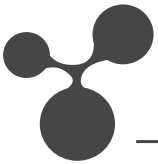
Technische Universität Dresden – Fakultät Informatik
Professur für Multimediatechnik, Privat-Dozentur für Angewandte Informatik

Prof. Dr.-Ing. Klaus Meißner
PD Dr.-Ing. habil. Martin Engelien
(Hrsg.)



GeNeMe '11

Gemeinschaften in Neuen Medien

an der
Fakultät Informatik der Technischen Universität Dresden

mit Unterstützung der

3m5. Media GmbH, Dresden
Communardo Software GmbH, Dresden
GI-Regionalgruppe, Dresden
FERCHAU Engineering GmbH, Dresden
IBM, Dresden
itsax.de | pludoni GmbH, Dresden
Kontext E GmbH, Dresden
objectFab GmbH, Dresden
queo GmbH, Dresden
Robotron Datenbank-Software GmbH, Dresden
SALT Solutions GmbH, Dresden
SAP AG, Resarch Center Dresden
Saxonia Systems AG, Dresden
T-Systems Multimedia Solutions GmbH, Dresden
Transinsight GmbH, Dresden
xima media GmbH, Dresden

am 07. und 08. September 2011 in Dresden

www.geneme.de
info@geneme.de

B.2 Worüber reden die Kunden? – Ein modelbasierter Ansatz für die Analyse von Kundenmeinungen in Microblogs

Andreas Schieber, Stefan Sommer, Kai Heinrich, Andreas Hilbert
Technische Universität Dresden

Kurzbeschreibung
Im Social Commerce entwickeln sich die Kunden zu einer bedeutenden Informationsquelle für Unternehmen. Die Kunden nutzen die Kommunikationsplattformen des Web 2.0 (z.B. Twitter), um ihre Meinungen und Erfahrungen über Produkte zu äußern. Diese Diskussionen können sehr wichtig für die Entwicklung von Produkten eines Unternehmens sein. Ein modellbasierter Ansatz soll es einem Unternehmen ermöglichen, die Meinungen zu seinen Produkten in Microblogs zu betrachten. Der erste Schritt dafür ist die Erkennung von Themen in einem spezifischen Kontext. In einem weiteren Schritt müssen die zu den Themen korrespondierenden Einträge bezüglich der geäußerten Meinungen analysiert werden. Für die Erkennung der Themen kommt ein Verfahren zum Einsatz, das auf der Latent Dirichlet Allocation basiert. Das Verfahren identifizierte eventbasierte Themen im Zusammenhang mit den 3D-TV-Anlagen von Sony.

Stichwörter: Social Commerce, Microblogs, LDA, Topic Models, Knowledge Discovery, Opinion Mining

1 Social Commerce

„Was mache ich gerade?“ beschreibt am besten die grundlegende Idee von Twitter. Über das soziale Netzwerk Twitter tauschen Personen Neuigkeiten oder Meinungen in kurzen Nachrichten aus. Twitter ist ein so genannter Microblog, das ist eine spezielle Art von Weblog, die einen gewöhnlichen Blog mit Funktionen eines sozialen Netzwerks kombiniert. Twitter war im Jahr 2009 die populärste Microblog-Applikation mit mehr als 1,8 Millionen Nutzern in Deutschland (Pattey und Stevens, 2009). Aufgrund der positiven Entwicklung von Microblogs (insbesondere von Twitter) werden diese Dienste zu einer wertvollen Quelle für Unternehmen (Pak und Paroubek, 2010, Bames und Böhringer, 2009).

„Worüber reden die Kunden?“ sollte die Frage für Unternehmen lauten. Heutzutage werden Kunden als wichtiger Kommunikationspartner angesehen, da sie wertvolles Feedback geben, Anforderungen an die Unternehmens-Performance stellen und Empfehlungsschreiben ausstellen (Richter, Koch und Krisch, 2007). Sie tauschen ihre Meinungen über die Kommunikationsplattformen des Web 2.0 aus und beeinflussen dabei den Prozess der Meinungsbildung anderer Kunden (O´Connor et al., 2010).

Dieses Phänomen wurde von Richter, Koch und Krisch (2007) als Weiterentwicklung des E-Commerce beschrieben und als Social Commerce bezeichnet. Dabei verändert sich die Kommunikation und Interaktion zwischen Unternehmen und Kunde. Insbesondere die Beziehungen und der Austausch von Informationen zwischen Unternehmen und Kunde werden wichtiger. Unternehmen im Social Commerce müssen wissen, wie sie sich nach Ansicht der Kunden verhalten sollen. Sie können diese Informationen z.B. zur Verbesserung ihrer Produkte und Dienstleistungen nutzen oder zur Einbindung der Kunden in die Produktentwicklung als so genannte Prosumers. In diesem Zusammenhang gewinnen die Kommunikationsplattformen des Web 2.0 an Bedeutung, da dabei die Interaktion zwischen den Kunden gefördert wird (Stephen und Toubia, 2010).

Das Ziel muss daher darin bestehen, die von Kunden geäußerten Meinungen auf solchen Kommunikationsplattformen zu analysieren. Wegen der großen Anzahl an Einträgen auf diesen Plattformen ist es allerdings sehr schwierig, die relevanten Inhalte ohne den Einsatz automatischer Prozeduren zu filtern. In diesem Zusammenhang ermöglicht Opinion Mining automatische Analysen von Textinhalten und unterstützt die Klassifikation von Einträgen, z.B. in positive, neutrale und negative Einträge (Liu, 2007). Vor dem Einsatz der Opinion-Mining-Verfahren müssen jedoch zunächst die relevanten Einträge identifiziert werden. Im begleitenden Beispiel ist ein Produktmanager von Sony besonders an Aussagen über Sony-Produkte interessiert, für ihn sind daher Einträge mit Bezug zu Sonys 3D-TV-Anlagen relevant.

2 Forschungsansatz

2.1 Forschungsziel und Vorgehen

In dieser Arbeit kommt der Design Science Ansatz von Hevner et al. (2004) zum Einsatz. Der Zweck von Hevner´s Ansatz ist die Entwicklung eines Artefaktes, das ein spezifisches Problem löst. In diesem Fall ist das spezifische Problem die Identifikation von Microblog-Einträgen innerhalb eines bestimmten Kontexts. Dazu sollen im Laufe der Arbeit folgende Fragestellungen beantwortet werden:

1) Welche Herausforderungen müssen bei der Analyse von Microblog-Einträgen bewältigt werden im Hinblick auf die limitierte Zeichenanzahl?

2) Wie können die Themen der Einträge automatisiert identifiziert werden?

Zur Beantwortung der Fragen werden Topic Models verwendet, welche es erlauben, automatisch Themen in einem textbasierten Datensatz zu finden. Als Datenquelle wurde der Microblogging-Service Twitter ausgewählt. Dies begründet sich darauf,

dass Twitter zum einen die populärste Microblogging-Plattform mit einer großen Anzahl an Benutzern ist, zum anderen sind die meisten auf Twitter veröffentlichten Einträge kostenlos verfügbar (Pak und Paroubek, 2010).
Im nächsten Abschnitt folgt ein Überblick über den Stand der Forschung. Anschließend werden die Charakteristika von Microblogs sowie das Vorgehensmodell dargestellt, mit dessen Hilfe die Einträge analysiert werden. Der letzte Abschnitt zeigt die Evaluierung des Ansatzes an exemplarisch ermittelten Themen in einem Twitter-Datensatz.

2.2 Stand der Forschung

Böhringer und Gluchowski (2009) beschreiben den Microblogging-Service Twitter und wie Benutzer untereinander durch die Nutzung von Web-2.0-Plattformen kommunizieren können. Die Einträge in Twitter, die sog. Tweets, enthalten verschiedene Inhalte, wie z.B. Meinungen oder Empfehlungen. Durch die Analyse dieser Inhalte können Unternehmen nützliche Einblicke in die Meinungen ihrer Kunden erhalten. Oulasvirta et al. (2010) und Tumasjan et al. (2010) zeigen, welche Einblicke das sein könnten: Oulasvirta et al. (2010) erläutern allgemeine Eindrücke, wie z.B. Studien über die Selbstoffenbarung der Nutzer; im Gegensatz dazu nutzten Tumasjan et al. (2010) Twitter, um die politischen Meinungen der Autoren zu enthüllen. Sie verwendeten 100.000 Tweets, um die politische Stimmung in Deutschland aufzuzeigen. Sie fanden heraus, dass die Mehrheit der analysierten Tweets die Präferenzen der Wähler wiederspiegelten und sogar annähernd an traditionelle Wahlumfragen heranreichten.

Der vorgestellte Ansatz nutzt Topic Models, die grundlegend von Blei und Lafferty (2009) beschrieben werden. Die ebenfalls genutzte Methode Latent Dirichlet Allocation (LDA) wurde von Blei, Ng und Jordan (2003) veröffentlicht. Seit der Publikation dieses Algorithmus wurde er erfolgreich von anderen Autoren zur Identifizierung von Themen verwendet. Ramage, Dumais und Liebling (2010) verwendeten Topic Models bereits, um Tweets zu analysieren.

3 Topic Models in Microblogs

3.1 Potenziale bei der Analyse von Microblogs

Um Microblogs zu analysieren, müssen einige Besonderheiten berücksichtigt werden. Böhringer und Gluchowski (2009) machten den Microblogging-Service Twitter und seine Funktionen bekannt: Twitter-Nutzer können miteinander kommunizieren, indem sie den Namens des Kommunikationspartners mit dem Präfix „@“ versehen. Beispielsweise schreibt Nutzer A einen Eintrag „@NutzerB“, um Nutzer B anzusprechen. Darüber hinaus können Nutzer den Eintrag eines anderen Nutzers

weiterleiten, indem sie ihn mit den Präfixzeichen „RT“ erneut veröffentlichen. Wenn bspw. Nutzer B den ursprünglichen Eintrag „Tweet“ von Nutzer A weiterleiten möchte, wird eine Nachricht mit „RT @NutzerA Tweet“ veröffentlicht. Auf diese Weise wird die Reichweite einer Äußerung erhöht, wovon letztendlich auch der Ruf des ursprünglichen Autors profitiert. Schließlich existiert mit dem Verschlagworten der Einträge noch eine äußerst wichtige Funktion von Microblogs. Diese Schlagwörter – die sogenannten Hashtags –, die vom Autor hinzugefügt werden, können durch das vorangestellte „#“ erkannt werden. Zusammenfassend ist festzuhalten, dass die technischen Funktionen von Twitter mehrere Möglichkeiten zur Analyse bieten, wie z.B. die Analyse von sozialen Netzwerken und Meinungsführern, das Web Content Mining, aber auch die Analyse des Konsumentenverhaltens.

Die Zeichenlimitierung auf 140 Zeichen ist eine weitere Besonderheit bei der Analyse von Microblogs. Um möglichst viele Informationen in einer Nachricht unterzubringen, tendieren die Nutzer zur Verwendung von Abkürzungen (z.B. wird „4ever“ als Abkürzung für „forever“ verwendet). Darüber hinaus verkomplizieren diese informelle Art des Sprechens, aber auch syntaktische Fehler den Analyse-Prozess. Bemingham und Smeaton (2010) sehen die Kürze aber als eine Stärke von Microblogs, weil die knappen Tweets kompakte und explizite Meinungen enthalten können. In ihrer Arbeit stuften sie die Klassifizierung von Meinungen in Microblogs einfacher ein als in Blogs.

3.2 Topic Models und Latent Dirichlet Allocation

Blei und Lafferty (2009) beschreiben Topic Models als eine leistungsstarke Technik zur unüberwachten Identifizierung von Strukturen in ansonsten unstrukturierten Dokumenten (z.B. Tweets). Blei und Lafferty (2009) verwendeten diese Technik für eine automatische Inhaltsverwaltung der digitalen Archive der Zeitschrift Science. Die Dokumente werden durch das Verfahren anhand der Verteilung der Wörter gruppiert, welche dazu tendieren, in ähnlichen Dokumenten gemeinsam aufzutreten. Diese Wortgruppen werden anschließend zu Themen (bzw. Topics) zusammengefasst.

3.3 Knowledge Discovery in Tweets

Der Prozess der Erkenntnisgewinnung (Knowledge Discovery in Databases, KDD) von Fayyad (1996) wurde als Grundlage für die Analyse der Twitter-Daten herangezogen und dazu in einigen Schritten modifiziert (siehe Abbildung 1: Der Prozess der Erkenntnisgewinnung bei Twitter-Daten).

Der erste Schritt ist die Auswahl der Zieldaten. Mit Hilfe der Twitter-Suche wurden die Zieldaten aus sämtlichen Twitter-Nachrichten selektiert. Wie eingangs erwähnt,wurden per Suchabfrage Tweets ausgewählt, die Stichwörter mit Bezug zu Sonys 3D-Fernsehern enthielten. Dabei wurden mehrere Abfragen durchgeführt,

die eine unterschiedliche Granularität des Kontextes aufwiesen: die erste Abfrage zielte auf Stichwörter, die allgemein im Zusammenhang mit der 3D-Technologie von Fernsehern stehen; die zweite Abfrage konzentrierte sich auf 3D-Fernseher des Herstellers Sony und die dritte Abfrage spezialisierte sich auf ein bestimmtes Produktmodell (KDL) von Sony.

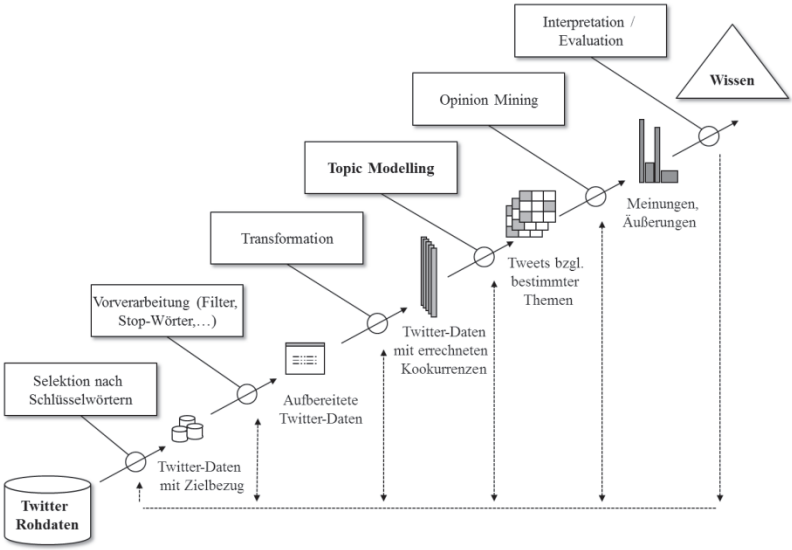


Abbildung 1: Der Prozess der Erkenntnisgewinnung bei Twitter-Daten

Nach der Auswahl der Zieldaten müssen einige Vorverarbeitungsaufgaben durchgeführt werden. Um sinnvolle Ergebnisse zu erhalten, wurden einige Elemente aus den Nachrichten entfernt. Dazu zählen Stoppworte, die Stichwörter aus dem Such-String, einzelne Zeichen und Querverweise zu anderen Nutzern (z.B. @ NutzerA). Anschließend wurde der Korpus durch Lexikalisierung und Bestimmung von Kookkurrenzen transformiert, um LDA durchzuführen zu können.
Der nächste Schritt beinhaltet die Erstellung der Topic Models. Dafür wurde der LDA-Algorithmus von Blei, Ng und Jordan (2003) implementiert, um Themencluster in den Twitter-Daten zu identifizieren. Die Ergebnisse der Analyse werden im weiteren Verlauf beschrieben.

4 Identifizierung von Themen in Twitter-Datensätzen

4.1 Konzept der Analyse

Wie bereits erwähnt wurde die Selektion der Quellendaten mit drei verschiedenen Abfragen durchgeführt, um Tweets mit Themen unterschiedlicher Granularität zu erhalten. Die Abfragen wurden zu zwei unterschiedlichen Zeiten durchgeführt, wodurch die Datensätze, die für den Modellansatz verwendet wurden, verdoppelt werden konnten. Der erste Korpus enthält ca. 1.500 Tweets, welche innerhalb von zwei Wochen (vom 16. bis zum 30. November 2010) gesammelt wurden. Der zweite Korpus enthält ca. 1.200 Tweets, welche ebenfalls innerhalb von zwei Wochen (vom 8. bis zum 22. Januar 2011) gesammelt wurden. Die Daten wurden transformiert, sodass die speziellen Elemente eines Tweets, wie z.B. Hashtags, Benutzernamen und URL´s, aus der Nachricht separiert wurden. Ein weiterer wichtiger Schritt in der Vorverarbeitung war die Entfernung der Suchbegriffe und doppelter Einträge. Anschließend wurde das LDA-Verfahren unter Nutzung des Gibbs-Sampling-Algorithmus (vgl. Ramage, Dumais und Liebling, 2010) durchgeführt.

4.2 Ergebnisse

Die Ergebnisse enthalten die Verteilungen der identifizierten Themen sowie Angaben zur Verteilung von Themen in einzelnen Dokumenten. Die Abbildung 2 zeigt die Verteilung aller Themen für den „Sony 3D“-Korpus, welcher 2010 gesammelt wurde. Die Top-8-Worte, welche das häufigste Thema X7 charakterisieren, sind ebenfalls in Abbildung 2 dargestellt.

Die weiteren Ergebnisse zeigen, dass die Verteilung aller Themen über die Tweets spezifischer wird, je detaillierter bzw. feingranularer die Such-Anfrage wird. Dabei kann eine nahezu gleichmäßige Verteilung der Themen im „3D“-Fall beobachtet werden, wohingegen die „Sony 3D“- und „Sony 3D KDL“-Datensätze sehr schiefe Verteilungen aufweisen. Dies weist darauf hin, dass das LDA-Modell in diesen Fällen häufige Themen stärker gewichtet hat. Dieses Ergebnis kann teilweise auch durch die kurze Zeitspanne der Datensätze erklärt werden. In einem bestimmten Zeitintervall ist es wahrscheinlicher, dass einige wenige Themen wie z.B. aktuelle Nachrichten oder Ereignisse kommentiert werden. In unserem „Sony 3D“-Beispiel können eindeutig die eventbasierten Einträge von anderen Tweets unterschieden werden.

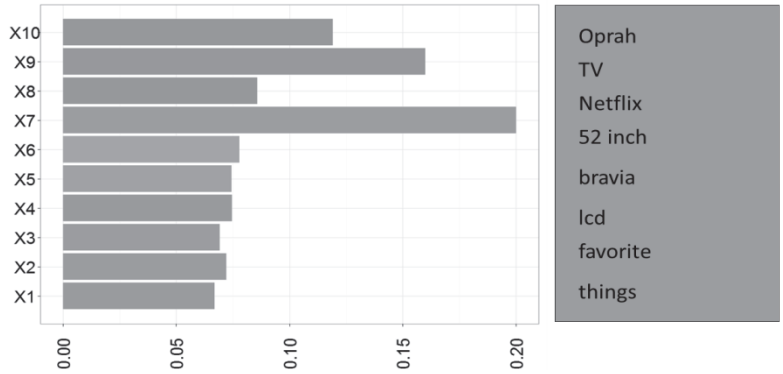


Abbildung 2: Themenverteilung im 2010er Korpus und Top-Wörter-Charakterisierung der Twitter-Einträge, die den Begriff „Sony 3D“ enthalten

Ein Beispiel dafür ist die Erwähnung der amerikanischen Talkshow-Moderatorin Oprah, die einen Sony-3D-Fernseher zu ihren Favoriten zählt (was sich im Thema X7 wiederspiegelt): Die Tweets in Abbildung 3 enthalten sowohl die Schlüsselwörter des Themas X7 (unterlegt) als auch verschiedene Äußerungen, welche zur Erreichung des zukünftigen Ziels, die Analyse von Kundemeinungen im Social Commerce, nützlich sein können.

1. 2OneQuestions: You better call oprah. RT@JamieFoxxy: I need that new Sony 3d52’ tv.

2. freeestufff: iPad tops Netflix, Sony 3D for Oprah’s ‘Favorite Things’ | How iLiving: Describing it as her “number one favorit... http://bit.ly/fWx4tV (expand)

3. GossipToday98: #NateBerkus, Did Oprah Hype 3D TV to Help Sony, Discovery? -http://ow.ly/1rKKLx

Abbildung 3: Beispielhaft ausgewählte Tweets, die in starkem Zusammenhang mit dem Thema X7 der „Sony 3D“-Daten stehen

Weiterhin konnte festgestellt werden, dass aufgrund der Wortlimitierung in den Twitter-Einträgen in einem einzelnen Dokument meistens nur ein Thema einen sehr hohen Anteil oder sogar einen Anteil von 100% aufweist.

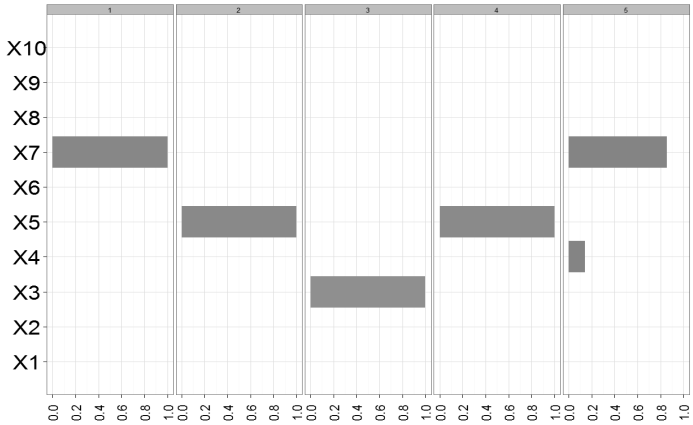


Abbildung 4: Themenverteilung von fünf zufällig ausgewählten Twitter-Einträgen, welche den Begriff „Sony 3D“ enthalten

Abbildung 4 zeigt solch eine Verteilung an fünf zufällig ausgewählten Dokumenten des „Sony 3D“-Korpus. Die Abbildung verdeutlicht, dass die Dokumente 1-4 zu 100% einem Thema zugeordnet werden, lediglich dem Dokument 5 werden zwei Themen, eines davon mit einem sehr hohen Anteil, zugeordnet. Diese Tatsache ist wichtig, da ein Modell für Twitter-Datensätze nur dann sinnvoll arbeitet, wenn es zu jedem Tweet ein einzelnes Thema – oder zumindest ein dominantes Thema – identifiziert. Nach der Betrachtung des Korpus von 2010, ist auch das Verhalten des Modells über einen größeren Zeitraum hinweg von Interesse. Dazu wurde der zweite Korpus von Anfang 2011 genutzt und die beiden Ergebnisse miteinander verglichen. Abbildung 5 zeigt die Themenverteilung für „Sony 3D“-Stichworte in dem 2011er Korpus. Das am häufigsten auftretende Cluster ist X9 und wird von fünf Stichwörtern repräsentiert.

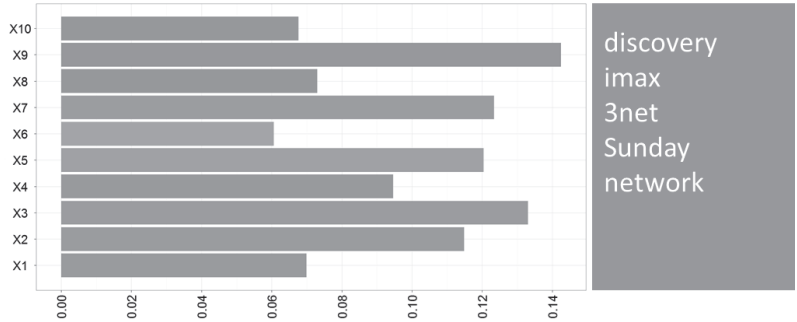


Abbildung 5: Themenverteilung im 2011er Korpus und Top-Wörter-Charakterisierung der Twitter-Einträge, die den Begriff „Sony 3D“ enthalten

Die Erkenntnisse aus dem 2010er Korpus über den kurzen Zeitraum und den großen Einfluss von besonderen Vorkommnissen, welche sich in den Wortgruppen wiederspiegeln, kann mit dem zweiten Korpus bestätigt werden (wie die Einführung des 3D-Netzwerkes „3net“ von Sony, Discovery und IMAX).

5 Fazit und Ausblick

Die Kundenkommunikation über Web-2.0-Technologien ist ein wichtiger, evolutionärer Schritt im Prozess der Meinungsbildung. Insbesondere Microblogs weisen Möglichkeiten auf, welche leistungsstarke Analysen im Bereich des Opinion Mining erlauben. Das Kennen und Verwenden von ermittelten Meinungen ist der Schlüssel, um den Kunden und seine Äußerungen zu Produkten zu verstehen. Dieses Wissen kann zur Verbesserung der Produkte oder der Produktpalette eines Unternehmens genutzt werden. Der vorgestellte Ansatz ermöglicht es, Beiträge zu identifizieren, die relevante Themen beinhalten. Durch die Anwendung von LDA kann zwischen Beiträgen, die nützlich für die Erforschung von Kundenmeinungen sind, und Beiträgen, die weniger nützliche Informationen enthalten, unterschieden werden. Die Gewinnung solcher neuer Sichtweisen auf Social-Network-Inhalte ist der erste Schritt, um zu wissen, um was sich die Diskussion wirklich dreht. Im verwendeten Beispiel kann der Sony-Produktmanager Beiträge mit interessanten Themen rund um Sony 3D analysieren. Der nächste Schritt ist die Erweiterung der Analyse durch die Implementierung eines passenden Algorithmus für Opinion Mining, um dem Produktmanager die Analyse der zum Ausdruck gebrachten Meinungen zu ermöglichen. Zusätzlich muss ein leistungsstarker Crawler entwickelt werden, um unabhängig von der Twitter API Tweets über eine längere Periode sammeln zu können.

Literatur

Barnes, S.J. und Böhringer, M., ‘Continuance Usage Intention in Microblogging Services: The Case of Twitter’, Proceedings of the 17th European Conference on Information Systems, 2009, 1-13

Bermingham, A. und Smeaton, A., ‚Classifying Sentiment in Microblogs - Is Brevity an Advantage?‘, Proceedings of the 19th ACM international conference on Information and knowledge management, 2010, 1833-1836

Blei, D. und Lafferty, J., Topic Models, [Online], URL: http://www.cs.princeton.edu/~blei/papers/BleiLafferty2009.pdf [30 Nov 2010], 2009

Blei, D., Ng, A. und Jordan, M., ‚Latent Dirichlet Allocation‘, Journal of Machine Learning Research, 2003, 933-1022

Böhringer, M. und Gluchowski, P.,‘Microblogging’,Informatik-Spektrum, 2009, 505-510

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

Hevner, A., March, S., Park, J. und Ram, S., ‚Design Science in Information Systems‘, MIS Quarterly, 28, 2004, 75-105

Liu, B., Web Data Mining, Berlin: Springer, 2007

O‘Connor, B., Balasubramanyan, R., Routledge, B. und Smith, N., ‚From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series‘, Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 2010, 122-129

Oulasvirta, A., Lehtonen, E., Kurvinen, E. und Raento, M., ‚Making the ordinary visible in microblogs‘, Personal and ubiquitous computing, Vol. 14 (3), 2010, 237-249

Pak, A. und Paroubek, P., ‚Twitter as a Corpus for Sentiment Analysis and Opinion Mining‘, Proceedings of the International Conference on Language Resources and Evaluation, 2010, 1320-1326

Pettey, C. und Stevens, H., Gartner‘s Hype Cycle Special Report for 2009, [Online], URL: http://www.gartner.com/it/page.jsp?id=1124212 [7 Dec 2010], 2009

Ramage, D., Dumais, S. und Liebling, D., ‚Characterizing Microblogs with Topic Models‘, Fourth International AAAI Conference on Weblogs and Social Media, 2010

Richter, A., Koch, M. und Krisch, J., ‘Social Commerce - Eine Analyse des Wandels im E-Commerce’, Bericht 2007/03, Fakultät Informatik, Universität der Bundeswehr München, 2007

Stephen, A.T. und Toubia, O., ‚Deriving Value from Social Commerce Networks‘, Journal of Marketing Research, Nr. 2 Vol. 67, 2010, 215-228

Tumasjan, A., Sprenger, T., Sandner, P. und Welpe, I., ‚Predicting Elections with Twitter - What 140 Characters Reveal about Political Sentiment‘, Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 2010, 178-185