Kursrelevanzprognose von Ad-hoc-Meldungen: Text Mining wider die Informationsüberlastung im Mobile Banking

Anja Schulz [1]
Humboldt-Universität zu Berlin

Myra Spiliopoulou
Otto-von-Guericke-Universität Magdeburg

Karsten Winkler [2]
HHL - Leipzig Graduate School of Management

Zusammenfassung: Internet-basierte Börseninformationsdienste erfreuen sich großer Beliebtheit bei privaten und institutionellen Anlegern. Die sprichwörtliche Flut täglich verfügbarer Unternehmensnachrichten erschwert jedoch eine effiziente Weiterleitung von tatsächlich kapitalmarktrelevanten Informationen insbesondere auf portable, technisch meist eingeschränkt ausgestattete Endgeräte im Rahmen des Mobile Banking. In dieser Fallstudie werden die von DAX100-Unternehmen über die Deutsche Gesellschaft für Ad-hoc-Publizität in den Jahren 1999 bis 2002 nach § 15 WpHG veröffentlichten Mitteilungen verwendet, um ein Klassifikationsmodell für die Prognose der Kursrelevanz einer Ad-hoc-Meldung zu generieren und zu testen. Dabei wird die Methode der Wissensentdeckung in textuellen Datenbanken bzw. des Text Mining angewendet, um tatsächlich kapitalmarktrelevante Ad-hoc-Meldungen für eine potenzielle Weiterleitung auf portable Endgeräte wie etwa Mobiltelefone oder persönliche digitale Assistenten zu selektieren. Im Gegensatz zu profilbasierten oder kollaborativen Relevanzfiltern wird ein Verfahren der empirischen Kapitalmarktforschung zur objektiven Bestimmung der Kursrelevanz von Trainings- und Testmitteilungen eingesetzt.

Schlüsselworte: Ad-hoc-Meldung, Kursrelevanz, Ereignisstudie, Mobile Banking, Text Mining, Textklassifikation, Informationsüberlastung, Börseninformationen

---

1 Diese Autorin wird durch die Deutsche Forschungsgemeinschaft im Rahmen des Sonderforschungsbereiches SFB 373 (Quantifikation und Simulation ökonomischer Prozesse) unterstützt.

2 Dieser Autor wurde durch die Deutsche Forschungsgemeinschaft im Rahmen des Projekts DIAsDEM unterstützt (DFG-Zuwendungen SP 572/4-1 und SP 572/4-3).

1 Mobile Banking und Informationsüberlastung

Die Abwicklung von Geschäftsprozessen unter Einsatz mobiler Endgeräte und geeigneter Informations- und Kommunikationstechnologien für eine standortunabhängige Kommunikation und Datenverarbeitung wird als Mobile Business bzw. m-Business bezeichnet [Kost02, S. 131]. Technische Basis für Mobile Business sind neben kabellosen Übertragungstechnologien (z. B. Bluetooth, GSM, UMTS) insbesondere portable Endgeräte wie bspw. Mobiltelefone, persönliche digitale Assistenten (PDAs), Kombinationsgeräte aus Mobiltelefon und PDA sowie tragbare Computer. Insbesondere die ersten drei Klassen mobiler Endgeräte weisen im Vergleich zu herkömmlichen Rechnern eine Vielzahl technischer Restriktionen wie z. B. miniaturisierte Bedienelemente, eingeschränkte Displaygröße und geringe Datenübertragungsrate auf, die bei der Entwicklung gebrauchstauglicher m-Business-Applikationen Berücksichtigung finden müssen [Kost02, S. 131-137].

Ubiquitäre Nutzbarkeit von m-Business-Applikationen und Lokalisierbarkeit der Anwender sind die wohl auffälligsten Merkmale des Mobile Business. Kandidaten für erfolgreiche Geschäftsmodelle des m-Business sind deshalb nach Petersmann und Nicolai insbesondere geschäftliche Aktivitäten, bei denen Ort, Zeitpunkt oder deren Kombination wichtige Einflussvariablen sind. Die Autoren systematisieren mögliche Pull- und Push-Anwendungen des Mobile Business in den Bereichen Informationsbereitstellung, Kommunikation, Informationsselektion sowie Anbahnung, Aushandlung und Abwicklung von Transaktionen [PeNi01, S. 12-17]. Pull-Applikationen stellen Informationen oder Dienste zur Befriedigung individueller, aktiver Nachfrage der Kunden bereit. Im Gegensatz dazu abonnieren Kunden von Push-Applikationen einmalig Informationen oder Dienste, die anschließend regelmäßig durch den Anbieter übermittelt bzw. ausgeführt werden.

Mobile Business in der Finanzdienstleistungsbranche sei hier als Mobile Banking bzw. m-Banking definiert, das Kreditinstituten mit den portablen Endgeräten der Kunden einen zusätzlichen Distributionskanal für Bankdienstleistungen eröffnet. Ein mobiles Bankportal kann z. B. zeitkritische Marktinformationen bereitstellen, kundenindividuell selektieren, an Abonnenten eines Push-Dienstes übermitteln und zugleich Wertpapiertransaktionen über mobile Endgeräte ermöglichen. Dabei schafft Mobile Banking einerseits Mehrwert durch die sofortige Weitergabe äußerst zeitkritischer, sehr spezifischer Informationen an den Kunden unabhängig von dessen Standort. Aus Kundensicht wird Mehrwert andererseits auch durch die Möglichkeit der sofortigen, standortunabhängigen Umsetzung von Finanzdispositionen aufgrund aktueller Marktentwicklungen generiert [BMO01, S. 179-198].

Welche Unternehmensnachrichten sind nun aber in Zeiten der sprichwörtlichen Informationsüberflutung zeitkritisch genug, um sie an Kunden mit eingeschränkter mentaler Informationsverarbeitungskapazität und technisch beschränkt ausgestatteten mobilen Endgeräten zu übermitteln? Allein die in den Jahren 1999 bis 2002 über die Deutsche Gesellschaft für Ad-hoc-Publizität veröffentlichten Ad-hoc-

Meldungen, ein Bruchteil der in diesem Zeitraum insgesamt verfügbaren Unternehmensnachrichten, haben bspw. ein Datenvolumen von etwa 94 MB.

Ein Ziel mobiler Bankportale muss die Vermeidung von negativem Stress bei Kunden durch Informationsüberlastung sein. Abonnenten sollten also nicht durch ein zu großes, teilweise irrelevantes und deshalb nicht mehr gänzlich wahrnehmbares Informationsangebot von ihren täglichen Aufgaben abgelenkt werden. In einer Studie von Farhoomand und Drury gaben von 124 Managern etwa 37 Prozent (bzw. 27 Prozent) an, täglich (bzw. häufig) Symptome der Informationsüberlastung zu spüren, deren häufigste Auswirkungen Zeitverlust, Negativeffekte in Bezug auf die Qualität der eigenen Arbeitsleistung, verminderte Effizienz sowie Frustation und Stress sind. Etwa die Hälfte der Befragten nutzt Techniken der Informationsfilterung zur Verminderung der Informationsüberlastung [FaDr02]. Darüber hinaus sollten mobile Portale eine kundenindividuelle Informationsselektion auch wegen technischer Besonderheiten der portablen Endgeräte (wie z. B. Displaygröße) und spezieller Nutzungsgewohnheiten der mobilen Benutzer anbieten. Persönliche digitale Assistenten werden bspw. im Gegensatz zur typischen stationären Internetnutzung (d. h. explorative und zeitlich oft unbeschränkte Recherche) tendenziell zur schnellen, aufgabenorientierten Problemlösung eingesetzt [AlKi00].

Die inhaltsbasierte Filterung von Informationen ist eine verbreitete Methode zur Verminderung von Informationsüberlastung, deren spezielle Techniken sich im Grad der Automatisierung und Personalisierung unterscheiden [BHKR98, S. 106]. Basis einer personalisierten Informationsselektion können einerseits nutzerspezifische Interessensprofile sein, die von Kunden explizit angelegt oder aber implizit aus deren beobachtetem Verhalten abgeleitet werden [BPC00; BBE+02]. Kollaborative Informationsfilter werten andererseits entweder explizite Kundenempfehlungen aus oder leiten implizit Empfehlungen aus dem aggregierten Verhalten von Benutzern innerhalb von Gruppen einander ähnlicher Kunden ab [BHKR98; ChJo02]. Im Gegensatz zu diesen inhaltsbasierten Informationsfiltern wird in diesem Beitrag ein objektives, an der Informationswirkung ausgerichtetes Verfahren für die automatisierte Informationsselektion im Mobile Banking vorgeschlagen. Dabei wird die Relevanz einer Nachricht durch die Aktienkursentwicklung des jeweiligen Unternehmens nach deren Veröffentlichung bestimmt.

Dieser objektive, kapitalmarktorientierte Relevanzfilter ermöglicht eine Selektion von Informationen, die nicht auf vorab erfassten bzw. beobachteten, meist sehr subjektiven inhaltlichen Kriterien (z. B. branchenbezogene Stichwörter) für die Interessantheit einer Ad-hoc-Meldung basiert. Ein für alle Kapitalmarktteilnehmer relevantes objektives Maß für die Interessantheit einer Nachricht ist die durch ihre Veröffentlichung induzierte Kursreaktion am Kapitalmarkt. Für eine erfolgreiche Einbettung in operative Informationssysteme muss der entsprechende Informationsfilter die höchstwahrscheinlich eintretende Aktienkursentwicklung nur auf Grundlage des Inhalts einer publizierten Unternehmensnachricht möglichst präzise vorhersagen. Welche Meldungsinhalte induzieren nun aber statistisch signifikante

Kursreaktionen? Da eine manuelle Evaluation sämtlicher Meldungen durch Experten aus Zeit- und Kostengründen nicht zu vertreten ist, wird in diesem Beitrag die Methode der Wissensentdeckung in textuellen Datenbanken zur automatisierten Erkennung statistisch signifikanter Muster angewandt.

Im Rahmen einer Fallstudie werden die von DAX100-Unternehmen über die Deutsche Gesellschaft für Ad-hoc-Publizität im Zeitraum 1999 bis 2002 veröffentlichten Ad-hoc-Mitteilungen analysiert, um mit Text-Mining-Verfahren ein Klassifikationsmodell für die Kursrelevanzprognose zu generieren und zu evaluieren. Dazu wird im nächsten Abschnitt der Begriff *Kursrelevanz einer Ad-hoc-Meldung* eingeführt. Im Anschluss daran wird in Abschnitt 3 die angewandte Methode der Wissensentdeckung in textuellen Datenbanken zusammengefasst. Nach Darstellung der durchgeführten Ereignisstudie zur Kursrelevanzbestimmung von Ad-hoc-Mitteilungen und einer prozessorientierten Präsentation der Fallstudie im vierten Abschnitt schließt der Beitrag mit einer Zusammenfassung und einem Ausblick.

2 Begriff und Kursrelevanz von Ad-hoc-Meldungen

Seit 1995 sind Emittenten von Wertpapieren, die im Amtlichen Handel oder am Geregelten Markt notiert sind, im Rahmen der Ad-hoc-Publizität gesetzlich verpflichtet, die Geschäftsleitung der zuständigen Börsen, das Bundesamt für Finanzdienstleistungsaufsicht und die Kapitalmarktteilnehmer unverzüglich über potenziell kursrelevante Tatsachen bzw. Unternehmensnachrichten zu informieren:

> § 15 Wertpapierhandelsgesetz (WpHG)
> Veröffentlichung und Mitteilung kursbeeinflussender Tatsachen
>
> (1) Der Emittent von Wertpapieren, die zum Handel an einer inländischen Börse zugelassen sind, muss unverzüglich eine neue Tatsache gemäß § 15 Abs. 3 Satz 1 veröffentlichen, die in seinem Tätigkeitsbereich eingetreten und nicht öffentlich bekannt ist, wenn sie wegen der Auswirkungen auf die Vermögens- oder Finanzlage oder auf den allgemeinen Geschäftsverlauf des Emittenten geeignet ist, den Börsenpreis der zugelassenen Wertpapiere erheblich zu beeinflussen, (...).

Die Deutsche Gesellschaft für Ad-hoc-Publizität mbH mit Sitz in Frankfurt am Main übermittelt diese gesetzlich geregelten Ad-hoc-Mitteilungen im Auftrag der Emittenten vorab an Börsen und Aufsichtsbehörde sowie nach einer festgelegten Frist an die angeschlossenen nationalen und internationen Nachrichtenagenturen [DGAP02]. Abbildung 1 zeigt eine über das Informationsportal der Deutschen Gesellschaft für Ad-hoc-Publizität (DGAP) veröffentlichte Ad-hoc-Meldung.

Ad-hoc-Meldungen nach § 15 WpHG sind öffentlich zugängliche Informationen, deren Inhalt nach dem Gesetz dazu geeignet sein sollte, die jeweiligen Aktienkurse wesentlich zu beeinflussen. Inwieweit Ad-hoc-Meldungen tatsächlich zu einer erheblichen Kursreaktion führen und somit kursrelevant sind, ist bisher noch

ungeklärt. Allerdings muss die halbstarke Form der Informationseffizienz unterstellt werden, um von der tatsächlich realisierten Kursreaktion eine Schlussfolgerung auf die Kursrelevanz des Meldungsinhaltes ziehen zu können.

Ein Kapitalmarkt wird als informationseffizient bezeichnet, wenn sämtliche verfügbaren, die Bewertung des jeweiligen Unternehmens betreffenden Informationen in den Kursen der Aktien enthalten sind. Das Auftreten neuer relevanter Informationen führt unverzüglich durch Kauf- und Verkaufsentscheidungen der verschiedenen Handelsakteure zu einer entsprechenden Anpassung der Kurse. Hinsichtlich der Menge der Informationen, die sich in den Aktienkursen widerspiegeln, werden drei Formen der Informationseffizienz unterschieden: die schwache, halbstarke und starke Informationseffizienz [Fama70]. Die halbstarke Form der Informationseffizienz liegt vor, wenn sämtliche historischen Marktdaten und alle öffentlich verfügbaren Informationen in den Aktienkursen berücksichtigt sind.

Für den deutschen Kapitalmarkt existieren bereits mehrere empirische Studien zur Informationswirkung von Ad-hoc-Meldungen. So untersucht Röder die Kursreaktionen von 912 Ad-hoc-Meldungen im Zeitraum 01.07.1996 bis 30.06.1997, die zuvor in inhaltsbezogene Kategorien eingeteilt wurden [Röde00]. Er stellt fest, dass Informationen über Dividendenzahlungen vergleichsweise zu schwachen, dennoch statistisch signifikanten Kursbewegungen führen, während die Kurse am stärksten durch Mitteilungen über Auftragseingänge beeinflusst werden. Ebenso scheint die Kursrelevanz einer Ad-hoc-Meldung von der Marktkapitalisierung des anzeigenden Unternehmens abzuhängen. In der von Röder untersuchten Stichprobe reagieren Aktien mit niedriger Marktkapitalisierung stärker aber auch langsamer, d. h. über den Veröffentlichungstag hinaus, auf positive oder negative Ad-hoc-Meldungen als Aktien des DAX- oder MDAX-Portefeuilles. Die aufgezeigte verzögerte Informationsverarbeitung bei kleineren Aktien wird von Röder in einer weiteren Untersuchung auf Basis von Umsatzdaten bekräftigt [Röde02].

Abbildung 1: Ad-hoc-Mitteilung der SAP AG vom 07.01.2000

Andere empirische Studien legen Indizien vor, dass nicht jede Ad-hoc-Meldung eine kursbeeinflussende Wirkung besitzt. Nowak weist bspw. darauf hin, dass nicht mehr als ein Drittel aller Ad-hoc-Meldungen im Zeitraum von 01.01.1995 bis 31.12.1996 statistisch signifikant kursrelevant waren [Nowa01, S. 465]. Bei guten Marktbedingungen scheinen besonders Unternehmen des ehemaligen Neuen Marktes Ad-hoc-Mitteilungen als Werbemedium zu nutzen, da sie in diesen Zeiten tendenziell mehr veröffentlichen als bei schlechter Marktlage [Gütt02, S. 14].

Die Autoren der vorgestellten Studien haben den Inhalt von Ad-hoc-Meldungen manuell entsprechend ihres Inhalts kategorisiert (z. B. Umsatz, Gewinn, Dividendenankündigung oder Kapitalerhöhung) und anschließend die Informationswirkung bzw. Kursrelevanz von Meldungen oder Meldungsgruppen untersucht. Im Gegensatz dazu erfolgt in diesem Beitrag keine ex-ante inhaltsbezogene Gruppierung der Ad-hoc-Meldungen. Es werden vielmehr Text-Mining-Methoden eingesetzt, um Beziehungen zwischen dem Inhalt einer Ad-hoc-Mitteilung und ihrer tatsächlich realisierten Kursreaktion zu entdecken bzw. ein Klassifikationsmodell zur automatischen Kursrelevanzprognose zu generieren. Ad-hoc-Meldungen sollen ohne manuellen Eingriff mit großer Wahrscheinlichkeit einer der zwei Klassen *kursrelevant* bzw. *kursirrelevant* zugeordnet werden, um eine automatisierte Informationsselektion innerhalb eines Mobile-Banking-Portals zu ermöglichen.

Abbildung 2 illustriert links eine positiv kursrelevante Ad-hoc-Meldung der SAP AG, die zu einer Kurssteigerung von 19,4 Prozent am Tag ihrer Wirksamkeit führte. Die Herausforderungen einer hochwertigen Kursrelevanzprognose wird in der rechts in Abbildung 2 dargestellten, auf den ersten Blick durchaus positiven Ad-hoc-Meldung der ThyssenKrupp AG angedeutet. Trotz sehr guter Unternehmensnachrichten in der gesamten Meldung ergab sich am Tag ihrer Wirksamkeit ein Kursverlust von 15,3 Prozent für die Aktie der ThyssenKrupp AG.

SAP schließt Geschäftsjahr 1999 mit stärkstem 4. Quartal in der Unternehmensgeschichte ab

Walldorf, 7. Januar 2000. Die SAP AG hat einer ersten Analyse der vorläufigen Geschäftszahlen zufolge einen Umsatz mit neuen Softwarelizenzen von nahezu 800 Mio. EUR im 4. Quartal 1999 erzielt. Dies entspricht einer Steigerung (...)

Emittent:	**SAP AG**
Veröffentlichung:	**07.01.2000, 08:27:00**
Schlusskurs 06.01.2000:	**432,00 EUR**
Schlusskurs 07.01.2000:	**516,00 EUR (+19,4%)**

Guter Start ins neue Geschäftsjahr / Ergebnis verdoppelt, Auftragseingang deutlich gestiegen

Beflügelt durch konjunkturellen Rückenwind erzielte ThyssenKrupp in den ersten sechs Monaten des Geschäftsjahres 1999/00 einen Auftragseingang von 18,7 Mrd Euro, 21,3 % mehr als im gleichen Zeitraum des Vorjahres (...)

Emittent:	**ThyssenKrupp AG**
Veröffentlichung:	**24.05.2000, 07:30:00**
Schlusskurs 23.05.2000:	**22,90 EUR**
Schlusskurs 24.05.2000:	**19,40 EUR (-15,3 %)**

Abbildung 2: Zwei Beispiele kursrelevanter Ad-hoc-Meldungen

Die Entwicklung des Aktienkurses ist als alleiniger Indikator für die Kursrelevanz einer Ad-hoc-Mitteilung nicht geeignet und wird deshalb in Abbildung 2 nur zur Illustration des Konzepts verwendet. Für die Ermittlung der Kursrelevanz einer Ad-hoc-Mitteilung erfolgt in dieser Fallstudie eine Bereinigung der Kursentwicklung um Einflüsse des Kapitalmarktes. Das dabei angewendete Verfahren wird in Abschnitt 4.2 detailliert vorgestellt.

3 Text Mining auf Ad-hoc-Meldungen

Im Gegensatz zu strukturierten Daten (z. B. Zeitreihen von Aktienkursen) werden textuelle Ad-hoc-Mitteilungen als unstrukturierte Daten bezeichnet. Schätzungen zufolge sind bis zu 80 Prozent betrieblicher Informationen nicht in strukturierter Form, sondern in unstrukturierten Textdokumenten abgelegt [Sull01, S. 56]. Diese großen Dokumentbestände erfordern aufgrund knapper Ressourcen intelligente Methoden der Datenanalyse, um daraus möglichst automatisiert entscheidungsrelevantes Wissen zu extrahieren. Analog zur Wissensentdeckung in Datenbanken (Data Mining) wurde dazu der Begriff Wissensentdeckung in textuellen Datenbanken (Text Mining) von Feldman und Dagan geprägt [FeDa95]. Das ultimative Ziel von Text Mining ist die Entdeckung von neuem, nicht-trivialem, interessantem und wirtschaftlich verwertbarem Wissen in großen Textbeständen.

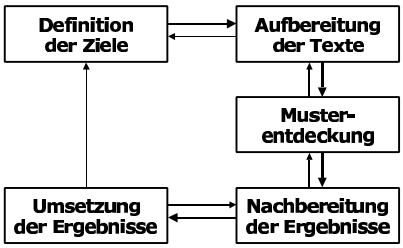

Abbildung 3: Prozess der Wissensentdeckung in textuellen Datenbanken

Angelehnt an [FPS96] wird in Abbildung 3 der generische Prozess der Wissensentdeckung in textuellen Datenbanken illustriert. Eine klare Definition der Zielstellung, meist Lösung eines betriebswirtschaftlichen Problems, bildet dabei den Ausgangspunkt für ein erfolgreiches Text-Mining-Projekt. Im Kontext der Vermeidung von Informationsüberflutung im Mobile Banking ist es Ziel dieser Fallstudie, ein Vorhersagemodell für die mit hoher Wahrscheinlichkeit richtige Prognose der Kursrelevanz von Ad-hoc-Mitteilungen zu generieren.

Natürliche Sprache ist naturgemäß hochdimensional, syntaktisch komplex und zudem semantisch oft mehrdeutig. Deshalb umfasst die Phase der Aufbereitung von Textdokumenten meist die folgenden Schritte: Sprachidentifikation, Extraktion der Texte aus Dokumenten, Zerlegung der Texte in einzelne Wörter, Wortreduktion auf grammatische Grundformen sowie Entfernung sinnleerer, besonders seltener und sehr häufiger Wörter. Für die Repräsentation textueller Daten wird häufig das Vektorraummodell des Information Retrieval [BaRi99, S. 27-30] verwendet. Dabei wird jedes Dokument in einen Vektor überführt, dessen Dimensionen den insgesamt in einem Textarchiv vorkommenden Wörtern entsprechen. Jede Dimension eines Dokumentvektors repräsentiert die Häufigkeit des entsprechenden Wortes innerhalb des korrespondierenden Textes. Somit wird jedes Dokument als ein Vektor in einem hochdimensionalen Raum modelliert.

Zwei Dokumente sind einander ähnlich, wenn ihre Vektoren topologisch nahe zueinander sind. Diese Modellierung ist aber suboptimal, weil die meisten Dokumente lediglich eine kleine Auswahl der Wörter des gesamten Wortschatzes enthalten und somit nur wenige vergleichbare Texte zu finden sind. Deshalb wird versucht, sämtliche Dimensionen entsprechend ihrer Relevanz zu gewichten. Diese Gewichtung erfolgt meistens auf Basis der Worthäufigkeit, wobei das Gewicht für ein Wort des Dokumentvektors mit zunehmender Häufigkeit des gleichen Worts im gesamten Archiv abnimmt. Viele Algorithmen der Musterentdeckung sind nicht speziell für Textdaten optimiert und können daher nicht so viele Dimensionen effizient verarbeiten, wie sie typischerweise im Vektorraum von Dokumenten auftreten. Deshalb wird vor der Musterentdeckung mittels herkömmlicher statistischer Verfahren oder Methoden des maschinellen Lernens bzw. der künstlichen Intelligenz meist eine automatische oder manuelle Reduktion der Vektordimensionen durchgeführt. Auch im Rahmen der Fallstudie sind sämtliche Ad-hoc-Mitteilungen zunächst aufzubereiten und in die von Klassifikationsalgorithmen vorausgesetzte Vektorform zu überführen.

Die in der Phase der Musterentdeckung einzusetzenden Algorithmen werden von den Anforderungen der jeweiligen Problemstellung diktiert. In der Fallstudie wird z. B. ein Klassifikationsverfahren genutzt, um Ad-hoc-Meldungen entsprechend ihres Inhalts in eine von zwei gegebenen Klassen einzuordnen: Eine Mitteilung kann entweder kursrelevant sein oder aber keine Kursrelevanz aufweisen.

Nach der Musterentdeckung müssen die Ergebnisse ausgewertet, aus betriebswirtschaftlicher Sicht interpretiert und hinsichtlich ihrer Qualität evaluiert werden. In dieser Phase werden meist Visualisierungsverfahren eingesetzt, um den Experten zu unterstützen. Eine anspruchsvolle Aufgabe ist dabei die Ableitung von Handlungsempfehlungen oder die Einbettung des entdeckten Wissens in die betrieblichen Geschäftsprozesse, um das Projektziel zu erreichen. Der gesamte Prozess der Wissensentdeckung ist iterativ, d. h. einzelne Phasen sollten bei Bedarf wiederholt werden. Eine Anwendung für das in der Fallstudie trainierte Klassifikationsmodell wäre die Einbindung in ein operatives Informationssystem, das nur höchstwahrscheinlich kursrelevante neue Ad-hoc-Mitteilungen auf portable Endgeräte der Kunden übermittelt. Nach Projektende ist die Einhaltung der Prämissen zu überwachen, um etwa bei Änderungen der spezifischen Inhalte von Mitteilungen aufgrund von Gesetzesnovellen das Klassifikationsmodell zu aktualisieren.

4 Kursrelevanzprognose von Ad-hoc-Meldungen

In diesem Abschnitt erfolgt zunächst die Beschreibung der verwendeten Datenbasis. Analog zum Prozess der Wissensentdeckung in textuellen Datenbanken wird anschließend die Erzeugung eines Klassifikationsmodells zur automatisierten Kursrelevanzprognose von Ad-hoc-Mitteilungen detailliert vorgestellt.

4.1 Datenbasis der Fallstudie

Grundlage der Fallstudie sind alle Ad-hoc-Meldungen, die von der Deutschen Gesellschaft für Ad-hoc-Publizität im Zeitraum vom 1. Januar 1999 bis 31. Dezember 2002 im Auftrag der Emittenten veröffentlicht wurden. Diese enthalten neben der nach § 15 WpHG potenziell kursrelevanten Tatsache in Textform auch das Datum und die Uhrzeit der Veröffentlichung.

Datenbasis I: Mitteilungen von Unternehmen mit Veröffentlichung durch die DGAP von 01.01.1999 bis 31.12.2002 in ca. 94 MB großer Textdatei	29.552
Mitteilungen in Englisch von Unternehmen mit Veröffentlichung durch die DGAP von 01.01.1999 bis 31.12.2002	- 8.332
Datenbasis II: Mitteilungen in Deutsch von Unternehmen mit Veröffentlichung durch die DGAP von 01.01.1999 bis 31.12.2002	= 21.220

Tabelle 1: Datenbasis der Fallstudie (Teil 1)

Ad-hoc-Mitteilungen müssen nach § 15 WpHG in Deutsch, können jedoch auch zeitgleich in Englisch veröffentlicht werden. Einige Meldungen enthalten sowohl eine deutsche als auch eine englische Fassung des Inhalts. Wie in Tabelle 1 zusammengefasst, wurden deshalb zunächst Meldungen und Meldungsteile in englischer Sprache aus dem Archiv der insgesamt im betrachteten Zeitraum durch die DGAP veröffentlichten Mitteilungen entfernt. Dazu wurde eine angepasste Version des Programms TextCat [Noor02] zur Sprachidentifikation eingesetzt.

Datenbasis II: Mitteilungen in Deutsch von Unternehmen mit Veröffentlichung durch die DGAP von 01.01.1999 bis 31.12.2002	21.220
Mitteilungen von Unternehmen ohne Listung im Index DAX100 zu irgendeinem Zeitpunkt zwischen 01.01.1999 und 31.12.2002	- 18.772
Datenbasis III: Mitteilungen (DAX100) in Deutsch mit Veröffentlichung durch die DGAP von 01.01.1999 bis 31.12.2002	= 2.448
Nicht auf Grundlage des § 15 WpHG veröffentlichten Mitteilungen (z. B. Unternehmensnachrichten oder Meldungen von sog. Director's Dealings)	- 134
Datenbasis IV: Ad-hoc-Meldungen (DAX100) in Deutsch mit Veröffentlichung durch die DGAP von 01.01.1999 bis 31.12.2002	= 2.314

Tabelle 2: Datenbasis der Fallstudie (Teil 2)

Anschließend wurden, wie in Tabelle 2 dargestellt, Ad-hoc-Mitteilungen von 136 Unternehmen extrahiert, die zu irgendeinem Zeitpunkt zwischen 1999 und 2002 dem DAX100 angehörten. Dieser Aktienindex umfasst annähernd die 100 bezüglich der Marktkapitalisierung größten und umsatzstärksten Unternehmen des Amtlichen Handels oder Geregelten Marktes der Frankfurter Börse. Es erfolgt eine Beschränkung auf größere Unternehmen, da deren Aktien im Vergleich zu kleineren Unternehmen häufiger gehandelt werden. Somit steht am Tag der Wirksamkeit einer Ad-hoc-Meldung auch häufiger ein Transaktionskurs zur Verfügung, dessen Vorliegen eine wichtige Voraussetzung für die korrekte Bestimmung der Kursrelevanz von Ad-hoc-Meldungen ist. Zusätzlich erfolgte eine Bereinigung der Datenbasis um sämtliche Mitteilungen, die nicht aufgrund § 15 WpHG veröffentlicht wurden, da diese keine Ad-hoc-Mitteilungen im Sinne des Gesetzes sind.

Für die zur Ermittlung der Kursrelevanz erforderliche Renditeberechnung werden tägliche, um Dividenden und andere Kapitalmaßnahmen bereinigte Schlusskurse aus Datastream verwendet. Aus dieser kommerziellen Datenbank stammen ebenfalls die verwendeten Indexstände des Composite DAX (CDAX).

4.2 Ermittlung der Kursrelevanz von Ad-hoc-Meldungen

Die Kursrelevanz einer Ad-hoc-Meldung ist Zielvariable bzw. abhängige Variable, deren Ausprägung (d. h. *kursrelevant* bzw. *kursirrelevant*) bei neuen Mitteilungen durch das Klassifikationsmodell nur auf Basis des textuellen Inhalts und ggf. weiterer Metadaten zu prognostizieren ist. Für das Training eines Klassifikationsmodells wird ein sog. Trainingsdatensatz benötigt, bei dem die Ausprägung der Zielvariable für alle Datensätze ex-ante gegeben ist. Die Schätzung der Klassifikationsqualität erfolgt anschließend durch Anwendung des Modells auf dem Testdatensatz, bei dem ebenfalls die Attributwerte der Zielvariablen bekannt sein müssen [EsSa00, S. 107-109]. In Ermangelung vorklassifizierter Trainings- und Testdaten muss deshalb in dieser Fallstudie zunächst die Kursrelevanz für Ad-hoc-Meldungen der Datenbasis IV (vgl. Tabelle 2) bestimmt werden. Nach Schätzung der Kurswirkung können die Ad-hoc-Mitteilungen dann als disjunkte Datensätze anteilig für Training und Test des Klassifikationsmodells verwendet werden.

In der vorliegenden Fallstudie wird zur Unterscheidung zwischen kursrelevanten und kursirrelevanten Meldungen die Ereignisstudien-Methodik genutzt. Die Ereignisstudie ist eine theoretisch fundierte Untersuchungsmethode der empirischen Kapitalmarktforschung, die sich besonders zur Bestimmung der kursbeeinflussenden Wirkung von Ad-hoc-Mitteilungen eignet [Nowa01, S. 450-451]. Dabei stellt die Veröffentlichung einer Ad-hoc-Mitteilung das Ereignis dar.

Im ersten Schritt muss der Ereignistag (Tag 0) bzw. der Tag der Wirksamkeit für jede Meldung festgelegt werden. Während Ad-hoc-Meldungen 24 Stunden am Tag abgegeben werden können, öffnet die Börse zu bestimmten Handelszeiten. Im Untersuchungszeitraum wurden diese Handelszeiten zweimal verlängert: am 18.

September 1999 von 9:00 bis 17:00 Uhr auf 9:00 bis 17:30 Uhr und am 2. Juni 2000 auf 9:00 bis 20:00 Uhr. Der Ereignistag entspricht in der Fallstudie dem Veröffentlichungstag einer Ad-hoc-Meldung, wenn deren Veröffentlichung am gleichen Tag vor Ende der Handelszeit erfolgte. Bei Abgabe einer Meldung nach Beendigung der Handelszeit wird der folgende Handelstag als Ereignistag definiert.

Zur Bestimmung der durch eine Ad-hoc-Meldung verursachten Kursreaktion muss die gesamte Kursveränderung der Aktie am Ereignistag um die Kursreaktion bereinigt werden, die ohne die Veröffentlichung der Ad-hoc-Meldung erwartet worden wäre. Zur Schätzung dieser erwarteten Rendite kann u. a. das Marktmodell verwendet werden, das einen bestimmten renditegenerierenden Prozess unterstellt [Shar63]. Das Marktmodell nimmt an, dass die Rendite $R_{i,t}$ einer Aktie des Unternehmens, das die Ad-hoc-Mitteilung i veröffentlichte,[3] am Tag t aus einem unternehmensspezifischen Bestandteil α_i und einem von der allgemeinen Marktentwicklung abhängigen Bestandteil $\beta_i \cdot R_{CDAX,t}$ sowie aus der Störgröße $e_{i,t}$ besteht. Die allgemeine Kapitalmarktentwicklung wird in der Fallstudie durch die Rendite des Composite DAX (CDAX) approximiert, der alle Unternehmen des Frankfurter Amtlichen Handels, des Geregelten Marktes und des Neuen Marktes umfasst.

$$\begin{aligned} R_{i,t} &= \alpha_i + \beta_i \cdot R_{CDAX,t} + e_{i,t} \\ t &= -55, \ldots, -6 \\ i &= 1, \ldots, N \end{aligned} \qquad (1)$$

Die in Formel (1) dargestellten Parameterwerte α_i und β_i werden für jede der N Aktien über einen 50 Tage langen Schätzzeitraum von Tag –55 bis Tag –6 durch die Kleinste-Quadrate-Schätzung bestimmt. Der Schätzzeitraum muss zur Abbildung des gewöhnlichen Zusammenhanges zwischen der Rendite der Aktie i und des Marktportefeuilles so definiert werden, dass zum einen an diesen Tagen keine ungewöhnlichen Ereignisse die Aktien betreffen und dass zum anderen die zu untersuchende Ad-hoc-Meldung noch keine Kursreaktionen bewirkt hat. Aktien, für die weniger als 35 Renditen im Schätzzeitraum zur Verfügung stehen, werden aus der Untersuchung ausgeschlossen, um eine möglichst genaue Schätzung zu gewährleisten. Ebenfalls erfolgt aufgrund der Vermutung, dass keine Transaktionskurse vorliegen, ein Ausschluss von Aktien mit einer Rendite von 0 % am Ereignistag, am vorhergehenden oder nachfolgenden Tag. Diese Vorgehensweise ist notwendig, da bei Kursen aus Datastream nicht zwischen einem im Vergleich zum Vortag konstanten Kurs oder einem vom Vortag fortgeschriebenen Kurs unterschieden werden kann. Die erwartete Rendite der Aktie i am Ereignistag $E(R_{i,0})$ wird als Summe aus dem geschätzten unternehmensspezifischen Bestandteil, $\hat{\alpha}_i$, und dem auf Basis der tatsächlich realisierten Rendite des

[3] Im Folgenden wird zur besseren Veranschaulichung der Laufindex i der Aktie direkt zugeordnet, die von der jeweilig veröffentlichten Ad-hoc-Meldung betroffen ist.

CDAX am Ereignistag geschätzten Markteinflusses, $\hat{\beta}_i \cdot R_{CDAX,0}$, berechnet. Die Differenz aus der gesamten Rendite $R_{i,0}$ und der erwarteten Rendite am Ereignistag $E(R_{i,0})$ stellt den Schätzwert für die kursbeeinflussende Wirkung der Meldung dar und wird als Überrendite oder abnormale Rendite $\hat{A}_{i,0}$ bezeichnet.

$$\hat{A}_{i,0} = R_{i,0} - E(R_{i,0}) \tag{2}$$

$$\hat{A}_{i,0} = R_{i,0} - (\hat{\alpha}_i + \hat{\beta}_i \cdot R_{CDAX,0}) \quad i = 1,\ldots,N \tag{3}$$

Um eine Aussage darüber treffen zu können, ob die geschätzten Überrenditen statistisch signifikant von null abweichen oder nur zufällig um null schwanken, wird ein t-Test durchgeführt. Meldungen, die eine zum Niveau von 10 % statistisch signifikante positive bzw. negative Überrendite bewirken, werden als positiv bzw. negativ kursrelevant bezeichnet. Kursirrelevant sind alle Ad-hoc-Meldungen mit Überrenditen, die nicht statistisch signifikant von null abweichen.

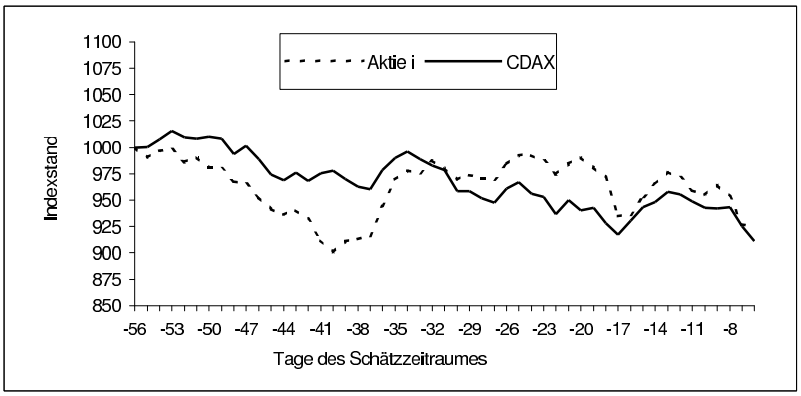

Abbildung 4: Tägliche Indexentwicklung einer Aktie *i* und des CDAX im Schätzzeitraum (Ausgangsbasis: Tag –56 = 1000)

Die obigen Ausführungen werden in den Abbildungen 4 und 5 an einem Beispiel graphisch verdeutlicht. Hierfür wird angenommen, dass die Rendite einer Aktie *i* am Ereignistag 3 % beträgt, während das Marktportefeuille in Form des CDAX eine Rendite von 1 % besitzt. Im Schätzzeitraum von Tag –55 bis Tag –6 weist die Aktie *i* gegenüber dem CDAX die in Abbildung 4 dargestellte tägliche Kursentwicklung auf. Abbildung 5 gibt für diese Zeitperiode die zugehörigen Renditekombinationen zwischen Aktie *i* und CDAX als Punktwolke an. Um die Rendite der Aktie *i* am Tag 0, die ohne Veröffentlichung der Ad-hoc-Meldung erwartet worden wäre, prognostizieren zu können, wird aus den verfügbaren Renditekombinationen das Marktmodell geschätzt. Da die Rendite des CDAX am Tag 0

annahmegemäß 1 % ist, berechnet sich nach dem geschätzten Marktmodell ($\hat{\alpha}_i = -0{,}002$, $\hat{\beta}_i = 0.799$) eine erwartete Rendite für die Aktie *i* von 0,8 %. Die tatsächliche Rendite der Aktie *i* ist allerdings 3 %, dargestellt in Abbildung 5 als Kreuz. Die sich daraus ergebene Überrendite von 2,2 % ist auf einem Niveau von 10 % statistisch signifikant von null verschieden, da die realisierte Renditekombination der Aktie *i* und des CDAX am Ereignistag außerhalb des Konfidenzintervalls liegt, welches das geschätzte Marktmodell umgibt.

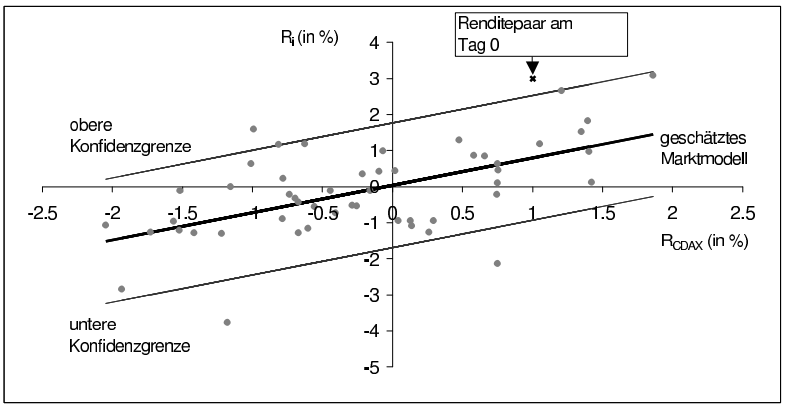

Abbildung 5: Schätzung des Marktmodells und des Konfidenzintervalls auf Basis der Renditekombinationen der Aktie *i* und des CDAX im Schätzzeitraum

Datenbasis IV: Ad-hoc-Meldungen (DAX100) in Deutsch mit Veröffentlichung durch die DGAP von 01.01.1999 bis 31.12.2002	2.314
Ad-hoc-Meldungen mit Veröffentlichungsdatum, an dem deren Emittenten mehr als eine Ad-hoc-Meldung veröffentlichten	- 320
Ad-hoc-Meldungen, für deren Kursrelevanzschätzung weniger als 35 Renditen in Schätzperiode oder Nullrenditen im Ereignisfenster vorliegen	- 534
Datenbasis V: Bereinigte Ad-hoc-Meldungen (DAX100) in Deutsch mit Veröffentlichung durch die DGAP von 01.01.1999 bis 31.12.2002	= 1.460

Tabelle 3: Datenbasis der Fallstudie (Teil 3)

Zur Isolierung ihrer Kurswirkung werden nur Ad-hoc-Meldungen mit einem Veröffentlichungsdatum berücksichtigt, an dem deren Emittenten genau eine Meldung veröffentlichten. Tabelle 3 gibt einen Überblick über zusätzlich vorgenommene Schritte der Datenbereinigung. In Datenbasis V gibt es unter den 1.460 (100 %) für die Kursrelevanzprognose verwendbaren Ad-hoc-Meldungen 235 (16,1 %) zum Niveau von 10 % statistisch signifikant positiv bzw. 161 (11,0 %) negativ kursrelevante sowie 1.064 (72,9 %) kursirrelevante Ad-hoc-Mitteilungen.

4.3 Aufbereitung der Ad-hoc-Meldungen

Die Vorverarbeitung der bereinigten Ad-hoc-Mitteilungen (Datenbasis V) umfasst den Einsatz der DIAsDEM Workbench [GSW01; WiSp01], um zusätzliche Metadaten in Form von häufigen semantischen Konzepten zu extrahieren. Das dabei genutzte DIAsDEM Vorgehensmodell zur semantischen Auszeichnung anwendungsspezifischer Textarchive beinhaltet einen interaktiven und iterativen Prozess der Wissensentdeckung. Ziel ist die inhaltsbezogene Annotation von strukturellen Textelementen (z. B. Sätzen) mit XML-Textmarken und die Ableitung einer XML-Dokumenttypdefinition (DTD). Diese DTD beschreibt die semantische bzw. inhaltliche Struktur des Archivs. Abbildung 6 illustriert auszugsweise links die in der Abbildung 1 gezeigte Ad-hoc-Mitteilung als inhaltsbezogen ausgezeichnetes XML-Dokument sowie rechts die darin referenzierte Dokumenttypdefinition.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE AdHocMitteilung SYSTEM 'AdHocMitteilung.dtd'>

<AdHocMitteilung> SAP schließt Geschäftsjahr 1999 mit
stärkstem 4. Quartal in der Unternehmensgeschichte ab.
Walldorf, 7. Januar 2000. <QuartalUmsatz Company="1010;
SAP AG" AmountOfMoney= "800000000 EUR"> Die SAP AG hat
einer ersten Analyse der vorläufigen Geschäftszahlen zu-
folge einen Umsatz mit neuen Softwarelizenzen von nahezu
800 Mio. EUR im 4. Quartal 1999 erzielt. </QuartalUmsatz>
<ErhoehungKennzahl> Dies entspricht einer Steigerung ge-
genüber dem Vorjahresquartal von ungefähr 40% und ge-
genüber dem Vorquartal von rund 150%. </Erhoehung
Kennzahl> <QuartalUmsatz> Der Gesamtumsatz im 4. Quar-
tal 1999 wuchs um rund 25% gegenüber dem Vorjahres-
quartal. </QuartalUmsatz> (...) Genaue Angaben zum vor-
läufigen Ergebnis des Geschäftsjahrs 1999 wird die SAP AG
am 24. Januar 2000 veröffentlichen. </AdHocMitteilung>
```

```
<?xml version="1.0" encoding="ISO-8859-1"?>

<!ELEMENT AdHocMitteilung ( #PCDATA |
AbschluszMasznahmePositiv | Abschreibung |
AdHocMitteilung | Akquisition | AkquisitionUmsatz |
AktieErhoehung | Ansprechpartner | AufsichtsratBeschlusz
| AufsichtsratDividende | AufsichtsratFunktionstraeger
| AufsichtsratJahresabschlusz | AufsichtsratZustimmung
| Auftragseingang | AuftragseingangPositiv |
AussageNegativ | ... | Gewinn | GewinnJeAktie |
GewinnPlanung | GewinnPositiv | GewinnRueckgang |
GewinnUmsatz | Grundkapital | ... | Zwischenbericht )* >

<!ELEMENT AbschluszMasznahmePositiv (#PCDATA)>
<!ELEMENT Abschreibung (#PCDATA)> ...
<!ELEMENT Zwischenbericht (#PCDATA)>

... <!ATTLIST QuartalUmsatz Company CDATA #IMPLIED>
<!ATTLIST QuartalUmsatz AmoutOfMoney CDATA #IMPLIED> ...
```

Abbildung 6: Semantisch annotierte Ad-hoc-Meldung mit XML DTD

Neben der nominal skalierten Branchenkennung und der Firma des Emittenten sowie Stunde, Wochentag und Kalenderwoche der Veröffentlichung sind die auf Satzebene entdeckten 129 semantischen Konzepte zusätzliche, strukturierte Metadaten. Diese binärskalierten Indikatorvariablen zeigen die Existenz bzw. Nichtexistenz des entsprechenden Konzepts in einer Ad-hoc-Mitteilung an.

4.4 Klassifikation von Ad-hoc-Meldungen

Für die Klassifikation von Ad-hoc-Mitteilungen hinsichtlich ihrer Kursrelevanz wurde der SAS Enterprise Miner [SAS02] aus der Vielzahl gegenwärtig am Markt verfügbarer allgemeiner Data-Mining-Software [WHM02] und dedizierter Text-Mining-Applikationen [Kamp02] ausgewählt. Der SAS Enterprise Miner integriert nahtlos spezielle Textanalysefunktionen (z. B. Textaufbereitung, Dimensionsreduktion und Segmentierung von Dokumenten) und anerkannte, nicht textoptimierte Data-Mining-Algorithmen (z. B. Klassifikationsverfahren). Abbildung 7 zeigt das nach vielen Iterationen des Wissensentdeckungsprozesses letztendlich für die Klassifikation der Meldungen verwendete Text-Mining-Diagramm. Die Knoten *Eingabedaten*, *Datenpartitionierung*, *Text Mining* für Textzerlegung und Dimen-

sionsreduktion sowie *Variablenselektion* und *Oversampling* des Graphen sind dabei der Aufbereitungsphase des in Abbildung 3 dargestellten Prozesses der Wissensentdeckung in textuellen Datenbanken zuzuordnen. Der Knoten *Logistische Regression* dient der Musterentdeckung bzw. der Generierung von Klassifikationswissen zur Unterscheidung von relevanten und irrelevanten Meldungen. Eine Bewertung der Qualität des gelernten Klassifikators und somit eine Nachbereitung der Ergebnisse ermöglicht der abschließende Knoten *Evaluation*. Die gerichteten Kanten des Graphen symbolisieren den Fluss von Trainings-, Test- und Metadaten zwischen den einzelnen Verarbeitungsschritten des Prozesses.

Abbildung 7: Text-Mining-Diagramm zur Klassifikation von Ad-hoc-Mitteilungen

Die Eingabedaten enthalten neben den in Abschnitt 4.3 genannten strukturierten Variablen den unstrukturierten, textuellen Inhalt der 1460 bereinigten Ad-hoc-Meldungen in Datenbasis V. Aufgrund der geringen Anzahl von Meldungen mit geschätzter Ausprägung der Zielvariable wurden 90 % der Ad-hoc-Mitteilungen für Trainings- und 10 % für Testzwecke verwendet. Dabei wurden zufällige, aber geschichtete Trainings- und Testpartitionen erzeugt, in denen der Anteil relevanter Meldungen jeweils dem in der Grundgesamtheit entspricht.

Der Knoten *Text Mining* kapselt einerseits elementare Textaufbereitungsfunktionalität. Diese umfasst die Zerlegung textueller Inhalte in individuelle, sämtlich kleingeschriebene Wörter, die Entfernung von Interpunktionszeichen und die Eliminierung sinnleerer Wörter anhand einer speziell angepassten Stoppwortliste mit 385 Termen wie z. B. „ein“, „die“ und „etwa“. Außerdem wurde eine spezielle Synonymliste mit 182 Termen eingesetzt, um synonyme Ausdrücke wie z. B. „Gewinnanstieg“ und „Gewinnerhöhung“ auf ein gemeinsames semantisches Konzept abzubilden. Auf die Identifizierung benannter Entitäten (z. B. Personen oder Datumsangaben) wurde bewusst verzichtet, um ein möglichst allgemeingültiges Klassifikationsmodell zu generieren. Trainings- und Testarchiv wurden danach in eine binäre Wort-je-Dokument-Matrix überführt. Durch Multiplikation der binären Worthäufigkeiten mit der Testgröße des χ^2-Unabhängigkeitstests auf gemeinsame Verteilung des jeweiligen Worts und der Zielvariable wurden Terme zusätzlich entsprechend ihrer Korrelation mit der Zielvariablen gewichtet.

Andererseits kapselt der Knoten *Text Mining* auch die erforderliche Dimensionsreduktion. Die verwendete Singulärwertzerlegung ist ein Verfahren der latentsemantischen Analyse [DDF+90; Sull01, S. 337-341], das die Dimensionen je Dokument auf eine vom Nutzer festzulegende Anzahl (meist ein- bis zweihundert) numerischer Attribute reduziert. Diese Methode der linearen Algebra verringert die Dimensionalität der gewichteten Wort-je-Dokument-Matrix durch Transformation in eine semantisch komprimierte Ergebnismatrix, die lediglich statistisch

bedeutsame Inhalte enthält. Dadurch wird die Relevanz bedeutungsloser Wörter gemindert und die Identifikation relevanter Termkombinationen ermöglicht.

Nach erfolgter Dimensionsreduktion wird der textuelle Inhalt einer Ad-hoc-Meldung durch maximal 200 numerische Singulärwerte beschrieben. Zusätzlich zeigen 100 neue Attribute das Auftreten der einhundert im gesamten Archiv am höchsten gewichteten Terme (z. B. „Wachstum“ und „veräußern“) in den jeweiligen Meldungen an. Der sich anschließende Knoten *Variablenselektion* kapselt die automatische Entfernung von Variablen vor der Modellbildung, die einen parametrisierbaren Schwellenwert der Korrelation mit der Zielvariable nicht überschreiten. Im Knoten *Oversampling* wird der Anteil kursrelevanter Meldungen auf 50 Prozent übergewichtet, da ein Klassifikationsmodell zur Vorhersage dieser unterrepräsentierten Klasse zu trainieren ist. Für die Klassifikation der Mitteilungen wurde ein Verfahren der logistischen Regression ausgewählt. Diese Methode prognostiziert die Wahrscheinlichkeit für das Auftreten des Attributwerts *kursrelevant* der Zielvariable bei gegebenen Attributwerten der Eingabevariablen. Ist diese Wahrscheinlichkeit größer gleich ein Drittel, so wird die entsprechende Ad-hoc-Meldung als kursrelevant klassifiziert. Die Parameterwahl für Variablenselektion und logistische Regression kann hier aus Platzgründen nicht diskutiert werden.

Um trotz des erforderlichen großen Anteils an Trainingsdaten eine möglichst genaue Schätzung des Klassifikationsfehlers durchführen zu können, wurde eine 10fache Überkreuz-Validierung durchgeführt [EsSA00, S. 109]. In zehn Durchläufen wurden jeweils 90 % der verfügbaren Daten für das Training und die verbleibenden 10 % für die Bewertung des jeweiligen Klassifikators verwendet. Tabelle 4 fasst die Ergebnisse der Evaluation der Klassifikationsgenauigkeit zusammen, wobei der durchschnittliche Klassifikationsfehler das arithmetische Mittel der Klassifikationsfehler der zehn im Rahmen der Überkreuz-Validierung generierten Klassifikationsmodelle ist. Die zweite Spalte der Tabelle 4 zeigt die Qualität des Klassifikators, der mit 1.460 Meldungen der gesamten Datenbasis V trainiert und getestet wurde. Die dritte Spalte zeigt zusätzlich das Evaluationsergebnis des Klassifikators, der mit gleichem Verfahren für die 339 vorliegenden Ad-hoc-Meldungen des Jahres 2002 trainiert und getestet wurde.

Zeitraum der Datenbasis	1999-2002	2002
Durchschnittlicher Klassifikationsfehler und 95 %-Konfidenzintervall [EsSa00, S. 111]	0,39 [0,365; 0,415]	0,45 [0,397; 0,503]
Durchschnittlicher Klassifikationsfehler der Klasse *kursrelevant* und 95 %-Konfidenzintervall	0,59 [0,565; 0,615]	0,50 [0,447; 0,553]
Durchschnittlicher Klassifikationsfehler der Klasse *kursirrelevant* und 95 %-Konfidenzintervall	0,31 [0,286; 0,334]	0,42 [0,368; 0,473]

Tabelle 4: Schätzung des Klassifikationsfehlers der zwei Klassifikatoren auf Testdaten

4.5 Interpretation der Ergebnisse

Die in Tabelle 4 dargestellten Ergebnisse der Fallstudie verdeutlichen die Herausforderungen einer Kursrelevanzprognose von Ad-hoc-Meldungen. Der durchschnittliche Klassifikationsfehler von 39 Prozent in der Datenbasis V im Zeitraum 1999 bis 2002 ist ggf. noch hinnehmbar. Zielstellung dieser Fallstudie ist jedoch die automatisierte Selektion kursrelevanter Ad-hoc-Mitteilungen zur Weiterleitung auf mobile Endgeräte im Umfeld des Mobile Banking. Der zur Beurteilung des Zielerreichungsgrades zu betrachtende durchschnittliche Klassifikationsfehler der Klasse *kursrelevant* liegt mit 59 Prozent jedoch noch weit entfernt vom Sollzustand eines automatischen Relevanzfilters für Unternehmensnachrichten im Mobile Banking: Von sämtlichen kursrelevanten Ad-hoc-Meldungen im Testarchiv wurden im Durchschnitt nur 41 Prozent tatsächlich als kursrelevant klassifiziert. Nach Training eines speziellen Klassifikators für 339, davon 33 Prozent kursrelevante Mitteilungen des Jahres 2002 konnten jedoch bereits die Hälfte der kursrelevanten Ad-hoc-Meldungen richtig klassifiziert werden.

Ein Grund für die besondere Schwierigkeit einer Kursrelevanzprognose könnte der oft mit dem Schlagwort „Informationsmüll“ artikulierte Vorwurf sein, viele Ad-hoc-Meldungen enthielten neben tatsächlich potenziell kursbeeinflussenden Tatsachen im Sinne von § 15 WpHG eher allgemeine Nachrichten oder sogar Werbung. Ebenso scheint es gängige Praxis der Finanzmarktkommunikation vieler Unternehmen zu sein, negative Nachrichten einerseits beschönigend darzustellen oder diese andererseits in „umhüllende“ positive Meldungen einzubetten. Beide Präsentationstechniken erfordern jedoch ein Lesen des kundigen Adressaten „zwischen den Zeilen“, um die Bedeutung sprachlicher Konstrukte im kommunikativen Zusammenhang zu betrachten. Diese pragmatische Perspektive der Sprachanalyse wurde, im Gegensatz zur semantischen Untersuchungsperspektive durch Einsatz einer Synonymliste, in dieser Fallstudie nicht berücksichtigt.

5 Zusammenfassung und Ausblick

In diesem Beitrag wurde die prinzipielle Möglichkeit wie auch die Komplexität einer automatisierten, kapitalmarktbasierten Selektion textueller Informationen im Mobile Banking vorgestellt. Die Methode der Wissensentdeckung in textuellen Datenbanken (Text Mining) wurde in dieser Fallstudie eingesetzt, um kapitalmarktrelevante von kapitalmarktirrelevanten Ad-hoc-Mitteilungen automatisiert zu unterscheiden. Die operative Umsetzung eines derartigen objektiven Relevanzfilters, d. h. des trainierten Klassifikationsmodells, kann dabei zur Minderung der Informationsüberflutung von Nutzern portabler Endgeräte beitragen. Die Qualität der automatischen Informationsselektion muss jedoch vor einer Einbindung in operative Systeme durch künftige Forschungsanstrengungen verbessert werden.

Weitere Erkenntnisse über die Relevanz von Ad-hoc-Mitteilungen könnten bspw. durch die zusätzliche Einbeziehung der Handelsvolumina in die empirische Untersuchung gewonnen werden. Zum einen deutet ein relativ hohes Handelsvolumen am Tag der Wirksamkeit der Ad-hoc-Meldung auf eine stattfindene Informationsverarbeitung hin. Somit könnte eine Mitteilung, die eine erhöhte Handelsaktivität verursacht, als für das Verhalten der Aktionäre relevant eingestuft werden, auch wenn sie keine Kursreaktion bewirkt. Mit Hilfe der Daten über das Handelsvolumen der jeweiligen Aktie könnte zum anderen eine Unterscheidung zwischen konstanten Kursen und von Datastream fortgeschriebenen Kursen erfolgen. Dieses würde zu einer Erhöhung der in der Fallstudie verwendbaren Datenmenge führen.

Ein wichtiger Ansatzpunkt zur Verbesserung der Klassifikationsqualität ergibt aus der in Abschnitt 4.5 dargestellten sprachlichen Komplexität von Ad-hoc-Mitteilungen. Als Ersatz einfacher Synonymlisten könnte etwa ein spezifischer Thesaurus erstellt und eingesetzt werden, der neben Synonymen auch weitere Beziehungen (z. B. „Umsatzrendite“ als Unterbegriff von „Rendite“) zwischen Wörtern und Konzepten abbildet. Gegenwärtig wird eine Stoppwortliste verwendet, um sinnleere Wörter nicht in die Wort-je-Dokument-Matrix aufzunehmen. Bei Einsatz eines Thesaurus könnte im Gegensatz dazu eine sog. Startwortliste erstellt werden, die nur gültige Terme und Konzepte enthält. Darüber hinaus sollte eine Auflösung der semantischen Bedeutung von Homonymen („Rücktritt“ als CEO oder „Rücktritt“ vom Fusionsvertrag) ebenso zur Senkung des durchschnittlichen Fehlers beitragen wie der Einbezug der pragmatischen Perspektive: Eine „Erhöhung“ des Gewinns ist z. B. im Gegensatz zur „Erhöhung“ des Verlustes durchaus positiv.

Die Ergebnisse der Fallstudie zeigen, dass ein speziell für Ad-hoc-Mitteilungen des Jahres 2002 trainierter Klassifikator einen geringeren durchschnittlichen Klassifikationsfehler für die Klasse *kursrelevant* aufweist als der Klassifikator für den Zeitraum von 1999 bis 2002. Hier sind temporale Aspekte des entdeckten Klassifikationswissens in Betracht zu ziehen, um den optimalen Zeitpunkt für die neue Generierung bzw. Aktualisierung eines Klassifikators zu ermitteln [BaSp02]. Insbesondere auf dynamischen Finanzmärkten könnte Klassifikationswissen schnell veralten. Zusätzlich sollten andere, häufig zur Textklassifikation verwendete Algorithmen eingesetzt werden, z. B. der Bayes-Klassifikator [EsSa00, S. 114-116].

Danksagungen

Die Autoren danken besonders Herrn Prof. Stehle, Ph.D., Institut für Bank-, Börsen- und Versicherungswesen der Humboldt-Universität zu Berlin, für wertvolle Hinweise und die Bereitstellung des Zugangs zu Datastream. Der Deutschen Gesellschaft für Ad-hoc-Publizität mbH wird für die Überlassung der Ad-hoc-Meldungen für Forschungszwecke herzlich gedankt. Dem SAS Academic Club danken die Autoren herzlich für die Bereitstellung des SAS Enterprise Miner.

Literatur

[AlKi00] Albers M. J.; Kim, L.: User Web Browsing Characteristics Using Palm Handhelds for Information Retrieval. In: Proceedings of IEEE Professional Communication Society International Professional Communication Conference and Proceedings of the 18th Annual ACM International Conference on Computer Documentation, Cambridge, MA, USA, September 2000: S. 125-135.

[BaRi99] Baeza-Yates, R.; Ribeiro-Neto, B.: Modern Information Retrieval. ACM Press: New York, 1999.

[BaSp02] Baron, S.; Spiliopoulou, M.: Monitoring the Results of the KDD Process: An Overview of Pattern Evolution. In: Meij, J. M. (Hrsg.). Dealing with the Data Flood: Mining Data, Text and Multimedia. STT/Beweton: Den Haag, 2002: S. 845-863.

[BBE+02] Billsus, D.; Brunk, C. A.; Evans, C.; Gladish, B.; Pazzani, M.: Adaptive Interfaces for Ubiquitous Web Access. In: Communications of the ACM 45 (2002) 5: 34-38.

[BPC00] Billsus, D; Pazzani, M. J.; Chen, J.: A Learning Agent for Wireless News Access. In: Proceedings of the 5th International Conference on Intelligent User Interfaces, New Orleans, LA, USA, January 2000: S. 33-36.

[BMO01] Böhner, G.; Mustafa, N.; Oberweis, A.: Strategische Positionierung von Finanzdienstleistern im Mobile Commerce. In: Petersmann, T.; Nicolai, A. T. (Hrsg.). Strategien im M-Commerce: Grundlagen, Management, Geschäftsmodelle. Schäffer-Poeschel Verlag: Stuttgart, 2001, S. 177-201.

[BHKR98] Borchers, A.; Herlocker, J.; Konstan, J.; Riedl, J.: Ganging up on Information Overload. In: IEEE Computer 31 (1998) 4: S. 106-108.

[ChJo02] Chandrasekaran, P.; Joshi, A.: MobileIQ: A Framework for Mobile Information Access. In: Proceedings of the Third International Conference on Mobile Data Management, Singapore, January 2002: S. 43-50.

[DDF+90] Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; Harshman, R.: Indexing by Latent Semantic Indexing. In: Journal of the American Society for Information Science 41 (1990) 6: S. 321-407.

[DGAP02] Deutsche Gesellschaft für Ad-hoc-Publizität mbH (Hrsg.): Finanzmarkt-Kommunikation aus einer Hand. 5. überarbeitete Auflage, Frankfurt am Main, September 2002. http://www.dgap.de/downloads/dgap_service_dt.pdf, Abruf am 2003-05-01.

[EsSa00] Ester, M.; Sander, J.: Knowledge Discovery in Databases: Techniken und Anwendungen. Springer-Verlag: Berlin, Heidelberg, 2000.

[FaDr02] Farhoomand, A. F.; Drury, D. H.: Managerial Information Overload. In: Communications of the ACM 45 (2002) 10: S. 127-131.

[Fama70] Eugene F. Fama: Efficient capital markets: A review of theory and empirical work. In: Journal of Finance 25 (1970) 2: S. 383-417.

[FPS96] Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.: The KDD Process for Extracting Useful Knowledge from Volumes of Data. In: Communications of the ACM 39 (1996) 11: S. 27-34.

[FeDa95] Feldman, R.; Dagan, I.: Knowledge Discovery in Textual Databases (KDT). In: Proceedings of the First International Conference on Knowledge Discovery and Data Mining, Montreal, Canada, August 1995: S. 112-117.

[GSW01] Graubitz, H.; Spiliopoulou, M.; Winkler, K.: The DIAsDEM Framework for Converting Domain-Specific Texts into XML Documents with Data Mining Techniques. In: Proceedings of the First IEEE International Conference on Data Mining, San Jose, CA, USA, November/December 2001: S. 171-178.

[Gütt02] Güttler, A.: Wird die Ad-hoc-Publizität korrekt umgesetzt? Eine empirische Analyse unter Einbezug von Unternehmen des Neuen Markts, Arbeitspapier, Fachbereich Wirtschaftswissenschaften, Johann Wolfgang Goethe-Universität: Frankfurt am Main, 2002.

[Kamp02] Kamphusmann, T.: Text-Mining: Eine praktische Marktübersicht. Symposion Publishing: Düsseldorf, 2002.

[Kost02] Koster, K.: Die Gestaltung von Geschäftsprozessen im Mobile Business. In: Hartmann, D. (Hrsg.). Geschäftsprozesse mit Mobile Computing. Vieweg Verlag: Braunschweig, Wiesbaden, 2002, S.127-145.

[Noor02] Noord, G. van: TextCat. http://odur.let.rug.nl/~vannoord/TextCat/index.html, Abruf am 2003-05-01.

[Nowa01] Nowak, E.: Eignung von Sachverhalten in Ad-hoc-Mitteilungen zur erheblichen Kursbeeinflussung. In: ZBB – Zeitschrift für Bankrecht und Bankwirtschaft 13 (2001) 6: S. 449-465.

[PeNi01] Petersmann, T.; Nicolai, A. T.: Der Möglichkeitenraum des Mobile Business - eine qualitative Betrachtung. In: Petersmann, T.; Nicolai, A. T. (Hrsg.). Strategien im M-Commerce: Grundlagen, Management, Geschäftsmodelle. Schäffer-Poeschel Verlag: Stuttgart, 2001, S. 11-26.

[Röde00] Röder, K.: Die Informationswirkung von Ad hoc-Meldungen. In: ZfB - Zeitschrift für Betriebswirtschaft 70 (2000) 5: S. 567-593.

[Röde02] Röder, K.: Intraday-Umsätze bei Ad hoc-Meldungen. In: Finanz Betrieb 4 (2002), S. 728-735.

[SAS02] SAS Institute Inc. (Hrsg.): SAS Text Miner: Distilling Textual Data for Competitive Business Advantage, A SAS White Paper. Cary, NC, USA, 2002. http://www.sas.com/apps/whitepapers/whitepaper.jsp, Abruf am 2003-05-01.

[Shar63] Sharpe, W. F.: A Simplified Model for Portfolio Analysis. In: Management Science 9 (1963) 2: S. 277-293.

[Sull01] Sullivan, D.: Data Document Warehousing and Text Mining. John Wiley & Sons: New York, 2001.

[WiSp01] Winkler, K.; Spiliopoulou, M.: Semi-Automated XML Tagging of Public Text Archives: A Case Study. In: Proceedings of EuroWeb 2001 "The Web in Public Administration". Pisa, Italy, December 2001: S. 271-285.

[WHM02] Wilde, K. D.; Hippner, H.; Merzenich, M. (Hrsg.): Data Mining: Mehr Gewinn aus Ihren Kundendaten. Verlagsgruppe Handelsblatt: Düsseldorf, 2002.