STATUS QUO DER TEXTANALYSE IM RAHMEN DER BUSINESS INTELLIGENCE

Autoren

Andreas Schieber, Andreas Hilbert
Technische Universität Dresden, Fakultät für Wirtschaftswissenschaften,
Lehrstuhl für Wirtschaftsinformatik | Business Intelligence Research
E-Mail: andreas.schieber|andreas.hilbert@tu-dresden.de

Zusammenfassung

Vor dem Hintergrund der Zunahme unstrukturierter Daten für Unternehmen befasst sich dieser Beitrag mit den Möglichkeiten, die durch den Einsatz der Business Intelligence für Unternehmen bestehen, wenn durch gezielte Analyse die Bedeutung dieser Daten erfasst, gefiltert und ausgewertet werden können. Allgemein ist das Ziel der Business Intelligence die Unterstützung von Entscheidungen, die im Unternehmen (auf Basis strukturierter Daten) getroffen werden. Die zusätzliche Auswertung von unstrukturierten Daten, d.h. unternehmensinternen Dokumenten oder Texten aus dem Web 2.0, führt zu einer Vergrößerung des Potenzials und dient der Erweiterung des Geschäftsverständnisses der Verbesserung der Entscheidungsfindung. Der Beitrag erläutert dabei nicht nur Konzepte und Verfahren, die diese Analysen ermöglichen, sondern zeigt auch Fallbeispiele zur Demonstration ihrer Nützlichkeit.

Keywords

Business Intelligence, Data Mining, Text Mining, Computerlinguistik, Customer Relationship Management

INHALTSVERZEICHNIS

1 EINFÜHRUNG

Seit vielen Jahren unterstützt die Business Intelligence (BI) Manager und Analysten dabei, mit Hilfe analytischer Informationssysteme fundierte Entscheidungen zu treffen (vgl. CHAUDHURI ET AL. (2011), S. 88; Fachgruppe Business Intelligence (2011), S. 5). Die Verfahren der BI sammeln, verarbeiten, konsolidieren und untersuchen relevante Daten eines Unternehmens und generieren daraus Informationen, die in den jeweiligen Entscheidungssituationen genutzt werden (vgl. GLUCHOWSKI ET AL. (2008), S. 90; KEMPER ET AL. (2006), S. 8).

In klassischen BI-Systemen stammen die entscheidungsrelevanten Daten einerseits aus bestehenden, operativen Informationssystemen des Unternehmens, andererseits auch aus externen Datenquellen (vgl. CHAUDHURI ET AL. (2011), S. 89f.; KEMPER ET AL. (2006), S. 10f.); in beiden Fällen handelt es sich dabei um strukturierte Daten, wie sie üblicherweise in relationalen Datenbanken vorkommen (vgl. BAARS & KEMPER (2008), S. 132). Diese strukturierten Daten können direkt verarbeitet und von Analyseverfahren ausgewertet werden. In Tabellen, Diagrammen und Grafiken lassen sich im Rahmen des Reporting z.B. finanzielle Kennzahlen aufbereiten, um das Rechnungswesen mit konsolidierten Informationen zu versorgen (vgl. BAARS & KEMPER (2008), S. 132; KEMPER ET AL. (2006), S. 110). Komplexere Analyseverfahren erkennen Muster in vergangenheitsbezogenen Daten und können daraus z.B. auf das zukünftige Verhalten von Kunden schließen – solche Verfahren werden u.a. dazu eingesetzt, Kündigungswahrscheinlichkeiten von Kunden in der Telekommunikationsbranche zu berechnen (vgl. CHAUDHURI ET AL. (2011), S. 89 und S. 97).

Schätzungen in der Literatur gehen jedoch davon aus, dass die große Mehrheit der Daten im Unternehmen, d.h. ca. 80%, nicht in strukturierter Form vorliegt (vgl. FELDEN ET AL. (2006), S. 1). Unstrukturierte Daten wie Dokumente, E-Mails usw. enthalten jedoch ebenfalls wichtige Informationen, die Entscheidungen beeinflussen können (vgl. HALPER (2013), S. 29; HEYER ET AL. (2006), S. 1, RUSSOM (2007), S. 1). Ein aussagekräftiges Beispiel liefern CHAUDHURI ET AL. (2011), S. 98 mit der Betrachtung einer Umfrage: Diese enthält zwar einerseits strukturierte Informationen (z.B. eine Skala von 1-5 als Antwortmöglichkeit auf eine Frage), aber ebenso freie Textfelder, in denen der Be-

fragte seine Antwort mit eigenen Worten formulieren kann; diese Felder enthalten oft wertvolle Informationen, die das Unternehmen – bspw. im Rahmen der Produktentwicklung – voranbringen. Diese Situation verstärkt sich noch mit der Entwicklung des World Wide Web zum Web 2.0, da nun Kunden ihre Erfahrungen mit Produkten und Dienstleistungen eines Unternehmens veröffentlichen und so anderen Nutzern zur Verfügung stellen (vgl. KAISER (2009), S. 90). Diese Erfahrungen können ebenfalls ausgewertet und für die marktorientierte Produktentwicklung herangezogen werden. Viele Autoren betonen deshalb die Wichtigkeit der Auswertung unstrukturierter Daten sowie deren Integration in BI-Systeme, denn durch die Kombination der Analyse von strukturierten und unstrukturierten Daten kann ein erweitertes Geschäftsverständnis realisiert werden (vgl. BAARS & KEMPER (2008), S. 133; GLUCHOWSKI ET AL. (2008), S. 326ff.; HIPPNER & RENTZMANN (2006b), S. 100). Anwendungsfelder lassen sich vor allem im Rahmen des Customer Relationship Management (CRM), bei der Wettbewerbsanalyse oder der Produktentwicklung identifizieren (vgl. BAARS & KEMPER (2008), S. 142ff.; CHAUDHURI ET AL. (2011), S. 98; THORLEUCHTER ET AL. (2010), S. 440ff.).

Vor diesem Hintergrund untersucht der vorliegende Beitrag auf Basis des aktuellen Forschungsstandes die Potenziale der Textanalyse im Rahmen der BI. In diesem Kontext ist unter der Bezeichnung Text Mining ein Forschungsfeld entstanden, das Verfahren und Erkenntnisse aus anderen Disziplinen kombiniert, um unstrukturierte Massendaten zu verarbeiten (vgl. MEHLER & WOLFF (2005), S. 5). Der Beitrag betrachtet in Abschnitt 2 zunächst das grundlegende Zusammenspiel der Komponenten eines BI-Systems und nimmt darauf aufbauend die Einordnung von Text Mining vor (Abschnitte 2.2 und 2.3). Anschließend wird das Forschungsfeld selbst vorgestellt, indem zunächst verdeutlicht wird, welche Wissenschaftsdisziplinen ihre Erkenntnisse im Rahmen von Textanalyseverfahren und -konzepten einbringen (Abschnitt 3.1); daran anknüpfend definiert Abschnitt 3.2 den Begriff Text Mining. Abschnitt 3.3 stellt im Anschluss ein Prozessmodell aus der Literatur vor, das anhand von Phasen und Aktivitäten darstellt, wie bei Text-Mining-Projekten vorzugehen ist (Abschnitt 3.3). Abschnitt 4 zeigt abschließend an zwei Fallbeispielen auf, wie Informationen aus Texten zur Entscheidungsunterstützung verwendet werden können.

2 BUSINESS INTELLIGENCE

Die BI als Disziplin der Wirtschaftsinformatik beschäftigt sich mit der betrieblichen Entscheidungsunterstützung (vgl. GLUCHOWSKI ET AL. (2008), S. 90; KEMPER ET AL. (2006), S. 8). Dazu werden relevante Daten des Unternehmens gesammelt, in einer zentralen Datenbank integriert und anwendungsspezifisch ausgewertet. Auf Basis dieser Auswertungen können sowohl strategische als auch operative Entscheidungen im betrieblichen Umfeld unterstützt werden (Fachgruppe Business Intelligence (2011), S. 4). Die nächsten Abschnitte legen eine konkrete Definition des Begriffs fest und erläutern vor allem die analyseorientierte Sichtweise auf die BI, der im Anschluss der Forschungsbereich Text Mining zugeordnet wird.

2.1 DEFINITION

Zur Definition von BI finden sich in der Literatur unterschiedliche Angaben, die von den Autoren der Fachgruppe Business Intelligence (2011), S. 2f. zusammengefasst werden. Die identifizierten Beschreibungen können verschiedenen Gruppen zugeordnet werden. Die einen bezeichnen mit BI eine IT-Architektur bzw. IT-Systeme (vgl. MOSS & ATRE (2003), S. 4; NEGASH (2004), S. 178), andere sehen darin einen Sammelbegriff für Technologien und Konzepte entscheidungsunterstützender Systeme (vgl. GLUCHOWSKI ET AL. (2008), S. 91; KEMPER ET AL. (2006), S. 8). Weitere Autoren wiederum fokussieren sich in ihren Definitionen auf den analytischen Charakter der BI und heben die Ableitung neuen Wissens aus vorhandenen Informationen hervor (vgl. CHAUDHURI ET AL. (2011), S. 88; GOLFARELLI ET AL. (2004), S. 1).

Trotz der verschiedenen Sichtweisen auf den Begriff der BI wird als Zweck aber einheitlich die Unterstützung von Entscheidungen gesehen. Die Autoren der Fachgruppe verstehen unter BI daher „Informationssysteme für alle Phasen betrieblicher Entscheidungsprozesse“ (Fachgruppe Business Intelligence (2011), S. 5) und subsumieren da-

runter sowohl Architektur- als auch Konzept- bzw. Technologieaspekte. Dieser Definition wird im Rahmen des vorliegenden Beitrags gefolgt.

2.2 ORDNUNGSRAHMEN

Die deutschsprachige BI-Forschung wird bis heute stark von Kemper und Gluchowski geprägt, die unter dem Begriff BI sämtliche Technologien und Konzepte zur Entscheidungsunterstützung verstehen. KEMPER & UNGER (2002), S. 665 haben daher einen Ordnungsrahmen entwickelt, der die unterschiedlichen Komponenten von BI-Systemen und der BI-Architektur zusammenfasst (siehe Abbildung 1).

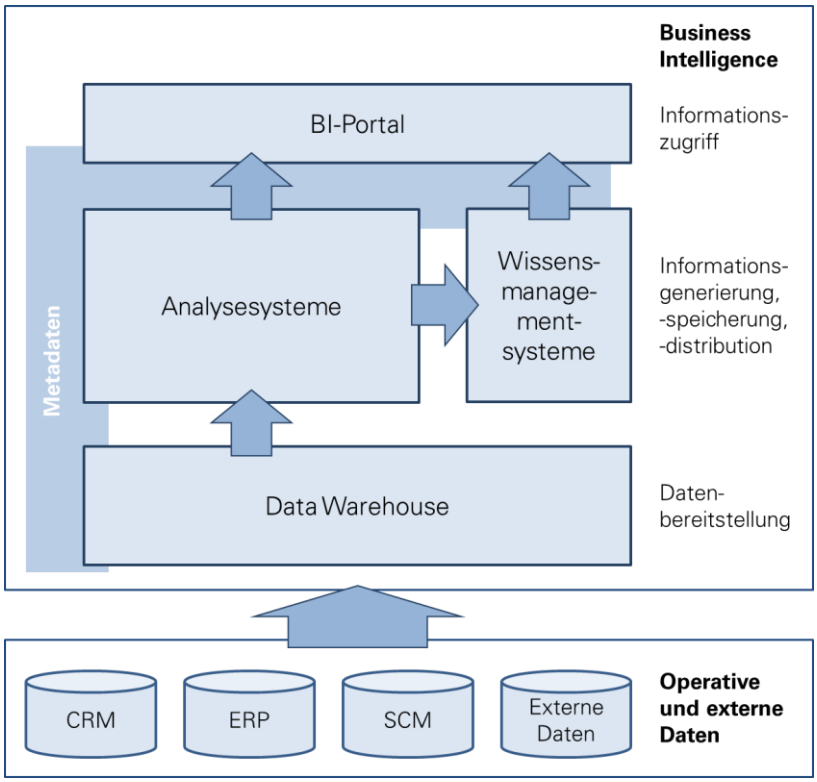

Abbildung 1: Ordnungsrahmen nach KEMPER & UNGER (2002), S. 665f.

Nach KEMPER ET AL. (2006), S. 10ff. beziehen BI-Systeme ihre Daten aus operativen Datenbanksystemen, in denen Informationen zu Prozessen, Abläufen usw. gespei-

chert werden. Diese Informationen werden mit Hilfe eines ETL[1]-Prozesses in ein zentrales Data Warehouse (DWH) übertragen, das die Daten für Analysesysteme bereitstellt. Die Analysesysteme generieren aus diesen Daten neue Informationen, indem sie in Kennzahlensystemen umgerechnet oder mit modellgestützten Verfahren untersucht werden. Über ein Portal kann der Nutzer das Ergebnis der Berechnungen einsehen und bei seinen Entscheidungen berücksichtigen.

Unter Verwendung von geeigneten Konzepten und Technologien lassen sich mit Hilfe des Ordnungsrahmens unternehmensspezifische BI-Systeme erstellen (vgl. KEMPER ET AL. (2006), S. 10). Diesem Beitrag dient der Rahmen später zur Positionierung und Einordnung von Text Mining.

2.3 ANALYSEORIENTIERTE BI UND DATA MINING

In Ergänzung zum Ordnungsrahmen aus Abschnitt 2.2 unterscheidet GLUCHOWSKI (2001), S. 7 drei Sichtweisen auf die BI, um Definitionen und Technologien voneinander abzugrenzen (siehe Abbildung 2).

Neben dem engen und dem weiten BI-Verständnis[2] existiert demnach die analyseorientierte BI, bei der vorhandene Daten mit Hilfe von speziellen Anwendungen gezielt ausgewertet werden, um neue Informationen zu generieren (vgl. GLUCHOWSKI ET AL. (2008), S. 90). Diese Anwendungen lassen sich den Analysesystemen der Informationsbereitstellungsschicht im Ordnungsrahmen von KEMPER & UNGER (2002) zuordnen.

[1] Die Abkürzung ETL steht für die Schritte Extraktion, Transformation und Laden; der ETL-Prozess extrahiert die Quelldaten aus den Vorsystemen, transformiert diese in ein einheitliches Format und lädt sie in das DWH (vgl. CHAUDHURI ET AL. (2011), S. 90).

[2] Gluchowski unterscheidet neben dem analyseorientieren BI-Verständnis auch eine enge bzw. eine weite Sichtweise: BI i.e.S. umfasst demnach nur Anwendungen für multidimensionale Auswertungen bzw. Darstellungen, z.B. Online Analytical Processing (OLAP); dagegen zählen zur BI i.w.S. sämtliche Komponenten, die Daten sammeln, aufbereiten, auswerten und präsentieren können, z.B. das DWH oder die ETL-Prozesse (vgl. GLUCHOWSKI ET AL. (2008), S. 90f.; GLUCHOWSKI (2001), S. 6ff.).

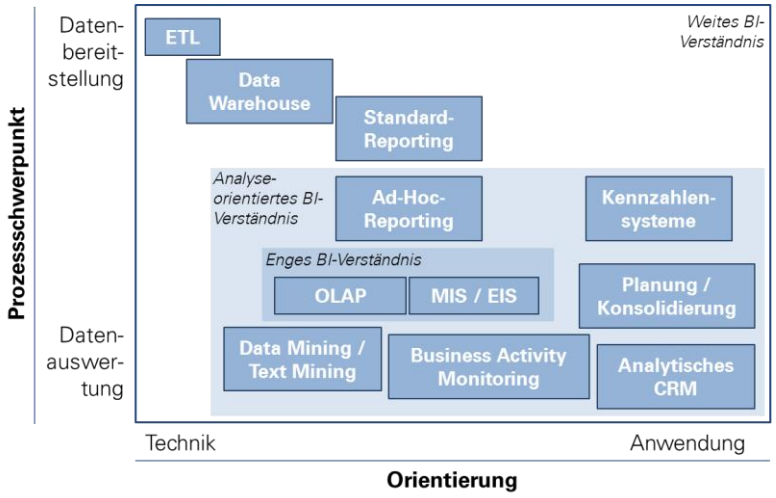

Abbildung 2: Abgrenzung des BI-Verständnisses in Anlehnung an GLUCHOWSKI ET AL. (2008), S. 92 und GLUCHOWSKI (2001), S. 7

Leistungsfähige BI-Anwendungen für diese Schicht sind u.a. Data-Mining-Systeme, die mit modellgestützten Analyseverfahren komplexe Auswertungen ermöglichen (vgl. KEMPER ET AL. (2006), S. 102). Auch wenn der Begriff Data Mining in der Literatur oft mit dem Prozess der Wissensgenerierung gleichgesetzt wird (vgl. KEMPER ET AL. (2006), S. 106), ordnet FAYYAD (1996), S. 9 Data Mining als Schritt in diesen übergeordneten Prozess des Knowledge Discovery in Databases (KDD) ein: „Data Mining is a step in the KDD process consisting of particular data mining algorithms that […] produces a particular enumeration of patterns". Ziel von Data-Mining-Verfahren ist demnach, Muster – wie z.B. Regelmäßigkeiten und Auffälligkeiten – in den zugrundeliegenden Daten zu entdecken und dadurch Strukturzusammenhänge abzubilden (vgl. BAUER & GÜNZEL (2004), S. 109; CHAUDHURI ET AL. (2011), S. 90; GLUCHOWSKI ET AL. (2008), S. 191). Den übergeordneten Prozess zur Entdeckung dieser Muster beschreibt FAYYAD (1996), S. 6 wie folgt: „Knowledge Discovery in Databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data". Ziel des KDD-Prozesses ist demnach die Ableitung neuen Wissens, das sinnvoll – d.h. entscheidungsunterstützend – genutzt werden kann. Abbildung 3 zeigt ein generisches Vorgehensmodell des beschriebenen Prozesses, der sich in mehrere, aufeinander folgende Schritte aufteilt (vgl. FAYYAD (1996), S. 6; KURGAN & MUSILEK (2006), S. 9ff.).

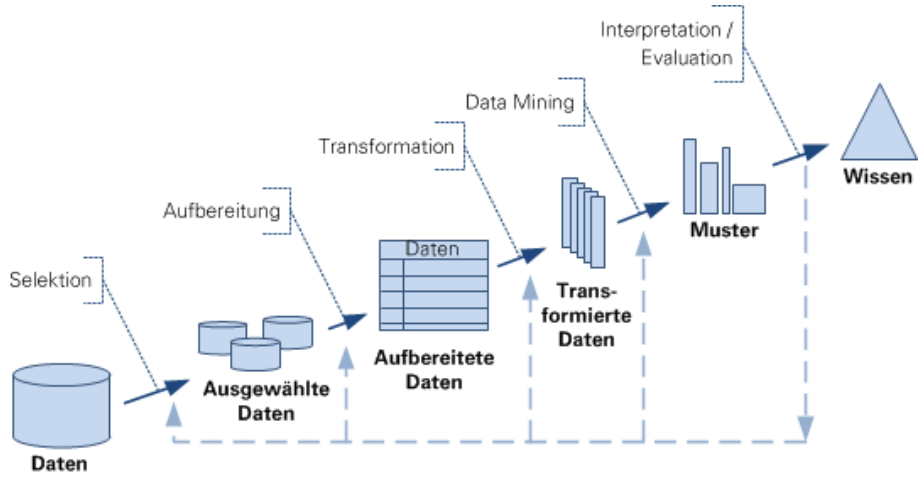

Abbildung 3: Der KDD-Prozess in Anlehnung an FAYYAD (1996), S. 6

Im ersten Schritt werden die Daten ausgewählt, die für die Analyse relevant sind. Die Grundvoraussetzung zur Anwendung der Data-Mining-Verfahren ist dabei eine strukturierte Datengrundlage auf Basis einer Datenbank oder eines Data Warehouse (vgl. BAARS & KEMPER (2008), S. 132; HAN & KAMBER (2006), S. 10; VOSSEN (2008), S. 528). Die Daten aus diesen Systemen sind anschließend gegebenenfalls aufzubereiten, bspw. bei Unterschieden in der Datenstruktur. Im nächsten Schritt findet eine (verfahrensspezifische) Vorverarbeitung der selektierten Daten statt, indem z.B. aus den bestehenden Attributen neue Variablen gebildet werden. Im Anschluss werden Data-Mining-Verfahren auf die vorbereitete Datenbasis angewendet, die als Ergebnis die Beziehungszusammenhänge zwischen den Daten aufdecken. Nach Evaluation und Interpretation der Resultate gehen diese in Wissen über (vgl. FAYYAD (1996), S. 10f.; KURGAN & MUSILEK (2006), S. 9ff.; PETERSOHN (2005), S. 11ff.). Zur Aufdeckung dieser Beziehungszusammenhänge im Teilschritt Data Mining stehen unterschiedliche Methoden zur Verfügung, die sich in voraussagende und beschreibende Verfahren klassifizieren lassen (vgl. BAUER & GÜNZEL (2004), S. 109; KEMPER ET AL. (2006), S. 108; VOSSEN (2008), S. 527):

- Zu den voraussagenden Verfahren gehören unter anderem die Klassifikation und die Regression. Während die zu untersuchenden Daten bei der Klassifikation (z.B. mit Hilfe eines Entscheidungsbaums) in vorab definierte Klassen eingeordnet werden, beschreibt die Regression Ursache-Wirkungs-Zusammenhänge zwischen einzelnen Merkmalen der zugrundeliegenden Da-

ten. Solche Verfahren werden z.B. bei der Berechnung der Kreditwürdigkeit eingesetzt (vgl. GLUCHOWSKI (2001), S. 9).

- Die Clusterbildung und die Assoziationsanalyse sind dagegen den beschreibenden Verfahren zuzuordnen. Bei der Clusterbildung werden die Daten zu ähnlichen, vorher unbekannten Gruppen zusammengefasst; die Assoziationsanalyse hingegen deckt Beziehungen zwischen den Ausprägungen der Variablen auf und bildet sie als Regeln ab. Während die Clusterbildung dazu eingesetzt wird, Kundensegmente zu bestimmen, kann die Assoziationsanalyse der Analyse von Warenkörben und damit der Layoutplanung von Supermärkten dienen (vgl. GLUCHOWSKI (2001), S. 9).

Für einen detaillierteren Einblick in die Verfahren des Data Mining sei an dieser Stelle auf entsprechende Literatur wie FAYYAD (1996), HAN & KAMBER (2006), PETERSOHN (2005) oder RUNKLER (2010) verwiesen.

3 TEXT MINING

Für die modellgestützte Auswertung sind relevante Daten jedoch nicht nur in strukturierter Form verfügbar, sondern liegen häufig auch als semi- oder gar unstrukturierte Texte vor (vgl. BAARS & KEMPER (2008), S. 132f.). Auch HEYER ET AL. (2006), S. 1 betonen: „Text ist ein bedeutender Wissensrohstoff, der im Zeitalter des Internet in großen Mengen in digitaler Form zur Verfügung steht". Allerdings können unstrukturierte Daten nicht ohne Weiteres analysiert werden (vgl. HAN & KAMBER (2002), S. 428; HOTHO ET AL. (2005), S. 19). Text-Mining-Werkzeuge sind auf diese Besonderheit der Datengrundlage spezialisiert und erlauben es – analog zu Data-Mining-Verfahren –, aus Texten Informationen und Zusammenhänge zu extrahieren (vgl. HEYER ET AL. (2006), S. 1; WEISS ET AL. (2010), S. 1). Dadurch und durch eine Reihe anderer Funktionen, z.B. das Erkennen inhaltlicher Strukturen oder Ähnlichkeiten zwischen Begriffen, eignet sich Text Mining besonders zur Wissensakquisition aus Texten (vgl. HEYER ET AL. (2006), S. 6f.).

Wie die Nachbardisziplin Data Mining (siehe Abschnitt 2.3) zielt Text Mining darauf ab, Wissen aus vorhandenen Daten zu generieren. Im Gegensatz zu Data-Mining-Verfahren, die auf Basis strukturierter Daten agieren, können Text-Mining-Algorithmen jedoch unstrukturierte Daten verarbeiten. Text-Mining-Tools sind daher ebenfalls zur analyseorientierten BI bzw. zu den Analysesystemen des BI-Ordnungsrahmens zu zählen. Die folgenden Abschnitte befassen sich zunächst mit einer Abgrenzung des Forschungsfeldes im Hinblick auf andere Disziplinen. Im Anschluss wird der Begriff Text Mining definiert und ein Prozessmodell für Text-Mining-Vorhaben dargestellt und erläutert.

3.1 BERÜHRUNGSPUNKTE MIT ANDEREN DISZIPLINEN

Das Forschungsfeld Text Mining kombiniert Techniken zur Verarbeitung und Analyse von Texten, die ihren Ursprung in unterschiedlichen Disziplinen haben (vgl. MINER ET

AL. (2012), S. 30f.). Wie bereits erwähnt, spielt das Forschungsfeld Data Mining in diesem Kontext eine große Rolle: Um Muster in (strukturierten) Massendaten erkennen zu können, werden bspw. ähnliche Datensätze gruppiert; solche Data-Mining-Verfahren werden auch im Text Mining angewendet (vgl. HOTHO ET AL. (2005), S. 30ff.; HUSSAIN ET AL. (2012), S. 9). Die Voraussetzung dafür ist jedoch, dass die zunächst unstrukturierten Textdaten in eine strukturierte Form überführt werden.

Diese Funktion steuert das Natural Language Processing (NLP) bei, ein Teilbereich der Computerlinguistik, der ebenfalls einen großen Beitrag im Rahmen des Text Mining leistet (vgl. MEHLER & WOLFF (2005), S. 9f.; MILLER (2005), S. 106f.). Mit Hilfe von NLP-Verfahren wird eine umfassendere Analyse der Texte ermöglicht, indem grammatikalische Regeln, Thesauren, Lexika und weitere linguistische Konzepte eine Strukturierung des Textes vornehmen: Dadurch können u.a. bestimmte Satzelemente (z.B. Substantive, Adjektive, usw.) identifiziert und gesondert untersucht werden (vgl. HOTHO ET AL. (2005), S. 25ff.; KAO & POTEET (2007), S. 1; WEISS ET AL. (2010), S. 16ff.). Während sich die Computerlinguistik mit der computerbasierten Verarbeitung von Text und Sprache beschäftigt (vgl. KÖHLER (2005), S. 1), nutzt Text Mining diese Erkenntnisse und wendet sie auf unstrukturierte Massendaten an, um neuartige Erkenntnisse aus den Texten zu gewinnen und daraus Handlungsmaßnahmen abzuleiten (vgl. HEARST (1999), S. 4f.; HIPPNER & RENTZMANN (2006a), S. 287; MEHLER & WOLFF (2005), S. 9). MEHLER & WOLFF (2005), S. 9 betonen dabei auch, dass beide Disziplinen voneinander lernen können, indem sie miteinander interagieren: Text-Mining-Ergebnisse können durch die Integration linguistischer Erkenntnisse verbessert werden, und linguistische Erkenntnisse können durch den Einsatz in Text-Mining-Projekten validiert werden.

Wie die Ausführungen zeigen, sind die Computerlinguistik und Data Mining in diesem Bereich die beiden wichtigsten Nachbardisziplinen. Aufgrund der Verarbeitung von großen Datenmengen kommt aber auch der Statistik eine wichtige Rolle zu; außerdem haben einige Verfahren ihre Wurzeln im Feld der Künstlichen Intelligenz bzw. im Information Retrieval (vgl. MINER ET AL. (2012), S. 31).

3.2 DEFINITION

In Abhängigkeit der Perspektive (siehe Abschnitt 3.1) wird auch der Begriff Text Mining auf unterschiedliche Weise betrachtet. HOTHO ET AL. (2005), S. 22f. erläutern drei Auffassungen: Text Mining im Sinne der Informationsextraktion, Text Mining im Sinne von textbasiertem Data Mining und Text Mining im Sinne eines (KDD-)Prozesses.

Als Werkzeug zur Informationsextraktion kann Text Mining eingesetzt werden, um Passagen aus einem Text zu extrahieren und mit bestimmten Attributen zu versehen; bspw. können dadurch (teil-)automatisiert Personen sowie deren jeweilige Funktion in einem Unternehmen identifiziert werden (vgl. HOTHO ET AL. (2005), S. 45ff.).

Die zweite Auffassung sieht den Begriff Text Mining – wie Data Mining – als Bezeichnung für Verfahren zur computergestützten Analyse und (semi-)automatischen Strukturierung von Texten (vgl. HE (2013), S. 501; HEYER ET AL. (2006), S. 3; HOTHO ET AL. (2005), S. 23).

Abschließend wird der Begriff Text Mining von einigen Autoren auch als Prozess zur Wissensgenerierung betrachtet (vgl. FELDMAN & DAGAN (1995), S. 112; HIPPNER & RENTZMANN (2006a), S. 287; HOTHO ET AL. (2005), S. 23). Wie in Abschnitt 2.3 erläutert wurde, wird dieser Prozess im Rahmen des Data Mining KDD genannt; in Analogie dazu prägten FELDMAN & DAGAN (1995), S. 112 den Begriff Knowledge Discovery in Textual Databases (KDT).

Der vorliegende Beitrag folgt der dritten Auffassung und versteht Text Mining als analytischen Prozess zur computergestützten Wissensgenerierung aus Textdaten. Daher fokussiert das Verständnis von Text Mining nicht nur auf die Analyseverfahren selbst, sondern auch auf die vor- und nachgelagerten Schritte des Prozesses wie die Sammlung und Aufbereitung der relevanten Daten sowie die Interpretation und Verwertung der Ergebnisse. Ein Modell dieses Prozesses wird bei HIPPNER & RENTZMANN (2006a), S. 288 diskutiert und im folgenden Abschnitt aufgegriffen.

3.3 PROZESSMODELL NACH HIPPNER & RENTZMANN (2006a)

Wie im vorigen Abschnitt erläutert wurde, verstehen auch HIPPNER & RENTZMANN (2006a), S. 287ff. Text Mining als Prozess zur Datenanalyse. In ihrem Beitrag beschreiben die Autoren ein anwendungsneutrales Vorgehensmodell für Text-Mining-Projekte und gliedern diesen iterativen Prozess in die nachfolgend beschriebenen Schritte (siehe Abbildung 1). Die Aktivitäten vieler Schritte decken sich dabei mit den Ausführungen zu Data Mining in Abschnitt 2.3; in der Phase der Dokumentaufbereitung unterscheiden sich die beiden Disziplinen jedoch besonders.

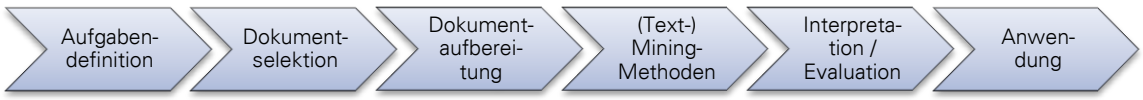

Abbildung 1: Der Text-Mining-Prozess nach HIPPNER & RENTZMANN (2006a), S. 288

3.3.1 Aufgabendefinition

Im ersten Schritt erfolgt die Aufgabendefinition, indem betriebswirtschaftliche Problemstellungen bestimmt und daraus Text-Mining-Ziele abgeleitet werden. Dazu gehören u.a. die Marktforschung (vgl. DAVIS & OBERHOLTZER (2008), S. 1ff.; KAISER (2009), S. 90), die Wettbewerbsanalyse (vgl. BAARS & KEMPER (2008), S. 144ff.; FENG & FUHAI (2012), S. 467ff.) oder die Produktentwicklung (vgl. KAISER (2008), S. 229ff.; THORLEUCHTER ET AL. (2010), S. 440ff.). Dies ist bereits eine wichtige Phase, da sich die gesetzten Ziele auf Aufgaben und Verfahren in den darauffolgenden Phasen – vor allem bei der Dokumentaufbereitung und beim Text Mining selbst – auswirken.

3.3.2 Dokumentselektion

Im Anschluss an die Zieldefinition erfolgt die Identifizierung der potenziell entscheidungsrelevanten Dokumente. HIPPNER & RENTZMANN (2006a), S. 288 beschreiben dabei, dass – analog zu einem DWH in klassischen BI-Systemen – ein Document Wa-

rehouse[3], das verschiedene Daten- und Dokumenttypen beinhaltet, von Nutzen sein kann. Dabei können die vorhandenen Daten aus verschiedenen Quellen zusammengeführt werden. Je nach Anwendungsfall könnten im Document Warehouse bzw. der Datenbank bspw. einerseits unternehmensinterne Texte wie Schriftverkehr mit Kunden, aber andererseits auch unternehmensexterne Daten aus dem Web gespeichert werden; auch eine Integration von internen und externen Daten wäre denkbar.

3.3.3 Dokumentaufbereitung

Aufgrund ihrer Beschaffenheit müssen die unstrukturierten Daten – anders als im Data-Mining-Prozess – gesondert aufbereitet und dadurch in eine strukturierte Form gebracht werden (vgl. HOTHO ET AL. (2005), S. 19). Dieser Schritt erfolgt in der Dokumentaufbereitung und kann aufgrund seiner Bedeutung über den Erfolg des Text-Mining-Projektes entscheiden. Ziel dieser Phase ist die Extraktion von Termen aus den Texten; diese Terme sind Grundlage für die anschließende Analyse und werden mit verschiedenen Techniken aus dem Forschungsfeld des NLP bestimmt (vgl. HIPPNER & RENTZMANN (2006a), S. 288; MILLER (2005), S. 106).

Der erste Ansatz ist die morphologische Analyse, die einzelne Wortformen und Wortbestandteile betrachtet. Dabei werden die Terme in den Texten nicht nur als Zeichenketten betrachtet, sondern es wird versucht, sie als bestimmte, konjugierte Form eines Wortes zu erkennen (vgl. FERBER (2003), S. 40f.). Bei diesem Vorgang wird zwischen der Grundformenreduktion und der Stammformenreduktion unterschieden. Die Grundformenreduktion bzw. Lemmatisierung beschreibt die Zurückführung von einzelnen Wörtern auf ihre Grundform, wie beispielsweise von Substantiven auf den Nominativ Singular und von Verben auf den Infinitiv. Die Stammformenreduktion, auch Stemming genannt, beschreibt hingegen die Zurückführung der einzelnen Wörter auf ihren Wortstamm. Als Beispiel kann die Zeichenkette „fand“ und „Gefundenes“ auf denselben Stamm „finden“ zurückgeführt werden (vgl. FERBER (2003), S. 41); je nach Verfahren können die Terme durch die Wortgruppierung zwar stark reduziert werden,

[3] In einem DWH werden üblicherweise Daten aus unterschiedlichen, operativen Systemen konsolidiert gespeichert (vgl. KEMPER ET AL. (2006), S. 17ff.). Ein Document Warehouse enthält demnach entscheidungsrelevante Dokumente des Unternehmens in analyseorientierter Speicherung (vgl. TSENG & CHOU (2006), S. 728).

jedoch leidet darunter unter Umständen die Interpretierbarkeit der Daten (vgl. NATARAJAN (2005), S. 34; SANJUAN & IBEKWE-SANJUAN (2006), S. 1533).

Eine weitere Technik des NLP ist die syntaktische Analyse, die eine Annotation einzelner Satzbausteine vornimmt. In diesem Zusammenhang wird unter einer Annotation eine Textauszeichnung – im angloamerikanischen Raum auch als Part-of-Speech-(POS)-Tagging bezeichnet – verstanden, die bestimmte Textbestandteile markiert, lexikonbasiert kategorisiert und damit Adjektive, Substantive, Verben und Eigennamen identifiziert (vgl. LOBIN (2004), S. 51). Darüber hinaus wird beim Parsing der Satzbau analysiert und somit jedes Wort entsprechend seiner Stellung im Satz als Subjekt, Objekt usw. untersucht und gekennzeichnet (vgl. CIRAVEGNA & LAVELLI (1999), S. 102ff.). Die syntaktische Analyse birgt großen Nutzen, da dadurch eine gezielte Extraktion von Informationen aus bestimmten syntaktischen Einheiten möglich ist.

Die semantische Analyse ist die dritte Technik des Natural Language Processing, die in Ergänzung kontextuelles Wissen verarbeitet. Dabei wird ein Satz in bedeutungsabhängige Einheiten zerlegt und die einzelnen Wörter dem Kontext entsprechend analysiert (vgl. HOTHO ET AL. (2005), S. 29). Ziel dabei ist zu erkennen, ob beispielsweise das Wort „Bank“ eine Sitzgelegenheit oder ein Geldinstitut bezeichnet.

Die genannten NLP-Verfahren können je nach Aufgabenstellung auch kombiniert werden. Welche der Ansätze zur Anwendung kommen, hängt jedoch vor allem von den Dokumenten und den Analysezielen ab.

Neben den NLP-Verfahren können in dieser Phase weitere Schritte zur Aufbereitung unternommen werden. Dazu zählen einerseits Term-Filtering-Verfahren, mit deren Hilfe die Menge an Termen und damit – wie bei der morphologischen Analyse – die Dimensionalität reduziert wird (vgl. HOTHO ET AL. (2005), S. 25; TSENG ET AL. (2007), S. 1222). Andererseits gehören dazu auch Verfahren, die die Datenbasis transformieren; üblich ist in diesem Kontext die Repräsentation der Dokumente in Form eines Vektorraummodells bzw. einer Term-Dokument-Matrix: In den Zeilen der Matrix sind dabei die Dokumente angeordnet, während in den Spalten die in den Texten vorkommenden Terme gespeichert sind; die Zellen der Matrix enthalten binäre Werte (Term kommt im Dokument vor oder nicht), Häufigkeitszahlen (Term kommt x-mal im Dokument vor) oder andere Kennzahlen zur Abbildung der Beziehung zwischen Dokument und Term (vgl. HIPPNER & RENTZMANN (2006a), S. 289; HOTHO ET AL. (2005), S. 25).

Nachdem mittels der verschiedenen Techniken die Terme des Textes extrahiert wurden, können diese als Variablen für die weitere Analyse verwendet werden. Die Ausführungen machen deutlich, dass an dieser Stelle die Weichen für sinnvoll interpretierbare Analyseresultate gestellt werden.

3.3.4 Text-Mining-Methoden

Im Anschluss an die Dokumentaufbereitung liegen die textuellen Daten in einer strukturierten Form vor, sodass auch Data-Mining-Verfahren angewandt werden können. Dazu zählen sowohl Klassifikationsverfahren, die die Texte in vorgegebene Kategorien einordnen (vgl. HOTHO ET AL. (2005), S. 30ff.; HUSSAIN ET AL. (2012), S. 9), als auch Segmentierungsverfahren, die ähnliche Texte in vorher unbekannte Gruppen zusammenführen (vgl. SOMMER ET AL. (2012), S. 10ff.; SANJUAN & IBEKWE-SANJUAN (2006), S. 1537), als auch Abhängigkeitsanalysen, welche das gemeinsame Auftreten von Termen untersuchen (vgl. DELGADO ET AL. (2002), S. 142ff.; NATARAJAN (2005), S. 36; TSENG ET AL. (2007), S. 1223). Inzwischen sind jedoch textspezifische Analyseverfahren entstanden, die versuchen, den Text auch inhaltlich zu erfassen. Dazu zählen u.a. Verfahren zur Meinungsanalyse (vgl. ARCHAK ET AL. (2011), S. 1490; PANG & LEE (2008), S. 1ff.), zur Zusammenfassung von Texten (vgl. CHOUDHARY ET AL. (2009); SARAVANAN & RAJ (2003), S. 465) oder zur Trendanalyse (vgl. CHOUDHARY ET AL. (2009), S. 731; HEINRICH ET AL. (2012), S. 1145ff.).

3.3.5 Interpretation / Evaluation

Die Resultate der Textanalyse werden anschließend interpretiert und hinsichtlich ihrer Relevanz im Sinne des Analyseziels aus der ersten Phase bewertet. Genügen die Ergebnisse den Anforderungen bzw. der Zielstellung nicht, müssen die vorangegangenen Phasen erneut durchlaufen und die Parametrisierung der eingesetzten Verfahren angepasst werden (vgl. CHOUDHARY ET AL. (2009), S. 730).

3.3.6 Anwendung

Sofern die Evaluation der Ergebnisse zufriedenstellend ausgefallen ist, folgt deren Anwendung: Als Abschluss des Prozesses müssen die Erkenntnisse aus der Analyse in Handlungsempfehlungen oder Maßnahmen umgesetzt werden (vgl. HIPPNER & RENTZMANN (2006a), S. 289); erst dadurch wird die Entscheidungsunterstützungsfunktion der BI erfüllt.

4 POTENZIALE DER TEXTANALYSE

Die Anwendung des Text Mining und die Umsetzung der Resultate sind vor allem in Bereichen sinnvoll, in denen viele Dokumente vorliegen und Wissen eine große Rolle spielt (vgl. HIPPNER & RENTZMANN (2006a), S. 289). Hierzu zählt u.a. das CRM, weil durch die Analyse von Schriftverkehr, digitalisierten Gesprächen oder sonstigen unstrukturierten Daten Kundeninformationen ergänzt oder gar validiert werden können. Inzwischen werden jedoch auch Meinungen aus Kundenrezensionen ausgewertet, um z.B. Verkaufszahlen eines Produktes abzuschätzen. An diesen beiden, ausgewählten Anwendungsfeldern erläutern die folgenden Abschnitte die Potenziale der Textanalyse im Rahmen der BI.

4.1 ERWEITERUNG DES CRM

Das CRM bzw. Kundenbeziehungsmanagement ist vor allem ein Teilgebiet des Marketing und bezeichnet die Ausrichtung des Unternehmens am Kunden (vgl. GNEISER (2010), S. 95f; THOMMEN & ACHLEITNER (2003), S. 123). Bei der Sammlung, Verwaltung und Auswertung von Kundeninformationen wird das CRM von (analytischen) Informationssystemen unterstützt, wodurch sich das Forschungsfeld auch in der Wirtschaftsinformatik etabliert hat (vgl. BAARS & KEMPER (2008), S. 143f.; STAHLKNECHT & HASENKAMP (2005), S. 326). In Verbindung mit der BI liegt der Anwendungsbereich insbesondere beim analytischen CRM (siehe auch Abbildung 2 in Abschnitt 2.3). Gegenstand ist hierbei die Analyse der gespeicherten Daten, um die Kunden besser kennenzulernen und deren Verhalten vorhersagen zu können (vgl. FAYERMAN (2002), S. 64). Data-Mining-Verfahren sind z.B. in der Lage, den Kundenstamm zu segmentieren, d.h. in heterogene Zielgruppen einzuteilen, und dadurch Kundenprofile zu erstellen (vgl. CHANG ET AL. (2009), S. 1433). Weitere Anwendungsgebiete sind die Prognose von Kündigungswahrscheinlichkeiten oder die Ableitung von Cross-Selling-Potenzialen aus der Einkaufshistorie (vgl. HIPPNER & RENTZMANN (2006b), S. 99).

In einem analytischen CRM-System nimmt das mitunter als Customer Data Warehouse bezeichnete DWH die Position der zentralen Datenbank ein, in der die Kundendaten gespeichert sind (vgl. BAARS & KEMPER (2008), S. 143). Diese konsolidierte Datenbasis ist Ausgangspunkt für die anschließende Analyse der Kunden und stellt daher bereits besondere Anforderungen an die BI: Um ein möglichst umfassendes Bild der Kunden erhalten zu können, müssen die Daten meist aus unterschiedlichen Systemen zusammengetragen werden (vgl. BAARS & KEMPER (2008), S. 143). Da gerade in diesem Kontext häufig auch Informationen aus dem Schriftverkehr oder Kundenmeinungen aus Online-Plattformen eine große Rolle spielen, schließt dies auch unstrukturierte Daten mit ein (vgl. BAARS & KEMPER (2008), S. 143f.; HIPPNER & RENTZMANN (2006b), S. 100). Wie in Abschnitt 1 erwähnt wurde, bieten nutzergenerierte Inhalte besondere Potenziale, Informationen über die Kunden des Unternehmens zu sammeln und im Hinblick auf Zielgruppen- oder Profilbildung auszuwerten. BAARS & KEMPER (2008), S. 144 schlagen in diesem Zusammenhang vor, wie die Integration unstrukturierter Daten in das Customer Data Warehouse idealerweise umgesetzt werden sollte und welche Chancen sich dadurch für das Unternehmen ergeben.

HIPPNER & RENTZMANN (2006b), S. 99ff. greifen diese Chancen auf und beschreiben einen konkreten Anwendungsfall in der Bankenbranche. Die Autoren analysierten in ihrem Beitrag die Transaktionsdaten von Bankkunden und dabei insbesondere deren Verwendungszwecke. Neben den Verwendungszwecken, die Freitexte enthalten, stehen dabei auch strukturierte Daten für die Analyse zur Verfügung: Kontoinhaber, Bankverbindung und Betrag (vgl. HIPPNER & RENTZMANN (2006b), S. 100). Durch die Auswertung der Verwendungszwecke in Kombination mit den strukturierten Angaben erhofften sich die Autoren, den Kunden und sein Verhalten besser beschreiben zu können. Dazu wurden im Rahmen des Projekts nur die Daueraufträge der Bankkunden analysiert, um möglichst belastbare Erkenntnisse zu gewinnen.

Bei der Aufbereitung der Texte wurden die identifizierten Terme mit Hilfe von NLP-Verfahren verarbeitet; dazu wurden kontextspezifische Wörterbücher und Ontologien entwickelt, um die Terme auf ihren Wortstamm zurückführen sowie synonym gebrauchte Begriffe bestimmen und zusammenfassen zu können. Anschließend wurde die Datenbasis in ein Vektorraummodell überführt, sodass sich für die Auswertung klassische Data-Mining-Verfahren einsetzen ließen. Im skizzierten Anwendungsfall wurden wichtige Terme extrahiert, hinsichtlich ihrer Aussagekraft zur Beschreibung

des Kunden bewertet und daraus entsprechende Handlungsempfehlungen abgeleitet; davon sollen beispielhaft zwei Ergebnisse genannt werden (vgl. HIPPNER & RENTZMANN (2006b), S. 104f.):

- Aus Termen wie Haushaltsgeld / Taschengeld / Unterhalt schlussfolgern die Autoren u.a. Hinweise auf die Haushaltsstruktur. Weitere Analysen, wie z.B. die Angaben zum Namen des Gegenkontoinhabers lassen demnach Rückschlüsse auf Kinder des Kunden zu; auch die Angabe der Gegenbank kann Handlungsmaßnahmen nach sich ziehen, um mit speziellen Konditionen sämtliche Familienmitglieder als Kunden bei der Bank zu gewinnen.
- Beim Term Bausparvertrag können in Verbindung mit Assoziationsanalysen außerdem zusätzliche Erkenntnisse gewonnen werden. Auch hier ist die Gegenbank eine wichtige Angabe, um Kunden mit maßgeschneiderten Angeboten auf Produkte des eigenen Hauses aufmerksam machen zu können.

Das Fallbeispiel zeigt auf, wie durch die Untersuchung von Texten Kundendaten ergänzt werden können – im Customer Data Warehouse könnte bspw. nach dieser Analyse die Information gespeichert werden, ob der Kunde einen Bausparvertrag bei einer Fremdbank besitzt – und sich daraus konkrete Handlungsempfehlungen für die Kundenansprache entwickeln lassen, z.B. indem solche Kunden mit gezielten Offerten beworben werden.

4.2 ALTERNATIVE ZUR MARKTFORSCHUNG

Ein weiteres Fallbeispiel demonstriert die Potenziale der Textanalyse im Rahmen der Marktforschung. DAVIS & OBERHOLTZER (2008), S. 1f. merken an, dass die Analyse von unternehmensexternen Texten als Ergänzung zur traditionellen Marktforschung dienen kann. Die Autoren beziehen sich darauf, dass immer mehr Kunden ihre Erfahrungen mit Produkten und / oder Dienstleistungen eines Unternehmens im Web 2.0 veröffentlichen und dadurch vielen anderen (vor allem potenziellen) Kunden zur Verfügung stellen (vgl. KAISER (2009), S. 90; MISHNE & GLANCE (2005), S. 1; PANG & LEE (2008), S. 1ff.). Bewertungen in Form von kurzen Texten werden in Onlineshops, z.B. bei Amazon, oder Bewertungsportalen, z.B. bei Ciao.de, von den Konsumenten genutzt und beein-

flussen direkt die Kaufentscheidung, da die Eigenschaften des Produkts vor dem Kauf transparenter werden (vgl. MISHNE & GLANCE (2005), S. 1). LIU (2008), S. 4f. beschreibt unter dem Begriff Opinion Mining – auch Sentiment Analysis genannt (vgl. Archak et al. 2011), S. 1486) – Verfahren, mit deren Hilfe Meinungen aus solchen Bewertungen extrahiert werden können. Dazu müssen einerseits die Produkteigenschaften (sogenannte „features") in den Texten erkannt und andererseits die zugehörige Polarität (d.h. positiv, neutral oder negativ) bestimmt werden (vgl. Archak et al. 2011), S. 1490); dafür eignen sich besonders Adjektive und Adverbien (vgl. BENAMARA ET AL. (2007), S. 1ff.), die mit Hilfe von POS-Tagging-Verfahren identifiziert werden können (siehe Abschnitt 3.3.3). Die Auswertung dieser Texte liefert ein aggregiertes Meinungsbild der Unternehmensleistungen im Web 2.0. Da sowohl positive als auch negative Eigenschaften sichtbar sind, lassen sich Maßnahmen für die folgende Produktserie oder die Preispolitik ableiten (vgl. THORLEUCHTER ET AL. (2010), S. 440ff.).

In diesem Kontext betrachten GRUHL ET AL. (2005) in ihrem Beitrag Kundenmeinungen zu Büchern, die in Weblogs veröffentlicht wurden, und vergleichen diese mit den realen Verkaufszahlen dieser Bücher im Online-Shop von Amazon. Durch ihre Erkenntnisse bei der Zeitreihenanalyse konnten sie auf Basis der Meinungen und Erfahrungen Spitzenwerte bei den Verkaufszahlen des beschriebenen Buches vorhersagen (vgl. GRUHL ET AL. (2005), S. 86).

MISHNE & GLANCE (2005) haben ebenfalls Erfolge bei der Vorhersage von Produktverkaufszahlen erzielt. In ihrem Beitrag zeigen die Autoren, wie sie auf Basis von Stimmungen in Weblogs Verkaufszahlen von Filmen vorhersagen können. Zum Aufbau des Modells wurden Angaben der Internet Movie Database mit Texten aus Blog-Beiträgen in Verbindung gebracht und ausgewertet. Die Autoren konnten zeigen, dass ein Zusammenhang zwischen positiven Bewertungen in Weblogs und dem finanziellen Erfolg der betrachteten Filme besteht (vgl. MISHNE & GLANCE (2005), S. 4).

Die Erkenntnisse aus der Meinungsanalyse ermöglichen dabei zum einen die Bildung eines Modells zur Prognose des Produkterfolgs – mit diesem Wissen lässt sich z.B. die Marketing-Strategie anpassen, um das Produkt positiv zu präsentieren. Zum anderen tragen diese Erkenntnisse aber auch dazu bei, im Rahmen der Produktentwicklung bestehende Produkte zu verbessern oder neue Produkte stärker an den Kundenwünschen auszurichten.

5 FAZIT UND AUSBLICK

Die Analysesysteme der BI sind in der Lage, mit Daten unterschiedlichen Strukturierungsgrades umzugehen. Strukturierte Daten werden schon seit vielen Jahren mit Data-Mining-Verfahren ausgewertet, um Erkenntnisse aus ihnen abzuleiten (siehe Abschnitt 2.3).

In neuerer Zeit ist daraus die Nachbardisziplin Text Mining entstanden, die die Auswertung unstrukturierter Daten ermöglicht. Text Mining führt dazu Verfahren aus verschiedenen Forschungsfeldern zusammen, allen voran Data Mining und Computerlinguistik (siehe Abschnitt 3.1), und nutzt diese Verfahren zur Strukturierung und anschließender Auswertung von Texten (siehe Abschnitt 3.2). Das Vorgehen bei Text-Mining-Projekten ist dabei an den KDD-Prozess angelehnt, unterscheidet sich jedoch in einigen Teilschritten – z.B. in der Phase der Datenaufbereitung – stark von diesem Data-Mining-spezifischen Ansatz (siehe Abschnitt 3.3).

Wie der Beitrag weiterhin zeigt, lassen sich durch Text Mining in vielen Anwendungsbereichen Potenziale realisieren (siehe Abschnitt 4). Insbesondere durch die stetige Zunahme von unternehmensexternen, nutzergenerierten Texten im Web 2.0 liegen entscheidungsrelevante, unstrukturierte Daten vor, die mit speziellen Verfahren gesammelt und verarbeitet werden müssen. Dies betrifft viele Bereiche der BI, deren reibungsloses Zusammenwirken überhaupt erst wertvolle Analysen erlaubt. Forschungsbedarf lässt sich daher z.B. für ETL-Prozesse ableiten, die auf die Extraktion von Texten aus Webseiten ausgerichtet werden müssen, aber auch für DWHs, die unstrukturierte Daten aufnehmen und integrieren sollen, oder Analyseverfahren, die ständig weiterentwickelt werden, um der Vielfalt der gesprochenen Sprache gewachsen zu sein.

LITERATURVERZEICHNIS

Archak, N.; Ghose, A.; Ipeirotis, P. (2011): Deriving the Pricing Power of Product Features by Mining Consumer Reviews, in: Management Science, Vol. 57, Nr. 8, S. 1485-1509.

Baars, H.; Kemper, H.-G. (2008): Management Support with Structured and Unstructured Data—An Integrated Business Intelligence Framework, in: Information Systems Management, Vol. 25, Nr. 2, S. 132-148.

Bauer, A.; Günzel, H. (2004): Data-Warehouse-Systeme, 2. überarbeitete und aktualisierte Auflage, dpunkt-Verlag, Heidelberg.

Benamara, F.; Cesarano, C.; Picariello, A.; Recupero, D.; Subrahmanian, V. (2007): Sentiment Analysis: Adjectives and Adverbs are Better than Adjectives Alone, in: Glance, N.; Nicolov, N.; Adar, E.; Hurst, M.; Liberman, M.; Salvetti, F. (Hrsg.): Proceedings of the First International Conference on Weblogs and Social Media, ICWSM 2007, Boulder, Colorado, USA, March 26-28, 2007.

Chang, C.-W.; Lin, C.-T.; Wang, L.-Q. (2009): Mining the text information to optimizing the customer relationship management, in: Expert Systems with Applications, Vol. 36, Nr. 2, Part 1, S. 1433-1443.

Chaudhuri, S.; Dayal, U.; Narasayya, V. (2011): An overview of business intelligence technology, in: Commun. ACM, Vol. 54, Nr. 8, S. 88-98.

Choudhary, A.; Oluikpe, P.; Harding, J.; Carrillo, P. (2009): The needs and benefits of text mining applications on post project reviews, in: Computers in Industry, Nr. 60, S. 728-740.

Ciravegna, F.; Lavelli, A. (1999): Full Text Parsing using Cascades of Rules: an Information Extraction Perspective (Hrsg.): EACL 1999, 9th Conference of the European Chapter of the Association for Computational Linguistics, June 8-12, 1999, University of Bergen, Bergen, Norway, The Association for Computer Linguistics, S. 102-109.

Davis, H.; Oberholtzer, M. (2008): What are they saying about us?, URL: http://greenfield-ciaosurveys.com/assets/pdfs/Davis_0108-Blogmining.pdf, Abruf am: 06.08.2009.

Delgado, M.; Martin-Bautista, M.; Sanchez, D.; Vila, M. (2002): Mining Text Data: Special Features and Patterns, in: Hand, D.; Adams, N.; Bolton, R. (Hrsg.): Pattern Detection and Discovery, Springer Berlin Heidelberg, S. 140-153.

Fachgruppe Business Intelligence (2011): Positionspapier des Leitungsgremiums der GI-Fachgruppe Management Support Systems, URL: http://fg-wi-bi.gi.de/uploads/media/GI_FG_BI_PositionspapierUmbenennung_2011.pdf, Abruf am: 25.11.2013.

Fayerman, M. (2002): Customer Relationship Management, in: New Directions for Institutional Research, Vol. 2002, Nr. 113, S. 57-68.

Fayyad, U. (1996): Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, California.

Felden, C.; Bock, H.; Gräning, A.; Molotowa, L.; Saat, J. (2006): Evaluation von Algorithmen zur Textklassifikation.

Feldman, R.; Dagan, I. (1995): Knowledge Discovery in Textual Databases (KDT), in: Fayyad, U.; Uthurusamy, R. (Hrsg.): Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal, Canada, August 20-21, 1995, AAAI Press, S. 112-117.

Feng, X.; Fuhai, L. (2012): Patent text mining and informetric-based patent technology morphological analysis: an empirical study, in: Technology Analysis & Strategic Management, Vol. 24, Nr. 5, S. 467-479.

Ferber, R. (2003): Information retrieval, 1. Aufl, dpunkt, Heidelberg.

Gluchowski, P. (2001): Business Intelligence - Konzepte, Technologien und Einsatzbereiche, in: HMD - Praxis der Wirtschaftsinformatik, Nr. 222, S. 5-15.

Gluchowski, P.; Gabriel, R.; Dittmar, C. (2008): Management Support Systeme und Business Intelligence, 2. vollständig überarbeitete Auflage, Springer-Verlag, Berlin.

Gneiser, M. (2010): Wertorientiertes CRM, in: Wirtschaftsinformatik, Vol. 52, Nr. 2, S. 95-104.

Golfarelli, M.; Rizzi, S.; Cella, I. (2004): Beyond Data Warehousing: What's Next in Business Intelligence? (Hrsg.): Proceedings of the 7th ACM International Workshop on Data Warehousing and OLAP, ACM, New York, NY, USA, S. 1-6.

Gruhl, D.; Guha, R.; Kumar, R.; Novak, J.; Tomkins, A. (2005): The predictive power of online chatter, in: Grossman, R.; Bayardo, R.; Bennett, K. (Hrsg.): Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, August 21-24, 2005, ACM, S. 78-87.

Halper, F. (2013): Text Analytics Hits the Mainstream, in: Business Intelligence Journal, Vol. 18, Nr. 2, S. 29-34.

Han, J.; Kamber, M. (2002): Data mining, Nachdruck, Morgan Kaufmann Publishing, San Francisco, California.

Han, J.; Kamber, M. (2006): Data mining, 2. Auflage, Morgan Kaufmann, San Francisco, California.

He, W. (2013): Improving user experience with case-based reasoning systems using text mining and Web 2.0, in: Expert Systems with Applications, Vol. 40, Nr. 2, S. 500-507.

Hearst, M. (1999): Untangling Text Data Mining, in: Dale, R.; Church, K. (Hrsg.): 27th Annual Meeting of the Association for Computational Linguistics, Univeristy of Maryland, College Park, Maryland, USA, 20-26 June 1999, ACL.

Heinrich, K.; Hilbert, A.; Kersten, M. (2012): Methoden für Trendanalysen im Web zur Unter-stützung des Customer Relationship Management, in: Mattfeld, D.; Robra-Bissantz, S. (Hrsg.): Multikonferenz Wirtschaftsinformatik 2012, GITO-Verlag, Berlin, S. 1145-1156.

Heyer, G.; Quasthoff, U.; Wittig, T. (2006): Text Mining: Wissensrohstoff Text, 1. Auflage, W3L-Verlag, Herdecke.

Hippner, H.; Rentzmann, R. (2006a): Text Mining, in: Informatik Spektrum, Vol. 29, Nr. 4, S. 287-290.

Hippner, H.; Rentzmann, R. (2006b): Text Mining zur Anreicherung von Kundenprofilen in der Bankenbranche, in: HMD - Praxis Wirtschaftsinform, Vol. 249, S. 99-108.

Hotho, A.; Nürnberger, A.; Paaß, G. (2005): A brief survey of text mining, in: LDV Forum, Vol. 20, Nr. 1, S. 19-62.

Hussain, I.; Koaawim, L.; Ormandjieva, O. (2012): Approximation of COSMIC functional size to support early effort estimation in Agile, in: Data & Knowledge Engineering, Vol. xxx.

Kaiser, C. (2008): Produkt-Mining im Web 2.0, in: Bichler, M.; Hess, T.; Krcmar, H.; Lechner, U.; Matthes, F.; Picot, A.; Speitkamp, B.; Wolf, P. (Hrsg.): Multikonferenz Wirtschaftsinformatik, GITO-Verlag, Berlin, S. 229-240.

Kaiser, C. (2009): Opinion Mining im Web 2.0 - Konzept und Fallbeispiel, in: HMD - Praxis der Wirtschaftsinformatik, Nr. 268, S. 90-99.

Kao, A.; Poteet, S. (2007): Overview, in: Kao, A.; Poteet, S. (Hrsg.): Natural Language Processing and Text Mining, 1. Auflage, Springer-Verlag, London, S. 1-7.

Kemper, H.-G.; Mehanna, W.; Unger, C. (2006): Business Intelligence - Grundlagen und praktische Anwendungen, 2. ergänzte Auflage, Vieweg, Wiesbaden.

Kemper, H.-G.; Unger, C. (2002): Business Intelligence, in: Controlling, Vol. 14, Nr. 11, S. 665-666.

Köhler, R. (2005): Korpuslinguistik – zu wissenschaftstheoretischen Grundlagen und methodologischen Perspektiven, in: LDV Forum, Vol. 20, Nr. 2, S. 1-16.

Kurgan, L.; Musilek, P. (2006): A survey of Knowledge Discovery and Data Mining process models, in: The Knowledge Engineering Review, Vol. 21, Nr. 01, S. 1-24.

Liu, B. (2008): Opinion Mining, URL: http://www.cs.uic.edu/%7Eliub/FBS/opinion-mining.pdf, Abruf am: 25.07.2011.

Lobin, H. (2004): Textauszeichnung und Dokumentgrammatiken, in: Lobin, H.; Lemnitzer, L. (Hrsg.): Texttechnologie, 1, Stauffenburg Verlag Brigitte Narr, Tübingen, S. 51-82.

Mehler, A.; Wolff, C. (2005): Perspektiven und Positionen des Text Mining, in: LDV Forum, Vol. 20, Nr. 1, S. 1-18.

Miller, T. W. (2005): Data and text mining, Internat. ed., Pearson Prentice Hall, Upper Saddle River, NJ.

Miner, G.; Delen, D.; Elder, J.; Fast, A.; Hill, T.; Nisbet, R. (2012): Practical text mining and statistical analysis for non-structured text data applications, 1, Academic Press, Waltham, MA.

Mishne, G.; Glance, N. (2005): Predicting Movie Sales from Blogger Sentiment, URL: http://www.nielsen-online.com/downloads/us/buzz/wp_MovieSalesBlogSntmnt_Glance_2005.pdf, Abruf am: 19.04.2010.

Moss, L.; Atre, S. (2003): Business Intelligence Roadmap: The Complete Project Life-cycle for Decision-Support-Applications, Addison-Wesley Professional.

Natarajan, M. (2005): Role of text mining in information extraction and information management, in: DESIDOC Bulletin of Information Technology, Vol. 25, Nr. 4, S. 31-38.

Negash, S. (2004): Business intelligence, in: Communications of the Association for Information Systems, Vol. 13, Nr. 1, S. 177-195.

Pang, B.; Lee, L. (2008): Opinion Mining and Sentiment Analysis, in: Foundations and Trends in Information Retrieval, Vol. 2, Nr. 1-2, S. 1-135.

Petersohn, H. (2005): Data Mining - Verfahren, Prozesse, Anwendungsarchitektur, Oldenbourg, München.

Runkler, T. (2010): Data-Mining, 1, Vieweg + Teubner, Wiesbaden.

Russom, P. (2007): The shifting continuum, URL: http://apps.teradata.com//tdmo/v07n04/pdf/AR5468.pdf, Abruf am: 29.10.2013.

SanJuan, E.; Ibekwe-SanJuan, F. (2006): Text mining without document context, in: Information Processing & Management, Vol. 42, Nr. 6, S. 1532-1552.

Saravanan, M.; Raj, P. (2003): Summarization and categorization of text data in high-level data cleaning for information retrieval, in: Applied Artificial Intelligence, Vol. 17, S. 461-474.

Sommer, S.; Schieber, A.; Hilbert, A.; Heinrich, K. (2012): What is the Conversation About? A Topic-Model-Based Approach for Analyzing Customer Sentiments in Twitter, in: International Journal of Intelligent Information Technologies, Vol. 8, Nr. 1, S. 10-25.

Stahlknecht, P.; Hasenkamp, U. (2005): Einführung in die Wirtschaftsinformatik, 11. vollständig überarbeitete Auflage, Springer-Verlag, Berlin, Heidelberg.

Thommen, J.-P.; Achleitner, A.-K. (2003): Allgemeine Betriebswirtschaftslehre, 4., überarb. u. erw. Aufl, Gabler, Wiesbaden.

Thorleuchter, D.; van den Poel, D.; Prinzie, A. (2010): Extracting Consumers Needs for New Products - A Web Mining Approach (Hrsg.): Third International Conference on Knowledge Discovery and Data Mining, S. 440-443.

Tseng, F.; Chou, A. (2006): The concept of document warehousing for multidimensional modeling of textual-based business intelligence, in: Decision Support Systems, Vol. 42, Nr. 2, S. 727-744.

Tseng, Y.-H.; Lin, C.-J.; Lin, Y.-I. (2007): Text mining techniques for patent analysis, in: Information Processing & Management, Vol. 43, Nr. 7, S. 1216-1247.

Vossen, G. (2008): Datenmodelle, Datenbanksprachen und Datenbankmanagementsysteme, 5. überarbeitete und erweiterte Auflage, Oldenbourg, München.

Weiss, S.; Indurkhya, N.; Zhang, T. (2010): Fundamentals of predictive text mining, Springer-Verlag, London, New York.