

Produkt-Mining im Web 2.0

Carolin Kaiser

Lehrstuhl Wirtschaftsinformatik II
Universität Erlangen-Nürnberg
Lange Gasse 20
90403 Nürnberg
Carolin.Kaiser@wiso.uni-erlangen.de

Abstract: Um im gestiegenen Qualitäts- und Kostenwettbewerb zu bestehen, setzen viele Unternehmen verstärkt auf Problemlösungen, die auf die Bedürfnisse der Kunden zugeschnitten sind. Dies setzt jedoch die Kenntnis der Kundenwünsche voraus. Das Web 2.0 stellt eine reichhaltige Quelle an Informationen über Produktbewertungen von Kunden dar. Eine manuelle Analyse ist jedoch sehr aufwändig und nur in geringem Umfang möglich. In dieser Arbeit wird ein Konzept zur automatischen Analyse von Produkt-Reviews auf Basis von Text Mining und Data Mining vorgestellt. Es erlaubt die Extraktion, Aggregation und Analyse von Produkteigenschaften und ihren Bewertungen und stellt damit wichtige Informationen für die Produktgestaltung und -verbesserung bereit.

1 Motivation

Gestiegene Kundenanforderungen und härterer Wettbewerbsdruck erfordern eine stärkere Kundenorientierung der Unternehmen. Um diesen Anforderungen gerecht zu werden und sich auf den im Zuge der Liberalisierung und Globalisierung hart umkämpften Märkten von den Wettbewerbern zu differenzieren, bieten immer mehr Unternehmen Problemlösungen an, die auf die Bedürfnisse der Kunden abgestimmt sind [Pi03, S. 181ff]. Voraussetzung für das Angebot von Problemlösungen ist eine detaillierte Kenntnis der Kundenwünsche. Die Bereitstellung der Informationen über die Wünsche und Meinungen der Kunden hinsichtlich der eigenen Produktpalette ist Aufgabe der Produktforschung, einem Teilbereich der Marktforschung [Ko96]. Sie greift zur Erhebung dieser Informationen zurzeit noch vorwiegend auf quantitative Methoden zurück [HB07, S. 3], die sich auf das empirische Überprüfen von vorher festgelegten Hypothesen konzentrieren. Quantitative Untersuchungsmethoden bringen zwar den Vorteil hoher Repräsentativität mit sich, gehen jedoch mit hohen Kosten einher und ermöglichen nur die Untersuchung vorgegebener Fragestellungen [Ke94]. Einen Gegensatz dazu stellen qualitative Methoden dar. Sie garantieren zwar keine Repräsentativität, zeichnen sich jedoch durch Offenheit und natürliche Kommunikationssituationen aus [Ke94]. Hierbei werden die Hypothesen erst auf Basis der erhobenen Daten erstellt und Personen verwenden eigene Ausdrucksformen, so dass die Inhalte realitätsgetreu erfasst werden können.

Eine reichhaltige und kostengünstige Quelle natürlicher Kundenkommunikation stellt das Web 2.0 bereit. Im Zuge der sozialen Orientierung des Internets kommunizieren immer mehr Kunden online und veröffentlichen auf Bewertungsplattformen ihre Meinungen zu Produkten in Form von Reviews. Eine manuelle Analyse der Reviews ist aufgrund der hohen Anzahl an Quellen und Beiträge sehr aufwändig und nur in begrenztem Umfang möglich. Zielsetzung dieses Projektes ist es daher, ein System zu entwickeln, das die automatische Analyse von Produkt-Reviews erlaubt. Hierbei sollen Produkteigenschaften und ihre Bewertungen extrahiert, aggregiert und analysiert werden, um Aufschluss über die Beurteilung der Eigenschaften eigener und konkurrierender Produkte zu erhalten, Abhängigkeiten zwischen Produkteigenschaften aufzudecken und ihren Einfluss auf die Gesamtbewertung zu messen. Zur Realisierung kommen Methoden des Text Mining zum Einsatz.

2 Text Mining

Text Mining zielt auf die Entdeckung und Extraktion von interessanten Zusammenhängen und Wissen aus unstrukturiertem Text ab [KP07, S. 1]. Es ermöglicht die Informationssuche, Textklassifikation, Text-Clusterung und die Informationsextraktion [We05, S. 7ff]. Neben diesen eher faktenorientierten Anwendungsbereichen entwickelt sich aktuell ein neuer Zweig, das affektive Text Mining. Die Informationsextraktion und das affektive Text Mining stellen verschiedene Methoden bereit, die grundsätzlich für die automatische Analyse von Produkt-Reviews geeignet sind. Ansätze aus dem Bereich der Informationsextraktion beschäftigen sich vorwiegend mit der Erkennung von Personen, Orten und Zeitangaben in Texten. Systeme des affektiven Text Mining behandeln vornehmlich die Erkennung von Standpunkten, emotionalen Dimensionen und Meinungen [Sh06, S. 11]. Nur wenige dieser Arbeiten befassen sich mit dem Mining von Produkt-Reviews. Dazu zählen die Systeme von [Da03], [Li05] und [PE07]. Diese ermöglichen zwar eine Extraktion und aggregierte Darstellung von Produkteigenschaften und ihren Bewertungen. Jedoch erlauben sie keine weitergehende Analyse der Produkteigenschaftsbewertungen. Für die Produktgestaltung und -verbesserung fehlen wichtige Informationen über Abhängigkeiten zwischen Produkteigenschaften und deren Einfluss auf die gesamte Einschätzung eines Produktes. Daher zielt dieses Projekt darauf ab, ein Produkt-Mining-System zu entwickeln, das sowohl die Extraktion und Aggregation als auch die Analyse der Produkteigenschaften und Bewertungen ermöglicht. Hierbei werden die Methoden des Text Mining um Methoden des Data Mining ergänzt.

3 Produkt-Mining

3.1 Überblick

Das Produkt-Mining-System (siehe Abbildung 1) ist in Java implementiert und besteht aus drei verschiedenen Komponenten, die in sequentieller Reihenfolge aufgerufen werden.

Die Extraktionskomponente erhält als Input Reviews, die mithilfe eines Crawlers von Produktbewertungsportalen herunter geladen werden. Aus diesen Reviews extrahiert sie mithilfe von Methoden des Open Source Framework GATE (<http://gate.ac.uk/>) die Produkteigenschaften und Produktbewertungen, die von den Kunden beschrieben sind, und speichert sie in einer MySQL-Datenbank ab. Die Extraktionskomponente transformiert damit unstrukturierten Text in strukturierte Daten. Die strukturierten Daten dienen als Input für die Aggregationskomponente, die eine Verknüpfung und Zusammenfassung der Daten ermöglicht. Im Rahmen der Verknüpfung werden die Bewertungen den Produkteigenschaften zugeordnet. Die Zusammenfassung erlaubt es gleichartige Produkteigenschaftsbewertungen über alle Benutzern hinweg zu aggregieren und grafisch darzustellen. Die verknüpften und aggregierten Daten werden wiederum in der Datenbank abgelegt. Auf Basis dieser Daten führt die Analysekomponente Auswertungen durch. Mithilfe von Assoziationsregeln und Entscheidungsbäumen können Abhängigkeiten zwischen Produkteigenschaftsbewertungen aufgedeckt und Hauptdeterminanten der gesamten Produktbewertung ermittelt werden.

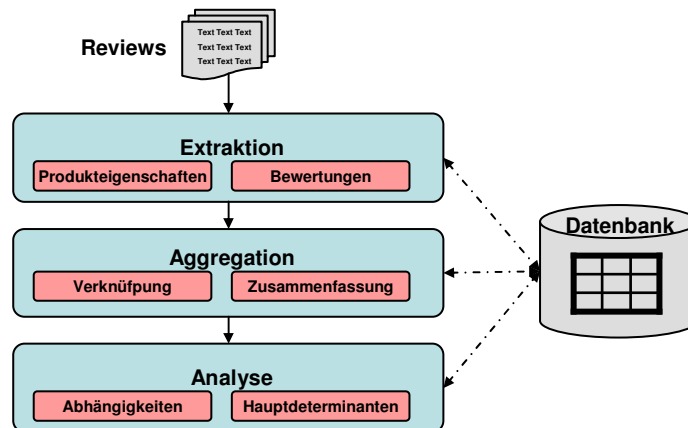


Abbildung 1: Produkt-Mining-System

3.2 Extraktion

3.2.1 Extraktion von Produkteigenschaften

Ziel der Extraktion von Produkteigenschaften ist es, die von den Kunden beschriebenen Charakteristika eines Produktes in Reviews zu erkennen. Bei der Bewertung von Digitalkameras können zum Beispiel die Bildqualität oder die Größe als Produkteigenschaften auftreten. Generell werden zwei Arten von Produkteigenschaften unterschieden: explizite und implizite Eigenschaften. Während explizite Eigenschaften als Wort oder Wortkombination in einem Review auftreten, können implizite Eigenschaften nur aus dem Kontext eines Satzes erschlossen werden [PE07, S. 10; Li05 S. 344]. So ist die Größe der Kamera in dem Satz „Ich liebe die Größe der Kamera.“ explizit enthalten, wohingegen der Satz „Die Kamera ist sehr klein.“ nur implizit auf sie hinweist.

Aufgrund der hohen Komplexität impliziter Eigenschaften werden in diesem Projekt zunächst nur explizite Eigenschaften betrachtet.

Zur Extraktion können im Anschluss an eine linguistische Analyse, die typischerweise eine Zerlegung von Texten in Sätze und Wörter sowie eine Erkennung von Wortformen und -kategorien enthält, prinzipiell Methoden des Überwachten [z. B. Ba00] und des Nicht-Überwachten Lernens [z. B. Et05] herangezogen werden. Beim Überwachten Lernen werden dem Extraktionsalgorithmus Trainingsbeispiele vorgegeben, anhand derer er die Unterscheidung von Eigenschaften und keinen Eigenschaften erlernen soll. Im Gegensatz dazu erhält ein Extraktionsalgorithmus des Nicht-Überwachten Lernens Suchmuster, anhand deren er die Produkteigenschaften in Texten ermittelt. Aufgrund der besseren Lernfähigkeit kommt in diesem Projekt Überwachtes Lernen zum Einsatz.

Im Rahmen des Überwachten Lernens kann die Extraktion von Eigenschaften als Klassifikationsaufgabe betrachtet werden [We05, S. 133]. Ziel der Klassifikation ist es, Objekte anhand ihrer Merkmale in Klassen einzuteilen. Bei der Extraktion von Eigenschaften wird jedes Wort eines Textes in die zwei Klassen „Eigenschaft“ und „Keine Eigenschaft“ eingeteilt. Als Merkmale können zum Beispiel die Form und Kategorie des Wortes selbst sowie umliegender Worte herangezogen werden. In nachstehender Abbildung wird das Wort „Größe“ der Klasse „Eigenschaft“ zugeordnet.

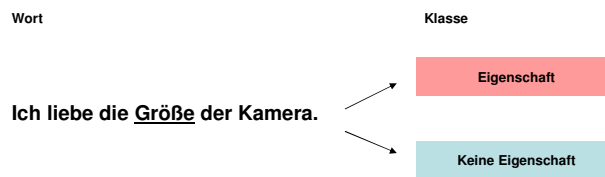


Abbildung 2: Wortklassifikation

Zur Klassifikation von Wörtern können verschiedene Methoden, wie zum Beispiel Hidden Markov Models oder Maximum Entropy, zum Einsatz kommen [We05, S. 135]. Aufgrund ihrer guten Generalisierungsfähigkeit und Möglichkeit eine große Anzahl an Merkmalen zu verarbeiten ist die Stützvektormethode [CV95] besonders zur Klassifikation von Worten geeignet [Mo00; KM00, S. 143]. Daher wird sie zur Extraktion der Produkteigenschaften herangezogen.

Die Stützvektormethode erhält als Input Trainingsbeispiele folgender Form [Ku00, S. 142]:

$$(x_1, y_1), \dots, (x_n, y_n) \quad x_i \in R, y_i \in \{+1, -1\} \quad (1)$$

Dabei stellt x den Vektor der Merkmale des Wortes i und y die Klasse des Wortes dar. Die Wortklasse y nimmt den Wert $+1$ an, wenn das Wort eine Eigenschaft beschreibt, den Wert -1 , wenn das Wort keine Produkteigenschaft bezeichnet. Anhand dieser Trainingsbeispiele erlernt die Stützvektormethode eine Klassifikationsregel, die im einfachsten linearen Fall folgende Form annimmt [Mo00, S. 532]:

$$h(x) = \text{sign}(b + \sum_{i=1}^n w_i x_i) \quad (2)$$

Aufgabe der Stützvektormethode [Mo00, S. 532f] ist es den Gewichtsvektor w und den Schwellwert b so zu wählen, dass $h(x)$ den Wert 1 annimmt, wenn eine Eigenschaft vorliegt, und den Wert -1 , sofern es sich um keine Eigenschaft handelt. Im zweidimensionalen Raum lässt sich die Klassifikationsregel grafisch als Gerade veranschaulichen. Alle Wörter auf einer Seite der Geraden werden als „Eigenschaften“ klassifiziert, alle Wörter auf der anderen Seite als „keine Eigenschaften“. Hierbei wählt die Stützvektormethode die Gerade so aus, dass sie beide Klassen mit maximalem Abstand separiert. Nachstehende Abbildung zeigt eine Gerade, die vier Trainingsbeispiele anhand der Merkmale „Besitzt Wort ‚Kamera‘ als zweites nachfolgendes Wort“ und „Besitzt Wort ‚der‘ als erstes nachfolgendes Wort“ optimal trennt.

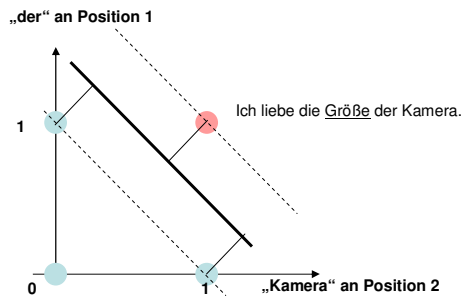


Abbildung 3: Stützvektormethode zur Eigenschaftsklassifikation

Im Rahmen der Eigenschaftsklassifikation werden folgende Merkmale herangezogen:

Merkmale
Wort an Position -5 bis 5
Wortstamm an Position -5 bis 5
Wortart an Position -5 bis 5
grammatikalische Funktion des Wortes an Position -5 bis 5
Haupttyp des Wortes an Position -5 bis 5

Abbildung 4: Merkmale der Eigenschaftsklassifikation

Im Rahmen eines Experimentes wurden 500 zufällig ausgewählte Reviews zu neun verschiedenen Digitalkameras von Amazon herangezogen¹ und alle 4405 darin vorkommenden Eigenschaften manuell markiert. Hierbei wurden 70% der Reviews als Trainingsdaten und 30% als Evaluationsdaten verwendet.

¹Da frei verfügbare linguistische Analyseprogramme für die englische Sprache am weitesten entwickelt sind, wurden englische Reviews verwendet. Nachdem das System prinzipiell genauso mit deutschen Reviews funktioniert, sind alle Beispiele und Abbildungen in Deutsch gehalten um eine einheitliche Sprache zu wahren.

Zur Evaluation werden die Maße Präzision (Precision) und Vollständigkeit (Recall) eingesetzt. Hierbei beschreibt die Präzision, wie viel Prozent der extrahierten Eigenschaften korrekt sind. Die Vollständigkeit gibt an, wie viel Prozent der tatsächlichen Eigenschaften extrahiert wurden. Bei einer dreifachen Kreuzvalidierung wurde eine Präzision von 70% und eine Vollständigkeit von 71% erreicht. Diese Ergebnisse können angesichts der relativ kleinen Review-Anzahl als gut bewertet werden. Mit steigender Review-Anzahl lassen sich die Ergebnisse noch weiter verbessern.

3.2.2. Extraktion von Bewertungen

Ziel der Extraktion von Bewertungen ist es, zu erkennen, wie Produkteigenschaften in Reviews eingestuft werden. Generell unterscheidet man zwei verschiedene Konzepte zur Einordnung von Bewertungen: Polarität und Stärke [Da03, S. 520]. Während die Polarität die Richtung der Bewertung (z. B. gut oder schlecht) bezeichnet, gibt die Stärke die Intensität der Bewertung (z. B. sehr, mittel, gering) an [HW00, S. 299]. Als Modelle können hierbei binäre Klassen [z. B. Pa02], diskrete Klassen [z. B. Li03], Ordnungsrelationen [z. B. WW04] oder Fuzzy-Mengen [z. B. SH01] herangezogen werden. Da die Stärke von Emotionen generell sehr subjektiv geprägt und schwer messbar ist, wird in diesem Projekt allein auf die Polarität abgestellt. Sie wird mithilfe der drei diskreten Klassen „positiv“, „negativ“ und „neutral“ modelliert.

Wie bei der Extraktion von Produkteigenschaften können auch bei der Extraktion von Bewertungen Methoden des Überwachten [z. B. Pa02] und Nicht-Überwachten Lernens [z. B. Tu02] zur Anwendung gelangen. Bei den Methoden des Nicht-Überwachten Lernens wird häufig die Anzahl emotionaler Wörter in Sätzen herangezogen, um die Polarität zu bestimmen [PL04, S. 9]. Dies ist jedoch sehr ungenau. Verneinungen oder irrelevante Inhalte in Sätzen können hierbei nicht oder nur schwierig berücksichtigt werden. Eine genauere Bestimmung der Polarität ermöglicht das Überwachte Lernen, das im Rahmen dieses Projektes zum Einsatz kommt. Hierbei kann die Polaritätsbestimmung als Klassifikationsaufgabe von Sätzen aufgefasst werden. Die Sätze der Reviews werden dabei anhand ihrer Merkmale in die drei Klassen „positiv“, „negativ“ und „neutral“ eingeteilt. Als Merkmale können das Auftreten von Wortformen, die Abfolge von Wörtern und die Struktur des Satzes herangezogen werden. Die Satzklassifikation erfolgt dabei unabhängig von dem Auftreten der Produkteigenschaften. Das Zusammenführen der Produkteigenschaften und Bewertungen wird erst im Rahmen der Aggregation durchgeführt.

Zur Klassifikation wird wie bei der Extraktion der Produkteigenschaften die Stützvektormethode herangezogen. Da die Stützvektormethode nur eine binäre Klassifikation erlernen kann, wird die Aufgabe auf drei Klassifikationen zurückgeführt: „positiv“ versus nicht „positiv“, „negativ“ versus „nicht negativ“, „neutral“ versus „nicht neutral“. Ein neu zu klassifizierender Satz wird mit jeder der drei Regeln klassifiziert und in die Klasse mit der höchsten Wahrscheinlichkeit eingeteilt.

Im Forschungsprojekt werden zur Satzklassifikation Wortstämme und grammatikalische Funktionen aufeinander folgender Wortpaare (Bigramme) als Merkmale herangezogen.

Zur experimentellen Überprüfung wurden alle 6843 Sätze der zufällig ausgewählten 500 Digitalkamera-Reviews manuell als „positiv“, „negativ“ oder „neutral“ markiert und der Stützvektormethode als Trainingsbeispiele übergeben. Bei einer dreifachen Kreuzvalidierung wurden eine Präzision von 69% und eine Vollständigkeit von 83% gemessen.

3.3 Aggregation

3.3.1 Verknüpfung

Die Verknüpfung zielt darauf ab, verschiedene Bezeichnungen einer Produkteigenschaft zusammenzuführen und die Bewertungen zu den Produkteigenschaften zuzuordnen.

In natürlich sprachlichen Texten werden dieselben Produkteigenschaften häufig mit verschiedenen Synonymen benannt. So wird zum Beispiel statt der Bildqualität einer Digitalkamera häufig nur von Bildern oder Fotos gesprochen. Für eine weiterführende Zusammenfassung und Analyse ist es sinnvoll, die Synonyme einer Produkteigenschaft zu finden und sie der Produkteigenschaft zuzuordnen. Zu diesem Zweck wurde manuell ein Lexikon mit Synonymen für Eigenschaften von Digitalkameras definiert und verwendet. Das Synonymen-Lexikon enthält zu jeder definierten Produkteigenschaft eine Liste zugehöriger Synonyme. Da auch häufig umgangssprachliche Begriffe verwendet werden, erscheint die Verwendung bestehender Lexika wenig geeignet. In der Zukunft wird angestrebt, Synonyme über Clusterverfahren aus den Reviews zu erkennen und zu gruppieren, um den manuellen Aufwand zu reduzieren.

Nachdem die Extraktion der Produkteigenschaften und die Extraktion der Bewertungen unabhängig voneinander durchgeführt werden, ist es die Aufgabe der Verknüpfung beide Größen zusammenzuführen. Hierbei werden folgende drei Annahmen getroffen: Produkteigenschaften, die aus positiv klassifizierten Sätzen extrahiert werden, sind positiv bewertet. Eigenschaften, die in negativ klassifizierten Sätzen stehen, werden als negativ bewertet angesehen. Als neutral werden diejenigen Produkteigenschaften eingestuft, die sich in neutral klassifizierten Sätzen befinden. Somit wird jeder Produkteigenschaft die Bewertung des Satzes zugeordnet, in dem sie enthalten ist. Die getroffenen Annahmen stellen zwar eine Vereinfachung dar, jedoch treffen sie in den meisten Fällen zu, so dass keine allzu großen Fehler dadurch entstehen.

3.3.2 Zusammenfassung

Aufgabe der Zusammenfassung ist es, gleichartige Eigenschaftsbewertungen eines Produktes zusammenzuführen und übersichtlich darzustellen. Sie dient dazu einem Hersteller eines Produktes einen Überblick darüber zu geben, welche Eigenschaften eines Produktes besonders oft positiv oder negativ bewertet wurden. Darüber hinaus sind Vergleiche mit anderen Konkurrenzprodukten möglich.

Im Rahmen der Zusammenführung werden zu einem gegebenen Produktnamen alle extrahierten Eigenschaften in der Datenbank gesucht und für jede Eigenschaft die Summe der positiven und negativen Bewertungen gebildet. Die Summen werden im Zuge einer Normierung jeweils durch die Anzahl aller Bewertungen dieses Produktes dividiert, um eine bessere Vergleichbarkeit unter den Produktbewertungen herzustellen.

Nach der Zusammenführung können Produkteigenschaften ausgewählt und visualisiert werden. Als Visualisierungsinstrumente stehen Balkendiagramme und Spider Charts zur Verfügung. Die Balkendiagramme zeigen die normierten Summen der positiven und negativen Bewertungen je gewählter Produkteigenschaft in Form von Balken. Zur besseren Unterscheidung von positiven und negativen Bewertungen, werden sie in entgegen gesetzter Richtung angetragen. Spider Charts sind ein beliebtes Instrument in der Akzeptanzforschung [AH04]. Sie stellen die Produkteigenschaften kreisförmig angeordnet dar. Die normierten Summen der positiven und negativen Bewertungen werden auf den Achsen der Eigenschaften als Punkte eingezeichnet und durch Linien verbunden. Dadurch ergeben sich zwei Flächen der positiven und negativen Bewertungen. Die übereinander liegenden Flächen ermöglichen einen guten Vergleich der positiven und negativen Bewertungen. Während Balkendiagramme besonders zum Vergleich von zwei Produkten geeignet sind, stellen Spider Charts die Summe der positiven und negativen Eigenschaftsbewertungen eines Produktes sehr anschaulich dar. Abbildung 6 stellt zwei Digitalkameramodelle X und Y mithilfe eines Balkendiagramms gegenüber und zeigt ein Spider Chart für das Digitalkameramodell X.

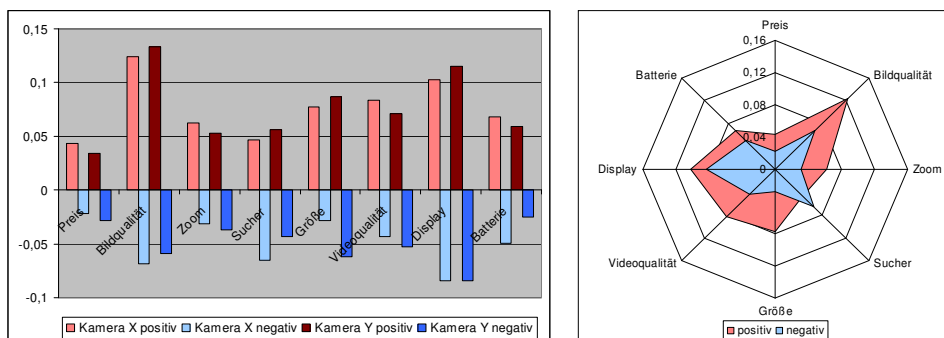


Abbildung 5: Grafische Darstellung der Zusammenfassung

3.4 Analyse

3.4.1 Ermittlung von Abhängigkeiten

Die Aufdeckung der Abhängigkeiten dient dazu Interdependenzen von Eigenschaftsbewertungen zu ermitteln. Sie gibt Aufschluss darüber, welche Produkteigenschaften und ihre Bewertungen oft gemeinsam wahrgenommen werden und sich bedingen. Damit stellt sie einen Ausgangspunkt für die Produktgestaltung und -verbesserung dar.

Zur Ermittlung dieser Gruppen sind die Assoziationsregeln des Tertius-Systems [FL01] gut geeignet. Sie geben an, wie Produkteigenschaften unter bestimmten Bedingungen in einem Review bewertet werden. Der Algorithmus erhält zu jedem Review alle auftretenden Produkteigenschaften mit den zugehörigen Bewertungen und erzeugt daraus Assoziationsregeln.

Die Regeln werden mittels eines unüberwachten Lernalgorithmus erstellt. Ziel des Algorithmus ist es, Regeln zu finden, die einen möglichst guten Kompromiss aus Neuheit und Erfüllung ermöglichen. Zu diesem Zweck wird ein heuristisches Konfirmationsmaß eingesetzt, das auf einer modifizierten X^2 -Statistik beruht. Das Konfirmationsmaß K misst dabei zum einen die Unabhängigkeit von Regelantezendenz und Regelkonsequenz (Neuheit) und zum anderen die Anzahl der Gegenbeispiele, bei denen der Regelantezendenz nicht zutrifft, die Regelkonsequenz aber dennoch gilt (Erfüllung) [FL01, S. 68]. Es ist auf den Bereich von -1 (minimale Konfirmation) bis zu $+1$ (maximale Konfirmation) normiert. Mithilfe einer optimalen Best-First-Suchstrategie werden die gemäß dem Konfirmationsmaß besten k Regeln gefunden [FL01, S. 61].

Die Eingabedaten des Regel-Lernalgorithmus werden auf Basis der Daten aus der Verknüpfungsphase generiert. Für jede Produkteigenschaft eines Reviews wird die Summe der positiven, negativen und neutralen Bewertungen ermittelt. Die Produkteigenschaft erhält diejenige Bewertung deren Summe am höchsten ist. Bei gleicher Anzahl wird die Bewertung als neutral eingestuft. Für 50 zufällig ausgewählte Reviews des Digitalkameramodells X können u. a. folgende Regeln erstellt werden:

Wenn Videoqualität = negativ oder Autofokus = negativ dann Bildqualität = negativ ($K=0,38$)

Wenn Einstellungen = positiv oder Autofokus = positiv dann Bildqualität = positiv ($K=0,36$)

Abbildung 6: Regeln zur Analyse der Abhängigkeiten von Eigenschaftsbewertungen

Die Regeln zeigen, dass die Bildqualität mit einer Konfirmation von 0,38 negativ beurteilt wird, wenn die Videoqualität oder der Autofokus negativ bewertet werden. Eine positive Einschätzung der Bildqualität geht häufig ($K=0,36$) mit einer positiven Beurteilung der Einstellungsmöglichkeiten oder des Autofokus einher.

3.4.2 Ermittlung von Hauptdeterminanten

Zielsetzung der Hauptdeterminantenanalyse ist es zu bestimmen, welche Eigenschaftsbewertungen den größten Einfluss auf die gesamte Beurteilung eines Produktes haben. Sie gibt Hinweise darauf, welche Eigenschaftsbewertungen zu einer insgesamt guten oder schlechten Bewertung des Produktes führen und stellt Ansatzpunkte für die Produktentwicklung und -verbesserung dar.

Zur Bestimmung der Hauptdeterminanten kommen ID3-Entscheidungs­bäume [Qu86] zum Einsatz. Entscheidungs­bäume dienen zur Klassifizierung von Objekten (Produkt­Reviews) anhand ihrer Merkmale. Als Eingabe erhält der Algorithmus zur Induktion des Entscheidungsbaums eine Menge von Reviews eines Produktes, bestehend aus Produkteigenschaftsbewertungen (Merkmale) und Gesamtbewertung (Klasse). Die Ausgabe ist ein Entscheidungsbaum, der aufzeigt, welche Produkteigenschaftsbewertungen zu welcher Gesamtbewertung führen. Die Induktion von Entscheidungs­bäumen [RN03, S. 655ff] erfolgt rekursiv top-down. In jedem Knoten wird zunächst die Produkteigenschaft mithilfe eines Maßes ausgewählt, das die Klassifikation am stärksten verbessert. Anschließend erhält der Knoten entsprechend der Bewertungen der Produkteigenschaft Kindknoten, die die Menge der zugehörigen Reviews erben. Die Rekursion endet, wenn alle Reviews eines Knotens die gleiche Gesamtbewertung haben oder alle Eigenschaften bereits verwendet sind. Der Blattknoten wird mit derjenigen Gesamtbewertung beschriftet, der alle bzw. die meisten seiner Reviews angehören. Zur Eigenschaftsauswahl wird der Informationsgewinn [Qu86] eingesetzt. Er besagt, wie viel Information man über die Gesamtbewertung gewinnt, wenn die Bewertung einer bestimmten Eigenschaft bekannt ist.

Die Eingabedaten des Induktionsalgorithmus werden auf Basis der Daten aus der Verknüpfungsphase erstellt. Die Ermittlung der Merkmale erfolgt analog zur Generierung der Eingabedaten der Abhängigkeitsanalyse. Die Gesamtbewertung wird auf der Grundlage der Anzahl der Anzahl der Sterne, die ein Review-Verfasser vergeben hat, eingeschätzt. Während Reviews mit fünf Sternen als insgesamt gut bewertet werden, erhalten Reviews mit drei Sternen eine mittelmäßige Gesamtbewertung und Reviews mit einem Stern eine schlechte Gesamtbewertung. Reviews mit zwei oder vier Sternen werden in Abhängigkeit der Anzahl der positiven und negativen Bewertungen der Klasse gute, mittelmäßige oder schlechte Gesamtbewertung zugeordnet. Für 50 zufällig ausgewählte Reviews des Digitalkameramodells X ergibt sich folgender Entscheidungsbaum.

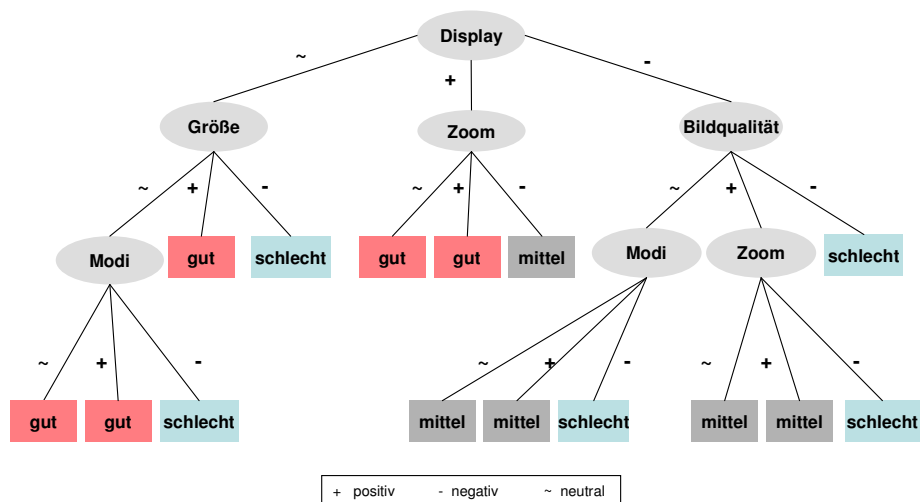


Abbildung 7: Entscheidungsbaum zur Analyse der Hauptdeterminanten

Der Entscheidungsbaum zeigt zum Beispiel, dass die Digitalkamera X in einem Review insgesamt gut eingestuft wird, wenn das LCD-Display und der Zoom positiv bewertet werden. Sie wird jedoch negativ beurteilt, wenn das LCD-Display und die Bildqualität negativ empfunden werden.

4 Zusammenfassung und Ausblick

Im Zuge gestiegener Kundenanforderungen und härterer Wettbewerbsbedingungen bieten immer mehr Unternehmen Problemlösungen an, die auf die Bedürfnisse der Kunden angepasst sind. Dies setzt voraus, dass die Kundenwünsche bekannt sind. Das Web 2.0 stellt eine wertvolle Ressource an Kundenmeinungen in Form von Produkt-Reviews bereit. Eine Analyse lässt sich manuell nur in sehr begrenztem Umfang durchführen. Daher ist eine automatisierte Analyse nötig. Ausgehend von bestehenden Produkt-Mining-Systemen und ihren Defiziten wurde ein automatischer Ansatz zur Analyse von Produkt-Reviews entwickelt, der die Extraktion, Aggregation und Analyse von Produkteigenschaften und ihren Bewertungen ermöglicht. Hierbei kommen Methoden des Text Mining und Data Mining zum Einsatz. Mithilfe der Stützvektormethode werden die Produkteigenschaften und Bewertungen extrahiert und mittels Assoziationsregeln und Entscheidungsbäumen weitergehend analysiert. Dadurch können nicht nur die Produkteigenschaftsbewertungen eigener und konkurrierender Produkte erkannt und miteinander verglichen werden, sondern auch die gegenseitigen Abhängigkeiten zwischen Produkteigenschaftsbewertungen und der Einfluss der Produkteigenschaftsbewertungen auf die Gesamtbewertung des Produktes aufgedeckt werden. Somit lassen sich wichtige Ansatzpunkte für die Produktentwicklung und -verbesserung gewinnen. Der Ansatz stellt ein Basisverfahren zur Analyse von Produkt-Reviews dar, der sich noch durch den Einsatz zusätzlicher Methoden erweitern lässt. Im weiteren Projektverlauf sollen Methoden zur Erkennung impliziter Produkteigenschaften erforscht und größere Trainingsmengen an Reviews zur Verbesserung des Extraktionsergebnisses verwendet werden. Darüber hinaus ist es geplant statistische Methoden, z. B. die Cluster- oder Faktorenanalyse, auf ihr Potential zum Einsatz in der Analysephase zu untersuchen. Ferner wird die Anwendbarkeit des Verfahrens auf andere Produktkategorien und Review-Arten überprüft.

Literaturverzeichnis

- [AH04] Amberg, M.; Hirschmeier, M.; Wehrmann, J.: The Compass Acceptance Model for the Analysis and Evaluation of Mobile Services. In: International Journal for Mobile Communications (IJMC) 2(3), 2004.
- [Ba00] Baluja, S.; Mittal, V. O.; Sukthankar, R.: Applying Machine Learning for High Performance Named-Entity Extraction. In: Computational Intelligence 16 (4), 2000. S. 586-595.
- [CV95] Cortes, C.; Vapnik, V. N.: Support Vector Networks. In: Maschine Learning, 20, 1995. S. 273-297.
- [Da03] Dave, K.; Lawrence, S.; Pennock, D. M.: Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In: Proceedings of the Twelfth International Conference on World Wide Web. ACM Press, Budapest, 2003, S. 519-528.

- [Et05] Etzioni, O; Cafarella, M.; Downey, D.; Popescu, A.M.; Shaked, T.; Soderland, S.; Weld, D.S.; Yates, A.: Unsupervised Named-Entity Extraction from the Web: An Experimental Study. In: *Artificial Intelligence* 165 (1), 2005, S. 191-134.
- [FL01] Flach, P. A.; Lachiche, N.: Confirmation-Guided Discovery of First-Order Rules with Tertius. In: *Maschine Learning* 42, 2001, S. 61-95.
- [HB07] Holzmüller, H. H.; Buber, R.: Optionen für die Marketingforschung durch die Nutzung qualitativer Methodologie und Methoden. In: Holzmüller, H. H.; Buber, R. (Hrsg.): *Qualitative Marktforschung*. Gabler Verlag, Wiesbaden, 2007, S. 3-20.
- [HW00] Hatzivassiloglou, V.; Wiebe, J. M.: Effects of adjective orientation and gradability on sentence subjectivity. In: *Proceedings of the 18th International Conference on Computational Linguistics, ACL*, New Brunswick, 2000. S. 299-305.
- [Ke94] Kepper, G.: *Qualitative Marktforschung. Methoden, Einsatzmöglichkeiten und Beurteilungskriterien*. Deutscher Universitäts-Verlag, Wiesbaden, 1994.
- [Ko96] Koch, J.: *Marktforschung: Begriffe und Methoden*. Oldenbourg Verlag, München, 1996.
- [KM00] Kudoh, T.; Matsumoto, Y.: Use of Support Vector Learning for Chunk Identification. In: *Proceedings of CoNLL-2000 and LLL-2000*, Lisbon, 2000; S. 142-144.
- [KP07] Kao, A.; Poteet, S.: Overview. In: Kao, A.; Poteet, S. R. (Hrsg.): *Natural Language Processing and Text Mining*. Springer Verlag, London, 2007. S. 1-7.
- [Li03] Liu, H.; Liebermann, H.; Selker, T.: A model of textual affect sensing using real-word knowledge. In: *Proceedings of the 17th International Conference on Intelligent User Interfaces*, 2003, S. 125-132.
- [Li05] Liu, B.; Hu, M.; Cheng, J.: Opinion Observer: Analyzing and Comparing Opinions on the Web. In: *Proceedings of the 14th international conference on World Wide Web*. ACM Press, New York, 2005. S. 342-351.
- [Mo00] Morik, K; Worbel, S.; Joachims, T.: *Maschinelles Lernen und Data Mining*. In: Görz, G.; Rollinger, C; Schneeberger, J. (Hrsg.): *Handbuch der Künstlichen Intelligenz*. 3. Aufl. Oldenbourg Verlag, München, 2000, S. 517-598.
- [Pa02] Pang, B.; Lee, L.; Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, 2002, S. 79–86.
- [Pi03] Piller, F. T.: *Mass Customization – Ein wettbewerbsstrategisches Konzept im Informationszeitalter*. 3. Aufl. Deutscher Universitätsverlag, Wiesbaden, 2003.
- [Qu86] Quinlan, J. R.: Induction of Decision Trees. In: *Maschine Learning* 1, 1986, S. 81-106.
- [PE07] Popescu, A. M.; Etzioni, O.: Extracting Product Features and Opinions from Reviews. In: Kao, A.; Poteet, S. R. (Hrsg.): *Natural Language Processing and Text Mining*. Springer Verlag, London, 2007. S. 9-28.
- [PL04] Pang, B.; Lee, L.: A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: *Proceedings of ACL*, 2004, S. 271–278.
- [RN03] Russel, S.; Norwig, P.: *Artificial Intelligence. A Modern Approach*. 2. Aufl. Pearson Education, Upper Saddle River, 2003.
- [SH01] Subasic, P.; Huettner, A.: Affect analysis of text using fuzzy semantic typing. In: *IEEE-FS 9*, 2001. S. 483–496.
- [Sh06] Shanahan, J. G.; Qu, Y.; Wiebe, J.: *Computing Attitude and Affect in Text: Theory and Applications*. Springer Verlag, Dordrecht, 2006.
- [Tu02] Turney, P. D.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, 2002, S. 417-424.
- [We05] Weiss, S. M.; Indurkha, N.; Zhang, T.; Damerau, F. J.: *Text Mining – Predictive Methods for Analyzing Unstructured Information*. Springer Verlag, New York, 2005.
- [WW04] Wilson, T.; Wiebe, J.; Hwa, R.: Just How Mad Are You? Finding Strong and Weak Opinion Clauses. In: *Proceedings of the 19th National Conference on Artificial Intelligence*, 2004, S. 761-769.