum einen die Bewertungen auf der eigenen Shopseite bereits kontinuerlich analyisiert werden, zum anderen durch die Analyse von konkurrierenden Shops Synergieeffekte realisiert werden können. Aufgrund des variierenden Produktsortiments werden nicht nur Insights über bekannte Produkte, Marken oder Kategorien erzielt, sondern auch über die Kundenwahrnehmung von Sortimenten, welche nicht im eigenen Leistungsportfolio angeboten werden. Sofern das Kundenfeedback hierbei positiv ausfällt, lassen sich hiermit Lücken im eigenen Angebot identifizieren und beheben. Neben des Feedbacks hinsichtlich konkreter Produkte, Marken oder Kategorien sehen die CM-Analysten insbesondere in der Identifikation von weiteren Kundenwünschen, auf welche in den Texten Bezug genommen und die gegenwärtig nicht erfüllt werden, großes Potential – und sehen das Opinion Mining hierdurch potentiell als automatisiertes, kundenzentriertes Marktforschungsinstrument.

Zur Untersuchung dieser Potentiale werden für diese Fallstudie nachfolgend Produktrezenensionen des führenden nationalen Online-Tierbedarfshändlers zooplus.de genutzt. Als notwendige Voraussetzungen für einen praktischen Einsatz von Opinion Mining sprachen sich die CM-Analysten neben der Erreichung einer hohen Ergebnisqualität für einen hohen Automatisierungsgrad bei möglicher Adaption an unterschiedliche Untersuchungsgegenstände aus. Diese Punkte wurden bereits im Rahmen der Zielsetzung dieser Arbeit berücksichtigt und werden nach Umsetzung der Fallstudie im Analyseteil und Fazit hinreichend diskutiert.

3.3 Data Understanding: Ausgangslage und Zielbild

Nach der Ausarbeitung der wirtschaftlichen Motivation und inhaltlichen Zielsetzung des zugrundeliegenden Fallbeispiels erfolgt nun eine tiefergehende Auseinandersetzung mit der entsprechenden Datenlandschaft. Die zu analysierenden Produktrezensionen liegen im Webshop zooplus.de separat auf den jeweiligen Produktseiten gemäß folgendem Ausschnitts ab:

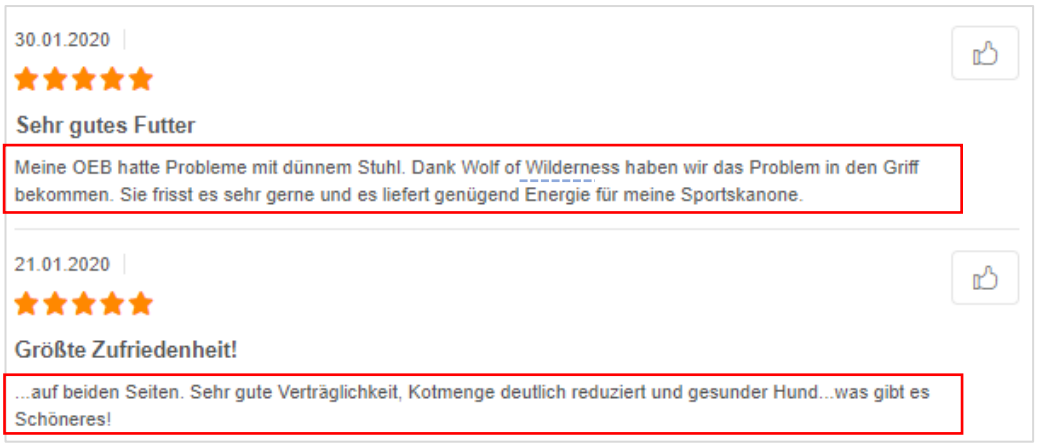

Abbildung 10: Produktrezensionen auf zooplus.de

Entsprechend des definierten Opinion Mining Prozess in 2.2.3 ist zunächst die Erstellung eines Datenkorpus notwendig, welcher alle notwendigen Texte zur späteren Analyse beinhaltet. Primär relevant sind die in Abbildung 10 rot markierten Freitexte, die nachfolgend im Rahmen der Data Preparation via Scraping aus dem World Wide Web zu extrahieren sind. Zusätzlich ist es notwendig dazugehörige Artikelstammdaten zu extrahieren, um die Analyseergebnisse entsprechenden Objekten – wie etwa Artikel, Kategorien oder Marken zuzuordnen. Abbildung 11 skizziert hierzu das Zielbild nach praktischer Umsetzung des Opinion Mining:

Kategorie: x									
Artikel	Feature 1			Feature 2			Feature 3		
	pos.	neg.	∑	pos.	neg.	∑	pos.	neg.	∑
a	w%	x%	y	w%	x%	y	w%	x%	y
b	w%	x%	y	w%	x%	y	w%	x%	y
c	w%	x%	y	w%	x%	y	w%	x%	y

Abbildung 11: Skizziertes Zielbild nach Feature-based Mining

Demnach liegen die Ergebnisse nach einem ausgewählten Attribut – beispielsweise Kategorie oder Marke – auf Artikelebene vor und umfassen in den Spalten attributindividuelle Features, welche die identifizierten Beurteilungskriterien der Kunden darstellen. In den Zeilen erfolgt die entsprechende Aggregation der Kundenbewertungen. Abbildung 12 zeigt hierzu ein mögliches Beispiel, welches sich aus den Kundenrezensionen aus Abbildung 11 ergeben könnte.

Kategorie: Hunde Trockenfutter									
Artikel	Energie			Verträglichkeit			Kotmenge		
	pos.	neg.	∑	pos.	neg.	∑	pos.	neg.	∑
12 kg Wolf of Wilderness	65%	35%	253	33%	67%	159	38%	62%	135

Abbildung 12: Ergebnisentwurf nach Feature-based Mining

Hierbei werden auf Kategorieebene Energie, Verträglichkeit und Kotmenge als Produktfeatures identifiziert, da diese Eigenschaften in den dargestellten Rezensionen explizit genannt und bewertet werden. Die Anzahl positiver wie negativer Bewertungen über die einzelnen Features werden über alle Rezensionen eines Artikels hinweg erfasst und ermöglichen eine Aussage darüber, ob das Leistungskriterium entsprechend erfüllt wird oder nicht. Zusätzlich lässt die Gesamtanzahl der Bewertungen je Feature eine Aussage über die Wichtigkeit des Features zu. Hierdurch lassen sich exakt jene Fragestellungen beantworten, die von Seiten des Category Managements als relevant eingestuft werden, so dass diese Ergebnisskizze als grobes Zielbild festzuhalten ist.

Nachdem hiermit die Motivation und das inhaltliche Verständnis für das Fallbeispiel geschaffen wurden, erfolgt mit der Data Preparation der erste technische Schritt der Umsetzung des gewählten Opinion Mining-Ansatzes.

3.4 Data Preparation: Erstellung des Datenkorpus

3.4.1 Datenbeschaffung und -aufbereitung

Die Prozessstufe der Data Preparation sieht die Extraktion und Aufbereitung der Rohdaten aus den Datenquellen vor. Für die vorliegende Studie sind demnach die Kundenrezensionen in Form von Freitexten sowie zusätzliche Artikelstammdaten aus dem zooplus-Webshop in jener Form aufzubereiten, dass die folgenden Analysetools die Daten einlesen und verarbeiten können. Abbildung 13 visualisiert hierzu die Umsetzung der Data Preparation:

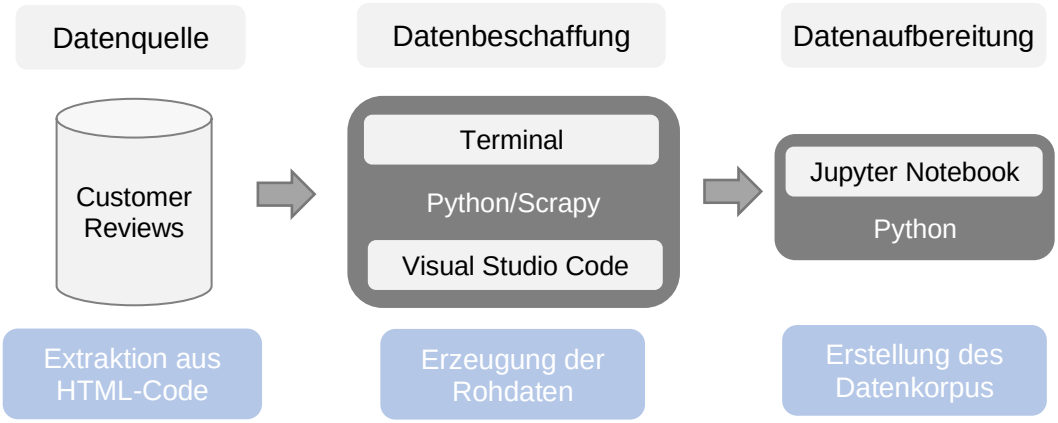

Abbildung 13: Prozess der Data Preparation

Im vorliegenden Fallbeispiel liegen die zu analysierenden Daten über den Zooplus Webshop in Form von Hypertext Markup Language (HTML) Dokumenten vor. Zur Extraktion der benötigten Daten bieten sich diverse Softwarelösungen an, wobei an dieser Stelle das Open Source Framework Scrapy genutzt wird. Dieses ist sowohl für Webscraping als auch Webcrawling anwendbar und somit universell für unterschiedliche Datenquellen geeignet. Über eine angebotene Shell wird dem Anwender die Möglichkeit gegeben, entwickelten Programmcode ohne Ausführung des kompletten Crawling-Prozesses zu testen, was die Codeerstellung erheblich beschleunigt und vereinfacht. Über Scrapy ist es mittels Cascading Style Sheets (CSS)-Ausdrücken möglich, Daten aus HTML-Dokumenten an dezidierten Stellen zu extrahieren. Abbildung 14 zeigt hierzu einen Ausschnitt des HTML-Codes einer Bewertungsübersicht aus dem Webshop von zooplus:

```
<h4 color="greyscale" class="H4-sc-3wlrl6-0 Review__ReviewTitle-sc-1pes9i0-1 furnkc">
Positiev</h4>
▼<p color="greyscale" class="Paragraph-up6p6k-0 icOrsx"> == $0
  "Sehr schnelle Lieferung, schmeckt unseren Hund sehr gut. Kann man auch sehr gut als
  Leckerchen benutzen.richtige Größe."
</p>
```

Abbildung 14: HTML-Code einer Produktbewertungsseite von zooplus.de

Zusätzlich zeigt Abbildung 15 den dazugehörigen Programmcode bei Nutzung von Scrapy zur Extraktion der dazugehörigen Rezension. Hierbei ist ersichtlich, dass lediglich die definierte Klasse im HTML-Dokument angesprochen werden muss, um die gewünschten Daten auszulesen.

```
'rez': str(rezensionen.css('p.Paragraph-up6p6k-0::text').extract()
```

Abbildung 15: Auszug des Programmcodes in Scrapy zur Datenextraktion

Nach dieser Methodik ist es in der Folge möglich, sämtliche weitere Daten, wie etwa Produkt- und Kategorietitel, aus dem Webshop auszulesen. Zwecks Komplexitätsreduktion werden vorliegend jedoch nur die Rezensionstexte eines Artikels behandelt. Im letzten optionalen Schritt der Data Preparation werden die extrahierten Rohdaten in Python zwecks verbesserter Formatierung und Strukturierung optimiert. Abbildung 16 zeigt hierzu einen Ausschnitt des erhaltenen Datenkorpus als Endprodukt:

RezAllnew.txt

```
Sehr schnelle Lieferung  schmeckt unseren Hund sehr gut. Kann man auch sehr gut als
Leckerchen benutzen.richtige Größe. Mein RR Rüde liebt dieses Futter und
verschlingt es wie ein riesiges Leckerli 🤣. Bei anderem Trockenfutter (einmal
Royal Canin  sonst schon auch Bosch) hat er immer gemeckert. Hier verschlingt er
sogar die doppelte Portion (bekommt er nicht jeden Tag. War ein Test und er soll
nebenbei auch noch ein bisschen zunehmen 😊). Das Futter hatte so gute Bewertungen
dass ich dachte für meinen Bodeguero  der normalerweise alles verträgt  absolut
super Preisleistungsverh.! Allerdings hat von diesem Futter sofort Durchfall u
Blähungen bekommen. (Trotz schonendem Übergang mit dem alten Futter!) Selbst nach 3
Wochen ist keine Besserung eingetreten. Schade. Wir werden wieder umstellen. Wir
haben einen sensiblen Malinois und waren länger auf der Suche nach Gründen für
Durchfall  mattes Fell und leichten Haarausfall. Dementsprechend mussten wir
```

Abbildung 16: Auszug des Datenkorpus mit Kundentexten

Wie dargestellt werden die Rezensionstexte in einer Text-Datei exklusive einer Zuordnung zu den jeweiligen Rezensenten gebündelt. Dies hat den Hintergrund, dass der dieser Arbeit zugrundeliegende Feature-based Mining-Ansatz keine Informationen auf Satz- oder Dokumentebene – welche in diesem Kontext der Kundenrezensionsebene entspricht – benötigt, sondern die Analyse auf der kleinstmöglicten Feature-Ebene stattfindet. Aus diesem Grund stellt die Bündelung der einzelnen Rezensionstexten den Dateninput für den nun umzusetzenden Opinion-Mining Prozess dar. Abbildung 17 zeigt hierzu die Rezensionstexte eines ausgewählten Artikels. Eine Bündelung aller Rezensionstexte mehrerer Artikel, beispielsweise auf Marken oder Kategorieebene ist über die manuelle Hinterlegung entsprechender Links zu den jeweiligen Produktbewertungsseiten im Programmcode mit einem moderaten manuellen Aufwand möglich, inhaltlich im Rahmen dieser Ausarbeitung jedoch nicht zielführend, da der CM-Fokus zur Identifikation von potentiellen Top-Performern eine Auswertung auf Artikelebene bedingt. Durch eine Aggregation der Artikeldaten ist eine konsolidierte Sicht auf Marken- oder Kategorien in der Folge ebenso möglich. Als

Musterartikel wird für die folgende Ausarbeitung ein Hunde-Trockenfutter verwendet.[1] Zum 12.08.2021 wurden 173 Produktbewertungen für diesen Artikel abgegeben, von denen die 40 neuesten im Datenkorpus abgebildet sind. Auf Rezensionen von 2018 oder älter wird hierbei verzichtet. Des Weiteren sei angemerkt, dass im Zooplus-Webshop jeweils 40 Produktbewertungen auf einer Seite dargestellt werden. Sofern mehr Rezensionen im Datenkorpus berücksichtigen werden sollen, müsste der Programmcode zur automatisierten Identifikation und Auslese der weiteren Bewertungsseiten um eine Pagination-Funktion erweitert werden.

Da im Zuge des Webscraping vorgestellte automatisierte Auslesung und Nutzung fremder Inhalte ohne unmittelbare Zustimmung des jeweiligen Seitenbetreibers erfolgt, schließt sich zunächst eine Einschätzung der Rechtslage hinsichtlich von Crawling- respektive Scrapingvorgängen an.

3.4.2 Exkurs: Rechtslage zum Scraping

Fragwürdig erscheint hierbei, ob die Auslese und Nutzung fremder Daten ohne Kenntnis oder Genehmigung der jeweiligen Seitenbetreiber juristisch haltbar ist. Ein Blick in die Rechtsprechung offenbart hierzu, dass das Oberlandesgericht (OLG) Köln einer Klage stattgab, bei der die angeklagte Partei Artikeltextbeschreibungen mittels Scraping ausgelesen, kopiert und für den eigenen Webshop punktuell übernommen hat. Hierzu urteilte das OLG (2020), dass durch die Nutzung eines Web-Scraper-Programms und die hierauf aufbauende Verwendung der Daten im eigenen Webshop eine „wiederholte und systematische Vervielfältigung, Verbreitung und öffentliche Wiedergabe von … Teilen der klägerischen Datenbank zuwiderlaufen und die berechtigten Interessen der Klägerin unzumutbar beeinträchtigen“ und somit ein Verstoß gegen das Urheberrecht vorliegt. Herrschende Meinung in der Rechtspraxis ist gegenwärtig nach Dury (2020) hierzu jedoch, dass „… Webscraping grundsätzlich zulässig ist, wenn von den eingesetzten Bots (Crawler, Scraper) keine technischen Schutzmaßnahmen überwundern werden und man auch keine eigene “Schattendatenbank“ mit den Daten bestückt“ werden. Übertragen auf das Vorgehen in dieser Arbeit resultieren hieraus zunächst keine Rechtsunsicherheiten, da zooplus.de keine technischen Schutzmaßnahmen zur Verhinderung von Crawling/Scrapingaktivitäten implementiert hat. Des Weiteren werden die Analyseergebnisse aus dem Opinion Mining nicht

[1] Vgl. https://www.zooplus.de/feedback/shop/hunde/hundefutter_trockenfutter/bosch/bosch_adult/567917 (zuletzt abgerufen am 28.08.2021)

kommerziell genutzt oder weitergegeben, sondern lediglich zur Beurteilung des Modells verwendet.

Allerdings ist zu konstatieren, dass bei der praktischen Anwendung und Nutzung von Opinion Mining die Rechtslage umfassend zu prüfen ist, sofern auf Daten externer Seiten zurückgegriffen wird. Weiterführende praktische Implikationen werden im Rahmen der Reflexion thematisiert.

3.5 Modelling: Umsetzung des Opinion Mining

Die praktische Umsetzung des Feature-based Mining erfolgt mittels der Analytics-Plattform RapidMiner. Die Ausarbeitung basiert hierbei auf der kostenfreien OpenSource-Version RapidMiner Studio, was wiederum kongruent mit dem zugrundeliegenden MVP-Ansatz dieser Arbeit ist. Ein wesentlicher Vorteil von RapidMiner liegt in der langjährigen Historie des Tools, so dass zu vielen Problemstellungen diverse Hilfestellungen und Lösungen aus einer breiten Community vorliegen. Generell zeichnet sich RapidMiner durch eine benutzerfreundliche graphische Benutzeroberfläche aus, welche dem Endanwender ein intuitives Modelling von Analytics mittels Prozessflussdiagrammen ermöglicht. Zentraler Bestandteil sind hierbei vordefinierte Operatoren, welche die unterschiedlichen analytischen Funktionen und Methoden abbilden und mit denen die Anwender ihre jeweiligen Modelle inhaltlich spezifizieren. Abbildung 17 zeigt hierzu einen Teilausschnitt der Modellierung für die vorliegende Fallstudie:

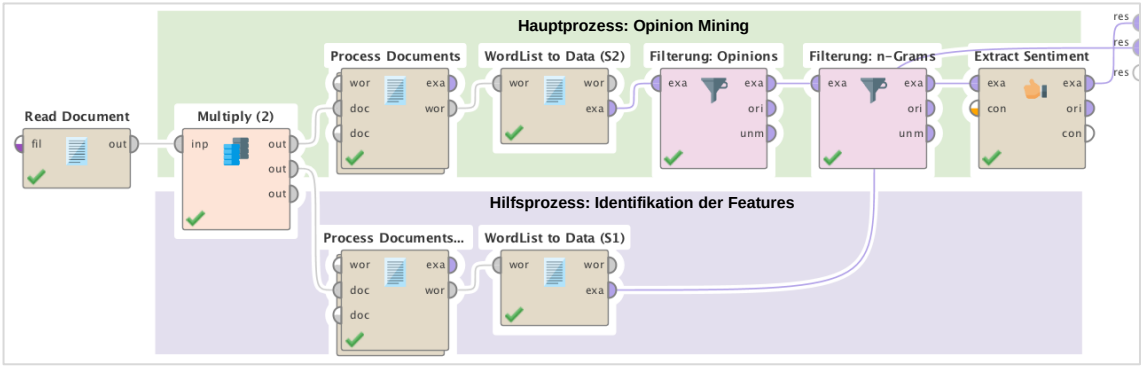

Abbildung 17: Modellierung des Opinion Mining in RapidMiner

Über den ersten Operator *Read Document* werden die zu analysierenden Daten in das Prozessmodell geladen. Vorliegend sind dies die gebündelten Textrezensionen auf Artikelebene aus der Data Preparation, so dass hiermit – bezugnehmend auf das definierte Opinion Mining Modell aus 2.2.3 – der Prozess der Datenbeschaffung abgebildet ist.

In der Folge ist ersichtlich, dass der Datenkorpus als Datengrundlage über den *Multiply*-Operator für zwei Prozesse genutzt wird, die inhaltlich sequenziell zu betrachten sind. Zunächst wird über den unteren Hilfsprozess die erste Stufe des Analyseprozesses im Opinion Mining vollzogen, indem aus den Rezensionstexten die Identifikation der Features erfolgt. Hierfür werden im Operator *Process Documents* folgende zwei Techniken gemäß Abbildung 18 angewendet:

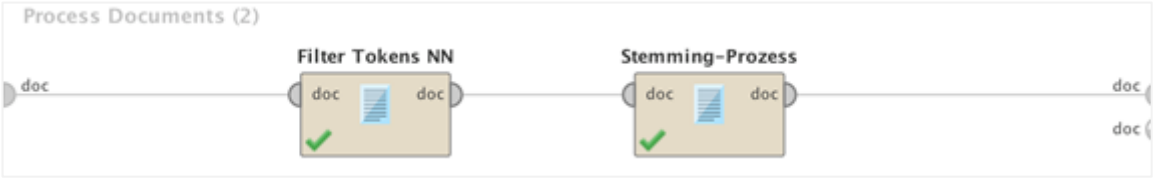

Abbildung 18: Umsetzung der Feature Identifikation in RapidMiner

Über den Operator *Filter Tokens NN* ermöglicht RapidMiner die Filterung von Tokens anhand definierter POS-Tags – für die deutsche Sprache wird hierfür das Stuttgart-Tübingen-Tagset (STSS) genutzt. Abbildung 19 zeigt hierzu die Konfiguration:

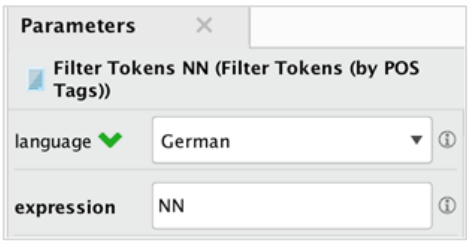

Abbildung 19: Part-of-Speech Tagging zur Feature-Identifikation

Durch die Auswahl des NN-Tags werden folglich nur Tokens identifiziert, welche gemäß STSS als „normale Normen“ klassifiziert sind. Für die vorliegende Untersuchung ist dies zweckmäßig, da zur Identifikation der Bewertungskriterien der Rezensenten einzig die Substantive methodisch von Interesse sind. Zusätzlich findet im zweiten Schritt das *Stemming* statt, um die jeweiligen Tokens zu clustern und somit unterschiedliche Deklinationen bei der Ermittlung der absoluten Häufigkeiten zu berücksichtigen. Grundsätzlich ist festzuhalten, dass die sequenziellen Phasen des zugrundeliegenden Opinion Mining-Prozessmodells in der Praxis nicht in der strikten Chronologie umzusetzen sind, sondern unter der Berücksichtigung des inhaltlichen Kontexts einer individuellen Ausgestaltung bedürfen. Somit erfolgt im vorliegenden Beispiel mit dem Stemming ein Verfahren des Natural Language Processing nach der Identifikation der Features und damit gemäß des originären Modells nach der Datenanalyse. Ebenso finden syntaktische und morphologische Analyseschritte durch die gleichzeitige Identifikation und Filterung via POS-Tags im Filter Token-Operator simultan statt. Aus

diesem Grund bleibt als erstes Zwischenfazit zu konstatieren, dass die praktische Umsetzung sich zwar nach den inhaltlichen Methoden des Ursprungsmodells richtet, jedoch sachlich-spezifische Anforderungen entsprechende individuelle Handhabung bedürfen, so dass der idealtypische Prozessablauf in der Praxis nicht unmittelbar umzusetzen ist. Als Resultat des dargestellten Hilfsprozesses ergeben sich folgende Häufigkeiten für identifizierte Tokens gemäß Abbildung 20:

Row No.	word	in docume...	total ↓
30	futt	1	33
49	hund	1	21
23	durchfall	1	13
28	fell	1	8
16	bosch	1	7
85	probl	1	7
13	blahung	1	6
52	jahr	1	5
69	mag	1	5
111	tag	1	5
131	woch	1	5
86	produk	1	4
107	stuhlgang	1	4

Abbildung 20: Häufigkeitstabelle über die identifizierten Tokens aus RapdiMiner

Aus Betrachtung der Ergebnisse ist ersichtlich, dass neben der Definition eines Schwellwertes zur Priorisierung der Tokens – vorliegend werden mindestens 4 Nennungen zugrunde gelegt – zunächst eine inhaltliche Auslese erforderlich ist, um die folgenden Analysen auf einer zweckmäßigen und effizienten Datengrundlage zu fußen. Hierfür sind rot markierte Tokens als inhaltlich bedeutungsvoll einzuschätzen, da diese wesentliche Produktmerkmale, -anforderungen oder -eigenschaften ausdrücken, wohingegen generische Begriffe wie Futter (‚fut') oder der Markenname (‚bosch') von keiner analytischen Relevanz sind. Aus diesem Grund werden rot markierte Tokens im Hauptprozess in Filterfunktionen überführt, um die Untersuchung somit auf diese Tokens einzuschränken. Eine automatisierte, KI-gestützte inhaltliche Auslese ist in diesem Zusammenhang nicht möglich, so dass der manuelle Aufwand im Rahmen der Token-Identifikation im vorliegenden Setting als notwendig und wiederkehrend einzustufen ist. Auf eine weiterführende Clusterung der identifizierten Tokens wird zunächst verzichtet.

Der zentrale Opinion-Mining Prozess findet nun Hauptprozess statt, der sich gemäß Abbildung 21 vollzieht:

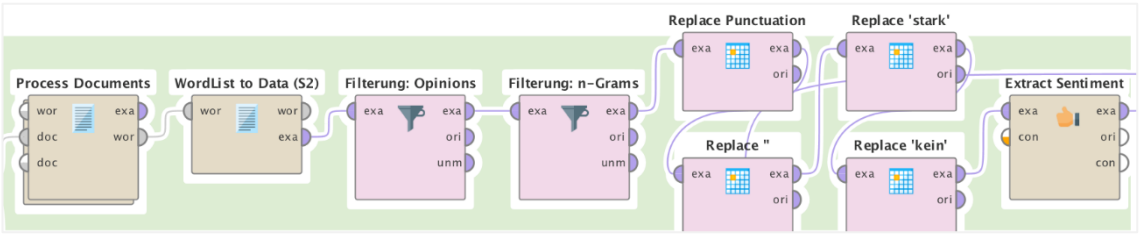

Abbildung 21: Kernprozess des Opinion Mining in RapidMiner

Nachdem die Features im Hilfsprozess identifiziert werden, geht es im Hauptprozess vornehmlich um die Extraktion der Opinion Words sowie die Analyse des Sentiments der Feature <> Opinion Words-Paare. Hierzu finden im Operator *Process Documents* die grundlegenden Verfahren des Natural Language Processing sowie Opinion Mining statt. Abbildung 22 gibt einen Überblick über die eingesetzten Verfahren:

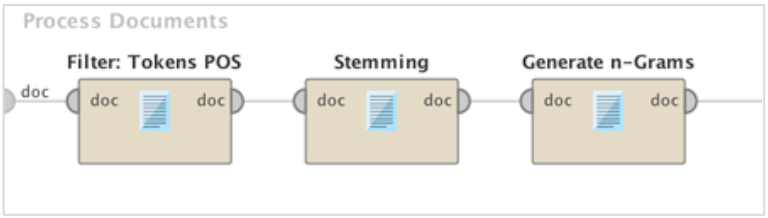

Abbildung 22: Prozess der Opinion Words-Extraktion

Zunächst werden über den *Filter Tokens by POS Tags*-Operator alle Tokens ausgelesen und anhand der definierten POS-Tags gefiltert – die zuvor angesprochene Token-Filterung findet zu einem späteren Zeitpunkt statt. Im Gegensatz zum Hilfsprozess, wo die Token nur die Features abbilden mussten und somit Nomen darstellten, werden im Hauptprozess neben Nomen auch diverse Tags über Adjektive und Adverbien in den Filter aufgenommen. Abbildung 23 zeigt hierzu die Konfiguration des entsprechenden Operators:

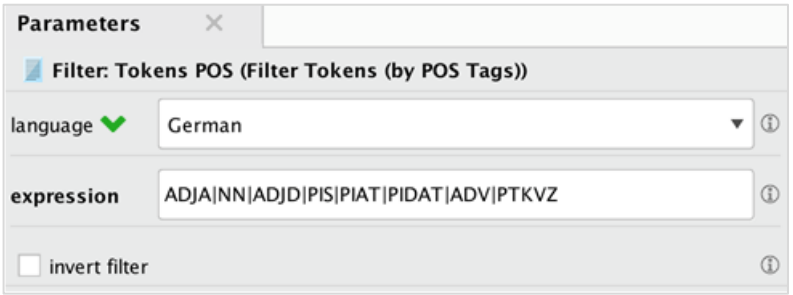

Abbildung 23: POS-Tagging zur Opinion Words-Extraktion

Mittels des folgenden *Stemming* wird sichergestellt, dass die jeweiligen Wortarten inhaltlich zweckmäßig geclustert werden, um spätere Häufigkeitsbetrachtungen im Rahmen der Summarization adäquat abzubilden. Anknüpfend findet mit der *Generierung von N-Grammen* die zentrale Methode zur Abbildung von Feature und Opinion Words-Paaren Anwendung. Hierbei werden alle vorkommenden Wortkombinationen entsprechend der konfigurierten Filter für den weiteren Prozess extrahiert. Ziel ist es hiermit, eine inhaltliche Zuordnung von Features und Opinion Words zu erreichen – basierend auf der Annahme, dass die Opinion Words in unmittelbarer Nähe vor oder nach dem Token im Text platziert sind. Konkret werden vorliegend 6-Gramme gebildet, so dass theoretisch alle Feature <> Opinion Word-Paare berücksichtigt werden, die maximal vier Wörter voneinander getrennt sind. Abbildung 24 zeigt einen Ausschnitt aus den erhaltenen N-Grammen:

Abbildung 24: Auszug der generierten 6-Gramme

Ehedem eine Sentimentanalyse auf die einzelnen Feature <> Opinion Word-Paare in Form der Grammen durchgeführt wird, erfolgt über diverse Filter- und Replace-Operatoren eine Optimierung der Datengrundlage. Über den Operator *Filterung: Features* wird sichergestellt, dass einzig die im Hilfsprozess identifizierten Features im weiteren Prozessablauf berücksichtigt werden – dies reduziert die Datenmenge und folglich die Rechendauer und erhöht gleichzeitig die Schärfe der Analyse. Abbildung 25 zeigt die entsprechende Konfiguration des Operators.

Abbildung 25: Filterung des Modells auf identifizierte Features

Es folgt eine *Filterung der N-Gramme*, um für die vorliegende Fallstudie einzig 6-Gramme in die Auswertung aufzunehmen. Im Anschluss bewirken zwei *Replace*-Operatoren durch die Entfernung von Punktationen und Symbolen eine saubere und lesefreundliche Datengrundlage. Zusätzlich ist ein *Replace*-Operator in das Modell implementiert, um das Wort „stark“ durch „schlecht“ zu ersetzen. Die Intention liegt hierbei in einer verbesserten Qualität der Sentimentanalyse, weil Probeläufe gezeigt haben, dass das Adjektiv „stark“ als positives Sentiment interpretiert und verarbeitet wird, wohingegen dieses im vorliegenden Beispiel in Verbindung mit den Features „Stuhlgang“ und „Blähungen“ konsistent negativ konnotiert ist. Aus diesem Grund wird mit einer Ersetzung durch das Adverb „schlecht“ eine erkennbar negative Konnotation für die automatisierte Sentimentanalyse erreicht. Analog hierzu wird das Adverb „kein“ durch das Adjektiv „gut“ ersetzt, um somit den umgekehrten, positiven Fall adäquat zu berücksichtigen.

Abschließend ermöglicht der Operator *Extract Sentiment* eine automatisierte semantische Erfassung der identifizierten Wortkombinationen, indem auf Basis des in RapidMiner hinterlegten Valence Aware Dictionary and sEntiment Reasoner (VADER)-Lexikons eine Zuordnung positiver wie negativer Sentiments an die Adjektive innerhalb eines 6-Gramms erfolgen. Hierbei erhält jedes im zugrundeliegenden Lexikon hinterlegtes Opinion Word einen graduellen Scoring-Wert zwischen -1 und 1, wobei Werte über 0 ein positives Sentiment kennzeichnen, Wert gleich 0 als neutral einzustufen sind und Werte kleiner 0 als negatives Sentiment interpretiert werden. Die Aufrechnung aller Sentiments innerhalb der einzelnen 6-Grammen führen zu einem Gesamtscore, welcher wiederum als Referenz für das Sentiment des Features herangezogen wird. Abbildung 26 zeigt hierzu zwei Beispiele.

word	Score	Scoring String	Negativity	Positivity
wass_jahr_hund_zunehm_probl_fell	0		0	0
weich_wenig_gar_gut_schuppig_fell	0.718	weich (0.18) gut (0.54)	0	0.718
wenig_gar_gut_schuppig_fell_mehr	0.538	gut (0.54)	0	0.538
wied_hund_futt_leid_schlecht_blahung	-0.513	schlecht (-0.51)	0.513	0

Abbildung 26: Auszug erhaltener Sentimente in RapidMiner

Das in Zeile 2 aufgeführte 6-Gramm „weich_wenig_gar_gut_schuppig_fell“ erhält einen positiven Gesamtscore und wird somit als positive Bewertung hinsichtlich des Features „Fell“ interpretiert, da mit „weich“ und „gut“ zwei positiv konnotierte Adjektive (=Opinion Words) mit „Fell“ in Verbindung gebracht werden. Demgegenüber enthält das 6-Gramm „wied_hund_futt_leid_schlecht_blahung“ in Zeile 4 eine negative Bewertung für das Feature „Blähungen“ aufgrund des Adjektivs „schlecht“. Auf Grundlage dieses Vorgehens lassen sich nun die ermittelten Scores für die einzelnen Features aggregieren und in der Folge ein Fazit darüber ziehen, wie die artikelindividuellen Produktfeatures von den Kunden entsprechend eingeschätzt werden. Einschränkend ist jedoch zu betonen, dass durch die sequenzielle Zusammenfassung der Textfragmente zur Generierung der 6-Gramme jedes Token einer Rezension bis zu sechsmal in der entsprechenden Sentimentanalyse berücksichtigt wird, was folglich bei der Analyse und Interpretation der Ergebnisse zwingend zu beachten ist.

3.6 Evaluation der Daten- und Modellqualität

3.6.1 Diskussion der Studienergebnisse

Für eine flexible Auswertung wurden die Daten aus RapidMiner in Excel überführt und aufbereitet. Abbildung 27 bietet die resultierende Ergebnisübersicht:

Artikel: bosch Adult Lamm & Reis															
Durchfall				Fell				Blähungen				Stuhlgang			
pos.	neg.	neu.	∑	pos.	neg.	neu.	∑	pos.	neg.	neu.	∑	pos.	neg.	neu.	∑
31	16	37	84	12	5	37	54	3	14	19	36	10	0	10	20
37%	19%	44%	100%	22%	9%	69%	100%	8%	39%	53%	100%	50%	0%	50%	100%

Abbildung 27: Ergebnis des Opinion Mining vor Klassifikation

Hierzu wurden die ermittelten Scoring-Werte der einzelnen 6-Gramme der jeweiligen Features in die Kategorien positiv (für Scoringwerte > 0), neutral (für Scoringwerte = 0) sowie negativ (für Scoringwerte < 0) eingeteilt. Aufgrund der inhaltlichen Nähe der

Features Durchfall, Blähungen und Stuhlgang werden diese nachfolgend gemäß Abbildung 28 unter dem Feature gastrointestinale Verträglichkeit aggregiert:

Artikel: bosch Adult Lamm & Reis							
Fellverträglichkeit				Gastrointestinale Verträglichkeit			
pos.	neg.	neu.	∑	pos.	neg.	neu.	∑
31	16	37	84	44	30	66	140
37%	19%	44%	100%	31%	21%	47%	100%

Abbildung 28: Ergebnis des Opinion Mining nach Klassifikation

Aus den Ergebnissen ist abzuleiten, dass der betrachtete Hundefutter-Artikel sowohl hinsichtlich der Fellverträglichkeit mit einem 2:1-Verhältnis positiver zu negativer Bewertungen als auch der gastrointestinalen Verträglichkeit mit einem 1,5:1-Verhältnis überwiegend positiv beurteilt wird. Hieraus wäre für das Category Management inhaltlich eine mögliche Schlussfolgerung, dass Fell- sowie gastrointestinale Verträglichkeit wesentliche Produkteigenschaften und -anforderungen darstellen, welche von Hunde-Trockenfutterprodukten explizit zu erfüllen sind. Bezugnehmend auf den betrachteten Artikel könnte eine Prüfung über die potenzielle Einlistung dieses Artikels für das eigene Sortiment stattfinden, aber auch eine tiefergehende Analyse der Produktnährstoffe zweckmäßig sein, um hieraus zusätzliche Daten und Insights zu generieren, welche zugrundeliegenden Nährstoffzusammensetzungen offensichtlich positive Resonanzen vom Markt erfahren. Zur Verfestigung und Ausbau des Kundenverständnisses wäre eine umfassende Analyse von Produkten gleicher, aber auch unterschiedlicher Marken & Kategorien ratsam, um somit die inhaltlichen Ableitungen sowohl auf valider Datenbasis zu fahren als auch durch Aggregation der erhaltenen Produktdaten auf Marken- und Kategorieebene entsprechende Insights auf weiteren Analyseebenen zu entwickeln. Somit ist die Extraktion und Aufbereitung der Daten gemäß den gestellten Category Management-Anforderungen durch den vorgestellten Opinion Mining-Prozess in nachgefragter Form und Struktur prinzipiell gelungen – die Evidenz der Studienergebnisse ist jedoch nachfolgend noch kritisch zu hinterfragen.

Hierzu gibt Abbildung 29 in den beiden angefügten Zeilen zusätzlich die Ergebnisse einer manuellen Prüfung der betrachteten Kundenrezensionen an.

Artikel: bosch Adult Lamm & Reis							
Fellverträglichkeit				Gastrointestinale Verträglichkeit			
pos.	neg.	neu.	∑	pos.	neg.	neu.	∑
31	16	37	84	44	30	66	140
37%	19%	44%	100%	31%	21%	47%	100%
6	*1*		*7*	*41*	*14*		*55*
86%	*14%*		*100%*	*75%*	*25%*		*100%*

Abbildung 29: Ergebnisabgleich von Opinion Mining und manueller Analyse

Methodisch wurden hierfür die 40 Textrezensionen manuell auf entsprechende Features und dazugehörige Kundenmeinungen untersucht. Aus den Ergebnissen wird offensichtlich, dass das Opinion Mining ein Vielfaches an Meinungen und Sentiments erfasst. Der Grund hierfür wurde bereits im Rahmen des Modelling offenkundig – das Verfahren der 6-Gramm-Generierung zur Beachtung der inhaltlichen Verknüpfung von Feature und Opinion Words führt zwangsläufig zu einer vervielfachten Berücksichtigung in den Analysen. Aus diesem Grund sind die absoluten Häufigkeiten aus dem Opinion Mining-Prozess nicht valide und von jeder unmittelbaren Interpretation auszuschließen. Bei Betrachtung der relativen Verteilung sowohl der Häufigkeit der bewerteten Features wie auch der jeweiligen Sentiments ist zu konstatieren, dass die grundlegenden inhaltlichen Ableitungen aus dem Opinion Mining bei manueller Prüfung gehalten werden können. Demnach zeigen die Daten hier ebenso eine deutlich erhöhte Bewertung der gastrointestinalen Verträglichkeit im Vergleich zur Fellverträglichkeit, so dass diese als wichtigstes Feature zu sehen ist. Außerdem werden beide Features im Rahmen ebenfalls überwiegend positiv bewertet, wenngleich relativ betrachtet deutlich positiver als im Zuge des Opinion Mining. Beim Abgleich der Sentimente auf Feature-Ebene werden kontextuelle Ungenauigkeiten des Opinion Mining-Prozesses ersichtlich. Exemplarisch angeführt sei folgende Artikelrezension:

> Wir haben einen sensiblen Malinois und waren länger auf der Suche nach Gründen für Durchfall, mattes Fell und leichten Haarausfall. Dementsprechend mussten wir oftmals das Futter wechseln/ausprobieren. Bei Bosch sind wir fündig geworden!Tolles Haar!Keine Probleme mit dem Magen. Der Hund frisst es gerne und die Sorte ist bis jetzt nicht relevant gewesen. Immer wieder

Abbildung 30: Kundenrezension mit fehlerhafter Sentimentanalyse

In dieser beschreibt der Rezensent die Probleme seines Hundes vor Nutzung des betrachteten Artikels, welche mit den zugrundeliegenden Features übereinstimmen. Da das vorliegende Opinion Mining-Modell jedoch kein Verfahren nutzt, welches befähigt ist einen solchen inhaltlichen Kontext zu erfassen, wird diese Rezension sowohl für das Feature Durchfall als auch für die Fellverträglichkeit gemäß Abbildung 31 dem betrachteten Artikel negativ zugeordnet.

word	Score	Scoring Stri...	Negativity	Positivity
durchfall_matt_fell_leich_haarausfall_dementsprech	-0.154	matt (-0.15)	0.154	0

Abbildung 31: Beispiel einer fehlerhaften Sentimentanalyse im Modell

Hierbei ist zusätzlich ersichtlich, dass durch die Länge des generierten N-Gramms ein Sentiment mehreren Features zugeordnet wird, obwohl teilweise nur ein Feature direkt angesprochen wird – in diesem Beispiel Fell. Eine Verkürzung der generierten N-Gramms würde wiederum die Anzahl der erkannten Feature <> Opinion Words reduzieren: Demzufolge ist letztlich eine Abwägung zwischen vollständigen und validen Daten zu treffen.

Darüber hinaus ist beim zusätzlichen Abgleich der identifizierten Features aufgefallen, dass der Geschmack des Futters – trotz 17-maliger Nennung bei manueller Prüfung – im Rahmen des Opinion Mining nicht als solches erkannt wurde. Dies resultiert aus den gewählten Formulierungen der Kunden, die eine Beurteilung des Geschmacks überwiegend über die Verbform vornehmen, wie an den aufgeführten Beispielen gemäß Abbildung 32 ersichtlich wird.

Das Futter ist von der Zusammensetzung super,da es kein Weizen enthält und sie darauf allergisch reagiert .Nur meine berner-sennen labrador Dame, rührt es leider nicht an.ich bin grad am umstellen auf Adultfutter. Nun muss ich leider weiter suchen.

Wir hatten schon viele Futtersorten ausprobiert,aber seit unser Hund dieses Futter futtert hat er keinen Durchfall mehr und viel weniger Muskelbeschwerden.Einfach nur super und ihm schmeckt es !!!!!

Abbildung 32: Kundenrezensionen mit fehlerhafter Feature-Identifikation

Da die Identifikation der Features auf Nomen basiert, schlägt an dieser Stelle die Identifikation fehl. Außerdem sei in Anbetracht der vielfältigen Ausdrucksmöglichkeiten darauf verwiesen, dass durch die Verwendung von Synonymen ebenso die Identifikation

von Features – aber auch die Analyse von Opinion Words – erschwert und stellenweise kaum realisierbar ist.

Aufgrund der mannigfaltigen Einschränkungen und Deviationen folgt eine detaillierte Reflexion der gewählten Methodik um abzuwägen, ob die Ergebnisse der umgesetzten Modellierung gemäß CRISP-DM in das Deployment übergeben werden könnten oder zunächst weitere Modifikationen am Modell notwendig erscheinen. Abschließend werden die gewonnenen Erkenntnisse aus der Fallstudie den Anforderungen einer potenziellen praktischen Implementierung gegenübergestellt.

3.6.2 Kritische Reflexion

Die Zielsetzung umfasst eine Evaluation des Feature-based Mining auf Basis eines MVP-Ansatzes, welcher per definitionem nicht die maximale Detailgenauigkeit zum Anspruch, jedoch zu grundsätzlich validen und reliablen Ergebnissen zu führen hat. Zur Untersuchung dieser Anforderung sind nachfolgend die einzelnen Verfahrensschritte kritisch zu reflektieren und auf potenzielle Alternativlösungen oder Erweiterungen aus der Literatur zu verweisen. Dabei ist kritisch festzuhalten, dass die Schlussfolgerungen aus der Fallstudie auf der Analyse eines Beispielartikels basieren, was insbesondere die Vollständigkeit der Konklusionen infrage stellt, jedoch dem vorgegebenen inhaltlichen Umfang entspricht.

So lassen sich im Rahmen der Data Preparation zwei Limitationen dieser Studie festhalten, die in Anbetracht einer möglichen praktischen Adaption als hinreichende oder notwendige Erweiterung des Modells zu überprüfen wären. Hierbei ist zunächst die Datenquelle respektive -anbindung zu thematisieren. In der vorliegenden Fallstudie liegt der Fokus rein auf einen bestimmten Webshop. Hierbei ist es jedoch denkbar, dass die benötigte Datengrundlage über mehrere Webshops oder auf unterschiedlichen Seitentypen im World Wide Web verteilt ist, so dass im komplexesten Fall eine Datenextraktion aus heterogenen Datenquellen notwendig ist, was wiederum die Datenbeschaffung deutlich erschwert. Verwiesen sei hierbei exemplarisch auf die tiefergehenden Forschungsarbeiten von Bao, Collier & Datta (2013, S. 240f.) sowie Gao, Li & Darwish (2012, S. 1173ff.) über die Berücksichtigung unterschiedlicher Datenquellen, etwa aus Social Media Streams. Eine weitere Vereinfachung vorliegender Fallstudie liegt darin, dass die Reliabilität der genutzten Produktbewertungen nicht geprüft wird, so dass einer potenziellen Verzerrung der Ergebnisse durch Fake Reviews nicht entgegengewirkt wird. Entsprechende Ansätze von Cardoso et al. (2018, S. 11ff.) und Mukherjee et al. (2014, S. 192ff.) zur automatisierten Entdeckung und Filterung von

Fake Reviews könnten zur Erhöhung der Datenqualität in ein Opinion Mining Modell eingebettet werden.

Auch im Modelling sind auf einige Limitationen des gewählten Ansatzes zu verweisen. Eine gravierende Problematik stellt sich bei der Identifikation von Features dar, da über entsprechenden POS-Tag nur Nomen erfasst werden. Somit werden implizite Features, die nicht unmittelbar in Form von Substantiven genannt und gemeinhin mit Verben umschrieben werden, nicht identifiziert. Zhang & Zhu (2013, S. 103f.) bieten hierfür einen potenziellen Lösungsansatz. Eine einfache Ausweitung des POS-Tags würde im vorliegenden Modell nicht genügen, da forthin das notwendige Textverständnis zur endgültigen Identifikation fehlt – und der manuelle Aufwand würde in Anbetracht der großen Anzahl an Verben unverhältnismäßig ansteigen. Zusätzlich werden in der vorliegenden Ausarbeitung Synonyme ausgeklammert, so dass die Ermittlung der absoluten Häufigkeiten der einzelnen Tokens unvollständig geschieht. Synonyme lassen sich jedoch über entsprechende Lexika und Wortlisten mit Zusatzaufwand prinzipiell in das Modell überführen. Eine fortgeschrittene Alternativlösung wäre die Nutzung des neuronalen Netzes word2vec zur Wortvektorisierung analog Zuo et al. (2018, S. 253): Durch die Methode des word embedding werden hierbei Textklassifikationen und -clustering ermöglicht, was sowohl die Identifikation von Token als auch von Opinion Words unterstützt. Entsprechende Extensions sind auch für RapidMiner verfügbar. Generell sind bei der Extraktion der Opinion Words die soeben angebrachten Kritikpunkte analog anzuführen: Für die Identifikation von Opinion Words werden Nomen exkludiert, wenngleich diese ebenfalls als meinungsausdrückende Worte denkbar sind. Somit sind vorliegend fälschlicherweise nicht-identifizierte Sentiments aufgrund nicht extrahierter Opinion Words nicht auszuschließen. Weitere Limitationen sind bei der Zuordnung von Opinion Words zu den jeweiligen Features festzuhalten. Die vorgestellte Methode der Textfragmentzusammensetzung über die Generierung von N-Grammen ermöglicht im Modell die Bildung von Tokens und Opinions Words-Paaren unter der Annahme, dass diese innerhalb einer zu definierenden Maximalmenge an Wörter beieinanderstehen. Dabei bringt diese Annahme jedoch auch die Problematik mit sich, dass Features oder Opinions Words redundant erfasst oder falsch gegenseitig zugeordnet werden, was letztlich die Ergebnisqualität reduziert. Eine entsprechende Reduktion der Fragmentisierung führt hingegen dazu, dass weniger Feature und Opinion Words-Paare identifiziert werden, so dass der abgebildete Prozess einen Zielkonflikt zwischen Vollständigkeit und Genauigkeit beinhaltet. In der Literatur werden für dargestellte Problemstellungen unterschiedliche regelbasierte Konzepte wie von

Somprasertsri & Lalitrojwong (2010, S. 942f.) sowie Machine Learning-Ansätze wie von Pang et al. (2002, S. 83f.) und Jin et al. (2009, S. 1196ff.) diskutiert, welche die abgebildete N-Gramm-Methodik mithilfe künstlicher Intelligenz optimieren. Bei der automatisierten Sentimentanalyse auf Basis des in RapidMiner hinterlegten VADER-Lexikon ist zu beachten, dass dieses nur Adjektiven ein entsprechendes Sentiment zuordnet, was wiederum Fehlerpotenziale oder manuelle Korrekturaufwände zur Folge hat. Für die vorliegende Ausarbeitung wurden exemplarisch Replace-Methoden in das Modell eingefügt, um hierdurch eine umfassendere Sentimentanalyse herbeizuführen, welche weithin jedoch keine vollständige Erfassung ermöglicht. Weiterführende Lexika oder Machine Learning-Methoden sind auch in diesem Kontext denkbar.

Im Zuge der Datenaufbereitung ist zu konstatieren, dass die Unterteilung der Sentimente in positiv wie negativ ohne graduelle Abstufungen den Informationsgehalt der Ergebnisse mindert. Bezugnehmend auf die Auswertung von Produktrezensionen gibt es in der Forschung umfassende Arbeiten von Zhang & Narayanan (2010, S.4ff.) sowie Rajeev & Rekha (2015, S.4) über komparative Analysen und Rankingmethoden, welche für eine Erweiterung eines Opinion Mining-Modells von Interesse sein könnten. Angesprochene Limitationen sind zusätzlich einer abschließenden Risikoeinschätzung in Tabelle 1 zusammengefasst.

Tabelle 1: Limitationen des vorgestellten Modells

Phase	Limitation (Erweiterung)	Risikoeinschätzung
Data Preparation	keine Berücksichtigung der Reliabilität von Produktrezensionen	gering bis mittel (abhängig v. Sachverhalt)
	Cross Media Selection	abhängig von Sachverhalt
Modelling	Unvollständige Identifikation von Features	sehr hoch
	Unvollständige Identifikation von Opinion Words	gering bis mittel (abhängig der Accuracy)
	Ungenaues Mapping von Features und dazugehörigen Opinion Words	mittel bis hoch (abhängig der Accuracy)
	Unvollständige Sentimentextraktion	gering bis mittel (abhängig der Accuracy)
	Summarization	gering (abhängig der Detailtiefe)

Demnach lässt sich konstatieren, dass insbesondere die unvollständige Erfassung von Features ein Risiko für Fehlschlüsse darstellt, da die Gefahr besteht, entscheidende Produkteigenschaften in der Sentimentanalyse unberücksichtigt zu lassen – wie in der vorliegenden Fallstudie der Geschmack des Hundefutters. Auch die Qualität des Mappings von Feature und Opinion Word birgt ein Risiko, wenngleich sich bei großen Datenmengen entsprechende Ungenauigkeiten nach dem Gesetz der großen Zahl nivellieren dürften, so dass grundlegende Schlussfolgerungen hierbei mit hoher Wahrscheinlichkeit als korrekt einzustufen sind – insbesondere unter der Prämisse eines MVP-Ansatzes. Die weiteren identifizierten Limitationen sind jeweils in Anbetracht des Zielkonflikts von Genauigkeit versus Aufwand abzuwägen. Eine Berücksichtigung angesprochener Erweiterungen aus der Forschung würden einen MVP-Ansatz jedoch bei weitem übersteigen, so dass diese keine Berücksichtigung im Modell erfahren. Angewandt auf die umgesetzte Fallstudie wäre nach CRISP-DM jedoch eine erneute Modelling-Phase unabdingbar, um das unberücksichtigte Feature „Geschmack“ in die Auswertung zu überführen.

4. Fazit & Ausblick

Das Ziel ist es, einen Opinion-Mining-Prozess auf Basis eines MVP-Ansatzes zu entwickeln, praktisch umzusetzen und abschließende Implikationen für dessen Praxistauglichkeit zu eruieren. Hierfür erfolgte zunächst eine Auseinandersetzung mit den in der Forschung diskutierten grundlegenden Verfahren und Techniken des Opinion Mining, ehedem infolge eines Austauschs mit potenziellen Anwendern der Feature-based Mining-Ansatz als zugrundeliegendes Untersuchungsobjekt definiert wurde. Hierzu wurde ein entsprechendes Prozessmodell zur Umsetzung entwickelt und im Rahmen einer Fallstudie praktisch angewandt, welche die Analyse wesentlicher Produktanforderungen im Online-Tierbedarf auf Basis von Kundenrezensionen umfasste. Hierbei wurden zunächst die entsprechenden Textrezensionen automatisiert aus dem World Wide Web ausgelesen, bevor im analytischen Teil die nachgefragten Produktanforderungen und deren Erfüllung mittels des konzipierten Opinion Mining-Prozesses ausgewertet wurden. Trotz der augenscheinlich zweckmäßigen Ergebnisse zeigten sich diverse Limitationen, die mit der Anwendung des vorgestellten Ansatzes einhergingen. Zur abschließenden Beantwortung der zugrundeliegenden Forschungsfrage, ob der entwickelte Feature-based-Mining Ansatz identifizierte Anwendungskriterien potenzieller Endanwender erfüllt und somit als praxistauglich einzustufen sei, sind nachfolgend den gewonnenen Erkenntnissen über Mehrwert und Limitationen den Anwendungskriterien gegenüberzustellen. Hierzu fasst Abbildung 33 die wesentlichen Teilaspekte zusammen:

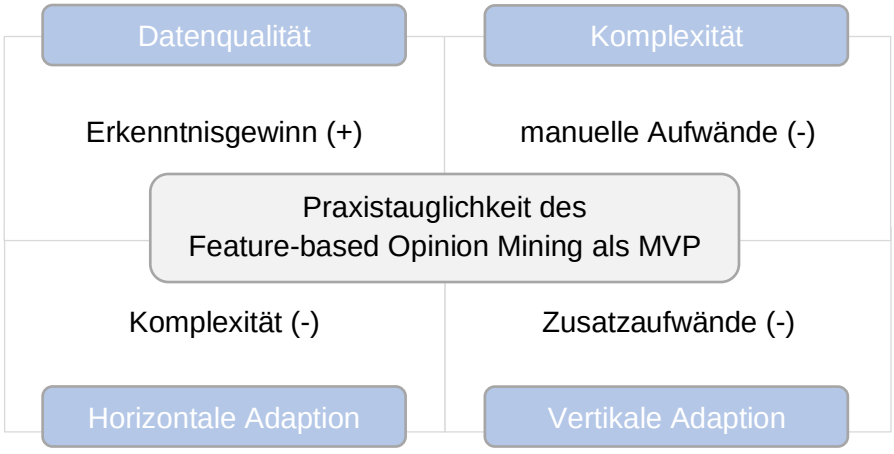

Abbildung 33: Praxistauglichkeit des Feature-based Opinion Mining

Die Ergebnisqualität ist hierbei positiv zu bewerten, da die Fallstudie den potenziellen inhaltlichen Mehrwert, der aus den Ergebnissen des vorgestellten Opinion Mining-Ansatzes resultiert, unterstreicht. Es wurden zwar Limitationen der Vorgehensweise identifiziert und kritisch proklamiert, diese sind jedoch mit den exemplarisch angeführten – teilweise manuellen – Methoden grundsätzlich auszumerzen. Demgegenüber stehen Erkenntnisgewinne über Markt- und Kundenanforderungen als Nutzen, welche aufgrund der hiesigen Menge an Kundenrezensionen sowie des hohen Aufwandes, Informationen aus Fließtexten manuell zu extrahieren und aufzubereiten, ohne unterstützende Opinion Mining-Methoden in der Praxis nicht effizient zu gewinnen wären. Gleichwohl ist jedoch anzuführen, dass der Feature-based Mining-Prozess in der vorgestellten Methodik manueller Interventionen – beispielsweise im Teilprozess der Feature-Auslese – bedarf, so dass kein vollständiger Automatisierungsgrad erreicht wird. Ausdrücklich betont sei hierbei die notwendige Ausrichtung der technischen Umsetzung an die zugrundeliegenden Daten sowohl bei der Datenbeschaffung wie auch der Datenanalyse, welche Expertenwissen bedarf und demnach Opinion Mining als innerbetrieblichen Geschäftsprozess komplex macht. Hier bedarf es Analytics-Unterstützung, so dass primäre Endanwender, wie beispielsweise Category Manager, nicht in der Lage sind, entsprechende Analysen selbstständig umzusetzen. Angesprochene Komplexität ist ebenso negativ hinsichtlich der Adaptionsmöglichkeiten zu bewerten: Sowohl die Anbindung an heterogene Datenquellen wie auch die Anpassung an variierende Analyseinhalten bedarf eines fortgeschrittenen Know-hows und manuellen Zusatzaufwendungen, so dass diese Kriterien unter der vorliegenden Prämisse eines MVP-Ansatzes als nicht gegeben angesehen werden können.

Somit lässt sich als Fazit konstatieren, dass das Feature-based Mining entsprechend nicht als MVP-Ansatz in der Praxis umzusetzen ist, sondern Fachwissen im Bereich Text Mining und erhöhten Zeitaufwänden für eine adäquate Modellierung mit validen Ergebnissen benötigt. Dies ist in Anbetracht der betrachteten Analyseebene auch nicht verwunderlich: Im Gegensatz zu den weiteren Methoden des Opinion Mining fußt das Feature-based Mining auf der kleinstmöglichen Betrachtungsebene, indem jedes einzelne Wort als potenzielles Feature einem individuellen Opinion Mining-Prozess zugrunde gelegt werden kann. Aus diesem Grund ist eine erhöhte Prozesskomplexität zwangsläufig, welche sich in den vielfältigen Diskussionen und Ansätzen in der Forschung zeigt. Als praktische Implikation und weiterführenden Ausblick ist folglich festzuhalten, dass eine tiefergehende Auseinandersetzung aus analytischen Gesichtspunkten zur Generierung von Insights über Kunden und Märkte große

Potenziale bietet, gegenwärtig jedoch hohe zeitliche Aufwände wie analytische Kapazitäten voraussetzt und für eine effiziente Ressourcenallokation prozesstechnisch schmalere und kostengünstigere Lösungen bedarf. Somit stellt insbesondere der Ansatz einer möglichst automatisierten, adaptionsfähigen technischen Lösung ein breites Forschungsfeld dar, welches unter anderem die Vereinigung bereits diskutierter separierten Methodenansätze zum Zweck haben sollte. Zusätzlich sind vor kommerzieller Nutzung juristische Feststellungen zu treffen, da die Verwendung von Daten auf Basis von Scraping- und Crawlingvorgängen grundsätzlich als rechtlich problematisch anzusehen ist.

Literaturverzeichnis

Bao, Y., Collier, N. & Datta, A. (2013): A Partially Supervised Cross-Collection Topic Model for Cross-Domain Text Classification. *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, 239-248. https://doi.org/10.1145/2505515.2505556

Chowdhury, G. (2005): Natural Language Processing. *Annual Review of Information Science and Technology*, *37*(1), 51-89. https://doi.org/10.1002/aris.1440370103

Cardoso, E., Silva, R. & Almeida, T. (2018): Towards automatic filtering of fake reviews. *Neurocomputing*, 309, 106-116. https://doi.org/10.1016/j.neucom.2018.04.074

Cui, G., Lui, H. & Guo, X. (2012). The Effect of Online Consumer Reviews on New Product Sales. *International Journal of Electronic Commerce, 17*(1), 39-58. https://doi.org/10.2753/JEC1086-4415170102

Dury, M. (2020): Webscraping, Screenscraping und das Datenbankurheberrecht. *https://www.dury.de/onlinerecht-blog/webscraping-screenscraping-und-das-datenbankurheberrecht*, zuletzt abgerufen am 21.08.2021.

File, K., Cermak, D. & Prince, R. (1994). Word-of-Mouth Effects in Professional Services Buyer Behaviour. *Service Industries Journal, 14*(3), 301-314. https://doi.org/10.1080/02642069400000035

Frow, P. & Payne, A. (2007). Towards the 'perfect' customer experience. *Journal of Brand Management, 15*(2), 89-101. https://doi.org/10.1057/palgrave.bm.2550120

Ganeshbhai, S. & Shah, B. (2015): Feature based opinion mining: A survey. *IEEE International Advance Computing Conference (IACC)*, 919-923. https://doi.org/10.1109/IADCC.2015.7154839

Gao, W., Li, P. & Darwish, K. (2012): Joint Topic Modeling for Event Summarization across News and Social Media Streams. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 1173-1182. https://doi.org/10.1145/2396761.2398417

Hippner, H. & Rentzmann R. (2006): Text Mining. *Informatik-Spektrum, 29*(4), 287-290. https://doi.org/10.1007/s00287-006-0091-y

Hu, M. & Liu, B. (2004): Mining and Summarizing Customer Reviews. *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 167-177. https://doi.org/10.1145/1014052.1014073

Jin, W., Hay Ho, H. & Srihari, R. (2009): OpinionMiner: A Novel Machine Learning System for Web Opinion Mining and Extraction. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1195-1204. https://doi.org/10.1145/1557019.1557148

Klass, E. (2019). Data Mining und Text Mining: kleine Unterschiede, große Wirkung. *Wirtschaftsinformatik & Management, 11*(4), 267-269. https://doi.org/10.1365/s35764-019-00178-6

Mukherjee, A., Liu B. & Glance, N. (2012): Spotting Fake Reviewer Groups in Consumer Reviews. *Proceedings of the 21st International Conference on World Wide Web*, 191-200. https://doi.org/10.1145/2187836.2187863

OLG Köln, Urt. v. 28.02.2020: 6 U 128/19. https://openjur.de/u/2201650.html

Pang, B., Lee, L. & Vaithyanathan, S. (2002): Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002),* 79-86. https://doi.org/10.3115/1118693.1118704

Popescu, AM. & Etzioni, O. (2005): Extracting Product Features and Opinion from Reviews. *Proceedings of the Conference Human Language Technology and Empirical Methods in Natural Language Processing,* 339-346. https://doi.org/10.3115/1220575.1220618

Rajeev, V. & Rekha, S. (2015): Recommending Products to Customers using Opinion Mining of Online Product Reviews and Features. *IEEE International Conference on Circuit, Power and Computing Technologies [ICCPCT 2015]*, 1-5. https://doi.org/10.1109/ICCPCT.2015.7159433

Reichheld, F. & Schefter, P. (2000). E-Loyalty: Your Secret Weapon on the Web. *Harvard Business Review, 78*(4), 105-113.

Shah, D., Rust, R., Parasuraman, A., Staelin, R. & Day, G. (2006). The Path to Customer Centricity. *Journal of Service Research, 9*(2), 113-124. https://doi.org/10.1177/1094670506294666

Singh, P., Sachdeva, A., Mahajan, D., Pande, N. & Sharma, A. (2014): An Approach towards Feature specific Opinion Mining and Sentimental Analysis across E-Commerce Websites. *5th International Conference - Confluence The Next Generation Information Technology Summit*, 329-335. https://doi.org/10.1109/Confluence201433936.2014

Somprasertsri, G. & Lalitrojwong, P. (2010): Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization. *Journal of Universal Computer Science, 16*(6), 938-955. https://doi.org/10.3217/jucs-016-06-0938

Zhang, K., Narayanan, R. & Choudhary, A. (2010): Voice of the Customers: Mining Online Customer Reviews for Product Feature-based Ranking. *Proceedings of the 3rd Workshop on Online Social Networks*.

Zhang, Y. & Zhu, W. (2013). Extracting Implicit Features in Online Customer Reviews for Opinion Mining. *Proceedings of the 22nd International Conference on World Wide Web*, 103-104. https://doi.org/10.1145/2487788.2487835

Zuo, Y., Wu, J., Hui, Z., Wang, D. & Xu, K. (2018): Complementary Aspect-Based Opinion Mining. *IEEE Transactions on Knowledge and Data Engineering*, 30(2), 249-262. https://doi.org/10.1109/TKDE.2017.2764084

Anhang

Datenbank	Suchtaxonomie	Auswahl	Publisher	H-Index
Google Scholar	„Loyalty“ AND „E-Business“	E-Loyalty: your secret weapon on the web	Harvard Business Review	179
SpringerLink	„Customer Experience“ AND „Loyalty“	Towards the ‚perfect‘ customer experience	Journal of Brand Management	50
Taylor & Francis	“Word of mouth” AND “Effects” AND “buying behaviour”	Word-of-Mouth Effects in Professional Services Buyer Behaviour	Service Industries Journal	66
Taylor & Francis	“Effect” AND (“Customer Reviews” OR “Consumer Reviews”) AND “online reviews”	The Effect of Online Consumer Reviews on New Product Sales	International Journal of Electronic Commerce	82
Sage	“Customer Centricity”	The Path To Customer Centricity	Journal of Service Research	122
SpringerLink	“Data Mining” AND “Text Mining”	Data Mining und Text Mining: kleine Unterschiede, große Wirkung	Wirtschaftsinformatik & Management	-
IEEE Explore	“Feature based” AND “opinion Mining”	Feature based opinion mining: A survery	2015 IEEE International Advance Computing Conference (IACC)	14
SpringerLink	“Text Mining”	Text Mining	Informatik-Spektrum	19

IEEE Explore	“Opinion Mining” AND “Process” AND “E*Commerce”	An approach towards feature specific opinion mining and sentimental analysis across e-commerce websites	Proceedings of the 5th International Conference on Confluence 2014: The Next Generation Information Technologiy Summit	11
Wiley Online Library	“Natural Language Processing”	Natural Language Processing	Annual Review of Information Science and Technology	56
ACM	“Extraction” AND “Product Features” AND “Customer Reviews”	Extracting Product Features and Opinions from Reviews	Proceedings of the Conference Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing,	35
IEEE Explore	“Opinion Mining” AND “Product Reviews” AND (“Classification” OR “Recommendation”)	Recommending Products to Customers using Mining of Online Product Reviews and Features	IEEE International Conference on Circuit, Power and Computing Technologies, ICCPCT 2015	13
ACM	“Customer Review” AND “Mining” AND “feature-based”	Mining and Summarizing Customer Reviews	Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining	36
researchgate	“information summarization” AND “social media” AND “stream”	Joint Topic Modeling for Event Summarization across News and Social Media Streams	Proceedings of the 21st ACM International Conference on Information and Knowledge Management	59

researchgate	(“Extraction” OR “Collection”) AND “Cross” AND “Text	A Partially Supervised Cross-Collection Topic Model for Cross-Domain Text Classification	Proceedings of the 22nd ACM International Conference on Information and Knowledge Management	59
Elsevier ScienceDirect	“Fake Reviews”	Towards automatic filtering of fake reviews	Neurocomputing	143
ACM	“Fake” AND “Reviews” AND (“Detect*” OR “Spot*”)	Spotting fake reviewer groups in consumer reviews	Proceedings of the 21nd International Conference on World Wide Web	90
ACM	“Features” AND “Customer Reviews” AND “Opinion Mining” AND “implicit”	Extracting implicit Features in Online Customer Reviews for Opinion Mining	Proceedings of the 22nd International Conference on World Wide Web	90
IEEE Explore	“Opinion Mining” AND “cross media”	Complementary Aspect-Based Opinion Mining	IEEE Transactions on Knowledge and Data Engineering	174
researchgate	“Customer Reviews” AND “Features” AND “Opinion” AND “Summarization”	Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization	Journal of Computer Science	28

ACM	“Opinion Mining” AND “Natural Language Processing” AND “Machine Learning” AND “Part of Speech”	OpinionMiner: A Novel Machine Learning System for Web Opinion Mining and Extraction	Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining	104
ACM	“Sentiment” AND “Machine Learning” AND “Naïve Bayes” AND “Support Vector Machines”	Thumbs up? Sentiment Classification using Machine Learning Techniques	Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)	132
ACM	“Customer Review” AND “Mining” AND “feature-based”	Voice of the Customers: Mining Online Customer Reviews for Product Feature-based Ranking.	Proceedings of the 3rd Workshop on Online Social Networks.	-

Ehrenwörtliche Erklärung

Hiermit versichere ich, dass die vorliegende Arbeit von mir selbstständig und ohne unerlaubte Hilfe angefertigt worden ist, insbesondere, dass ich alle Stellen, die wörtlich oder annähernd wörtlich aus Veröffentlichungen entnommen sind, durch Zitate als solche gekennzeichnet habe. Ich versichere auch, dass die von mir eingereichte schriftliche Version mit der digitalen Version übereinstimmt. Weiterhin erkläre ich, dass die Arbeit in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde/Prüfungsstelle vorgelegen hat. Ich erkläre mich nicht damit einverstanden, dass die Arbeit der Öffentlichkeit zugänglich gemacht wird. Ich erkläre mich damit einverstanden, dass die Digitalversion dieser Arbeit zwecks Plagiatsprüfung auf die Server externer Anbieter hochgeladen werden darf. Die Plagiatsprüfung stellt keine Zurverfügungstellung für die Öffentlichkeit dar.

Philipp Lukasewycz,

Krefeld, 28.08.2021