Social Media Analysis

Kai Heinrich
Technische Universität Dresden
Förstereistraße 14
01099 Dresden
kai.heinrich@tu-dresden.de

Zusammenfassung

Die Analyse von Kundenmeinungen über Produkte und Technologien ist ein absoluter Schlüsselfaktor für Unternehmen. Schnelle Reaktionen auf Datenanalysen, welche aus Echtzeitmeinungsbildern gewonnen werden, können einem Unternehmen einen erheblichen Wettbewerbsvorteil verschaffen. Im Rahmen der evolutionären Entwicklung von Web 2.0 Anwendungen können die steigenden Nutzerzahlen solcher Social Media Anwendungen und die darin generierten Meinungsbilder nicht ignoriert werden. Besonders Web und Microblogs erfreuen sich großer Beliebtheit unter den Nutzern. Microblogs, wie z.B. Twitter ermöglichen die Kommunikation über kurze Statustexte im Rahmen eines sozialen Netzwerks. Während diese Art des Datenstreams viele Vorteile hat, stellen sich dem Analysten auf der anderen Seite auch Probleme in den Weg. Ein Problem ist die Masse der täglichen Posts und die daraus resultierende, nicht überschaubare Menge der relevanten Daten. Eine Lösung dieses Problems ist es, sich auf Meinungsführer zu konzentrieren. In diesem Beitrag werden Verfahren zur Identifikation von Meinugsführern im Microblogging-Dienst Twitter vorgestellt. Dabei werden inhalts- und strukturbezogene Analysen kombiniert um ein möglichst gutes Ergebnis zu erzielen. Die Ergebnisse werden anschließend mit passenden Metriken verglichen und ausgewertet.

1 Einleitung

„Was mache ich gerade?“ beschreibt am besten die grundlegende Idee von Twitter[1]. Über das soziale Netzwerk Twitter tauschen Personen Neuigkeiten oder Meinungen in kurzen Nachrichten aus. Twitter ist ein sogenannter 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 (Pettey & Stevens, 2009 [22]). Aufgrund der positiven Entwicklung von Microblogs (insbesondere von Twitter) werden diese Dienste zu einer wertvollen Quelle für Unternehmen ((Pak & Paroubek, 2010 [21]), (Barnes & Böhringer, 2009 [1])). Denn mittels der Microblogs geben Kunden wertvolles Feedback zu Produkten und Dienstleitungen, indem sie ihre Meinungen untereinander austauschen und dabei auch die Meinungsbildung anderer Kunden beeinflussen (O'Connor, Balasubramanyan, Routledge, & Smith, 2010 [20]).
Das Ziel muss für Unternehmen 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

[1] http://www.twitter.com

Inhalte ohne den Einsatz automatischer Prozeduren zu filtern. Der Beitrag widmet sich dieser Problemstellung, indem er ein Konzept präsentiert, das die Einträge in einem iterativen Verfahren anhand mehrerer Kriterien analysiert. Im Ergebnis sollen daraufhin thematisch relevante Meinungsäußerungen leichter aufzuspüren sowie im Zeitverlauf zu betrachten sein.

2 Forschungsziel und Methodik

Das Ziel dieser Forschungsarbeit ist die Erstellung eines Konzepts, mit dessen Hilfe die Einträge von Microblogs bzgl. mehrerer Kriterien analysiert werden sollen, um relevante Äußerungen von der unübersichtlichen Masse zu trennen. Die Analyse soll dabei einerseits das Referenzpotenzial der Microblog-Nutzer, also von zentralen Personen innerhalb der Kommunikationsplattform, berücksichtigen. Andererseits sollen auch die Inhalte der veröffentlichten Einträge untersucht und zu Themenclustern zusammengefasst werden.

Zur Erreichung des Ziels sollen folgende Forschungsfragen beantwortet werden:

> Wie können Meinungsführer in Microblogs identifiziert werden?
>
> Wie können die Inhalte der Einträge in Microblogs zu Themenclustern zusammengefasst werden?

Das Konzept soll unter Nutzung des Design-Science-Forschungsansatzes von (Hevner, March, Park, & Ram, 2004 [12]) entwickelt werden, indem ein Artefakt implementiert wird, das die Evaluation der in Frage kommenden Methoden ermöglicht.
Als Datenquelle soll der Microblogging-Dienst Twitter dienen: Twitter bietet die populärste Microblogging-Plattform und verfügt daher über eine große Anzahl an Nutzern. Diese Nutzer generieren täglich eine hohe Zahl an Nachrichten, die meist öffentlich zugänglich publiziert werden (Pak & Paroubek, 2010 [21]).
Im weiteren Verlauf des Beitrags wird das Problem zunächst in das Forschungsfeld eingeordnet, und wichtige Begrifflichkeiten werden geklärt. Im Anschluss werden die Charakteristika von Microblogs beschrieben. Abschnitt 4 präsentiert auf Basis dieser Erkenntnisse ein mehrstufiges Prozessmodel für die Analyse von Kundenmeinungen in Microblogs.

3 Einordnung der Problemstellung

Der folgende Abschnitt ordnet die Arbeit in das Opinion Mining ein und zeigt auf, welche Forschungsfelder zur Lösung der Problemstellung beitragen können.

3.1 Opinion Mining

Die Methoden des Opinion Mining erlauben die Extraktion von Meinungen aus Texten. Das Opinion Mining kann daher wertvolle Informationen liefern, wenn Meinungen von Kunden über die Produktpalette eines Unternehmens ausgewertet werden sollen. Die

Methoden des Opinion Mining werden wie Data- und Text-Mining-Methoden dem analyseorientierten Business-Intelligence-Verständnis zugeordnet (Gluchowski, 2001 [10]).
In der Wissenschaft finden sich zahlreiche Definitionsversuche zum Begriff Opinion Mining (Xia, Gentile, Munro, & Iria, 2009 [25], Dave, Lawrence, & Pennock, 2003 [8], Kaiser & Bodendorf, 2009 [15], Dey & Haque, 2009 [9]). (Liu, 2008 [18]) definiert das Opinion Mining als Extraktion von Attributen und Komponenten eines kommentierten Objekts; die Extraktion erfolgt dabei aus Textdokumenten, die positive, negative oder neutrale Meinungen über das Objekt enthalten.

(Liu, 2007 [17]) identifiziert im Rahmen des Opinion Mining drei Bereiche mit unterschiedlichen Aufgabenstellungen:

- „sentiment classification“: Die Analyse einer Aussage auf Dokumentenebene und Einstufung der Aussage als positiv oder negativ.
- „feature-based opinion mining and summarization“: Satzbasierte Identifikation von Eigenschaften (features) eines Objekts und Einstufung der Eigenschaften als positiv oder negativ.
- „comparative sentence and relation mining“: Der Vergleich von Objekten und Eigenschaften sowie Identifikation des präferierten Objekts.

Da es sich bei den Nachrichten in Microblogs grundsätzlich um Textdokumente handelt, eignen sich alle genannten Verfahren für die Analyse der Inhalte. Aufgrund der beschränkten Länge der Nachrichten ist es jedoch praktikabel, vor allem die Verfahren der sentiment classification einzusetzen: Über Microblogs werden meistens kurze Statements abgegeben, die selten über einen oder zwei Sätze hinausgehen; dementsprechend dürfte eine Analyse auf Dokumentenebene im Rahmen von Microblogs ähnliche Ergebnisse wie eine Analyse auf Satzebene erbringen.

3.2 Topic Modeling

(Blei & Lafferty, 2009 [3]) beschreiben Topic Models als eine leistungsstarke Technik zur unüberwachten Identifizierung von Strukturen in ansonsten unstrukturierten Dokumenten. 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 zusammengefasst.
In einer vorgelagerten Forschungsarbeit wurde diese Methode im Rahmen der Analyse von Microblogs eingesetzt. (Schieber, Sommer, Hilbert & Heinrich, 2011 [24]) haben mit Hilfe der Verfahren des Topic Modelling Microblog-Einträge über mehrere Zeiträume analysiert. Dabei zeigen die Ergebnisse, dass Themencluster identifiziert werden konnten, welche mit bestimmten, realen Ereignissen in Zusammenhang standen; so verwies ein Themencluster auf den Zusammenschluss von Sony, IMAX und Discovery zu einem Unternehmensnetzwerk (Schieber et al., 2011 [24], S. 6).
Entgegen der Auswertung eines feststehenden Datensatzes von (Schieber et al., 2011 [24]) beschreiben (Blei & Lafferty, 2009 [3]) die Möglichkeit, neben den Inhalten der Einträge auch die Zeitkomponente zu untersuchen. Die sogenannten „dynamic topic

models“ erfassen dabei die Evolution von Themen(clustern) im Verlauf der Zeit und zeigen auf, wie sich die Wortgruppen innerhalb eines Themas während der betrachteten Zeiträume verändern (Blei & Lafferty, S. 14ff.).

3.3 Soziale Netzwerke und Meinungsführer

Microblogs werden zu den sozialen Netzwerken zugeordnet. Ein soziales Netzwerk ist dabei die Komposition einer großen Anzahl von Akteuren, welche durch das Muster der zu Grunde liegenden Interaktion charakterisiert ist. (Iacobucci & Hopkins, 1992 [13], S. 5). Die Nutzer von Microblogs veröffentlichen Nachrichten, die von anderen Nutzern gelesen werden können. Durch entsprechende Funktionen können sich Nutzer zu Netzwerken zusammenschließen, innerhalb derer ihre Nachrichten schneller kursieren. Nutzer, die eine zentrale Rolle in diesen Netzwerken einnehmen, besitzen deswegen eine mitunter große Zuhörerschaft und können Verhalten und Einstellungen von anderen Netzwerkmitgliedern beeinflussen (Kröber-Riel & Weinberg, 2003 [16], S. 518). Im Forschungsfeld der Sozialen Netzwerkanalyse sind Verfahren bekannt, mit denen solche Meinungsführer aufgespürt werden können (vgl. Chin & Chignell, 2007 [7] und Brin & Page, 1998 [5]).

3.4 Aufbau und Eigenschaften von Microblogs

Abgeleitet von der Nutzung als private Tagebücher bzw. Journale sind Microblogs eine besondere Form von Weblogs. Im Unterschied zu Weblogs kann hier nur eine beschränkte Anzahl von Zeichen zur Kommunikation genutzt werden. (Java, Finn, & Tseng, 2007 [14]) beschreiben Microblogging als eine neue Form der Kommunikation, bei welcher ein Nutzer in kurzen Mitteilungen seinen Status bekanntgeben kann.
Die Funktionsweise eines Microblogs soll an dem größten Vertreter Twitter erläutert werden. (Böhringer & Gluchowski, 2009 [4]) geben einen Überblick über die Funktionen des Dienstes: Ein Nutzer hat die Möglichkeit neben dem Verfassen von den auf 140 Zeichen begrenzten Statusmeldungen auch auf andere Statusmeldungen zu reagieren. Dazu gibt es zum einen eine Antwortfunktion (REPLY) und zum anderen eine Verbreitungsfunktion (RETWEET). Die Verbreitung der Einträge erfolgt dabei über die Netzwerkstruktur, welches sich durch die Folge-Funktion (FOLLOW) ergibt. Ein User hat die Möglichkeit, die Meldungen eines anderen Nutzers im Auge zu behalten, indem er ihm folgt. So bildet sich das Netzwerk um einen User aus den entsprechenden FOLLOW-Beziehungen.

4 Meinungsanalyse in Microblogs

Im Sinne der in Abschnitt 0 erwähnten Netzwerkstruktur ist es sinnvoll, die Meinungen derer zu analysieren, welche eine zentrale Stellung einnehmen und damit ein hohes Referenzpotenzial besitzen (Brüne, 1989 [6]). Nach der Anwendung eines entsprechenden Filters werden die Verfahren aus Abschnitt 0 zur Meinungsanalyse in Microblogs herangezogen. Im Sinne des Modells der zweistufigen Kommunikation (Brüne, 1989

[6]) stehen die resultierenden Meinungen dann stellvertretend für die Masse. Die Begrenzung der Datenmenge ist eine Notwendigkeit, da die algorithmischen Beschränkungen einiger Verfahren die Analyse solch enormer Datenmengen, wie sie bei der Verwendung von Microblogs anfallen, nicht zulassen. Neben der eigentlichen Analyse von Meinungen sollen ebenfalls Themen extrahiert werden, um abschließend eine Trendbetrachtung der unterschiedlichen Meinungen in gewissen Themenbereichen untersuchen zu können.
Entsprechend der Ziele, eine Trendmodellierung mit einer Themenstruktur unter Berücksichtigung der Besonderheiten von Microblogs bzgl. Inhalt und Struktur zu vereinbaren, wird das Modell von (Mei, Ling, Wondra, Su, & Zhai, 2007 [19]) auf die abgegrenzten Subnetzwerke der Meinungsführer angewendet.

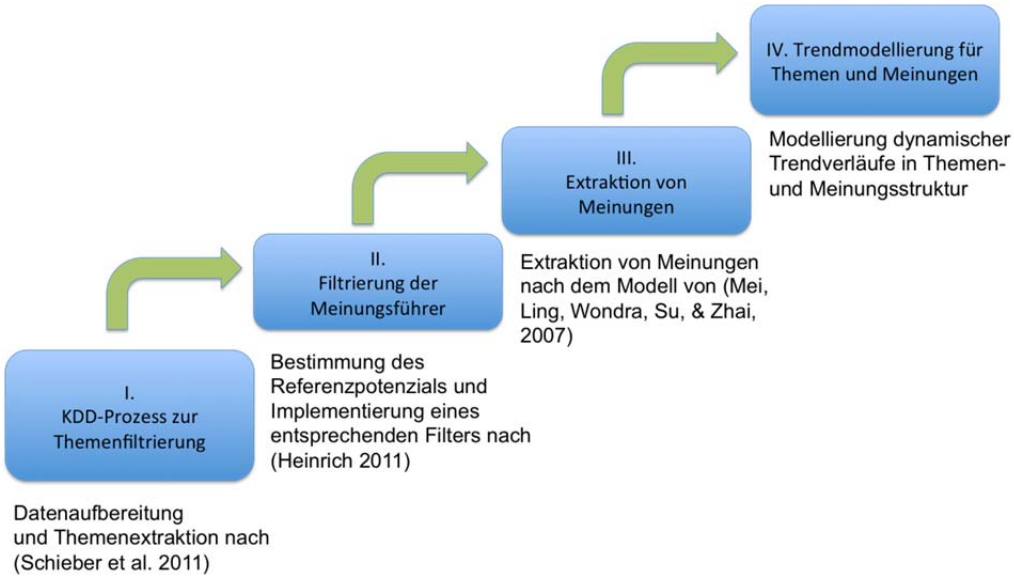

Abbildung 1: Prozessmodell zur Meinungsanalyse in Microblogs

Ziel nach der Datenaufbereitung und Themenfiltrierung (Schieber, Sommer, Heinrich, & Hilbert, 2011 [24]) in Schritt I ist die Filtrierung von Meinungsführern (Heinrich, 2011 [11]) in Schritt II. Danach können in Schritt III Meinungen extrahiert werden (Mei, Ling, Wondra, Su, & Zhai, 2007 [19]). Abschließend wird in Schritt IV die Abbildung von Meinungstrends und vor allem deren Prognose im Zeitverlauf angestrebt. Das komprimierte Prozessmodell ist in Abbildung 1 dargestellt.

5 Analyseergebnisse und verwendete Software

Für den Analyseprozess wurde ein Crawlprozess angesetzt und jeweils über zwei Zeiträume zu gängigen Keywords wie „Google“ und „Twitter“ Daten gesammelt. Nach der Bereinigung wurden Meinungsanalysen zu den jeweiligen Themen durchgeführt.
Abbildung 2 zeigt die verwendete Software in den verschiedenen Analyseschritten. Die eigentliche Meinungsanalyse wurde mit dem SAS Sentiment Analysis Toolkit durchgeführt. Die Ergebnisse sind einmal mit vorgeschalteter Meinungsführeranalyse und einmal ohne durchgeführt. Während die Ergebnisse beim „ Google“ Stichwort sich noch

die Waage halten, ist bei dem Twitter Thema aufgrund der stark verminderten Stichprobe kein hinreichendes Datenmaterial mehr vorhanden um eine ausbalancierte Analyse durchzuführen.

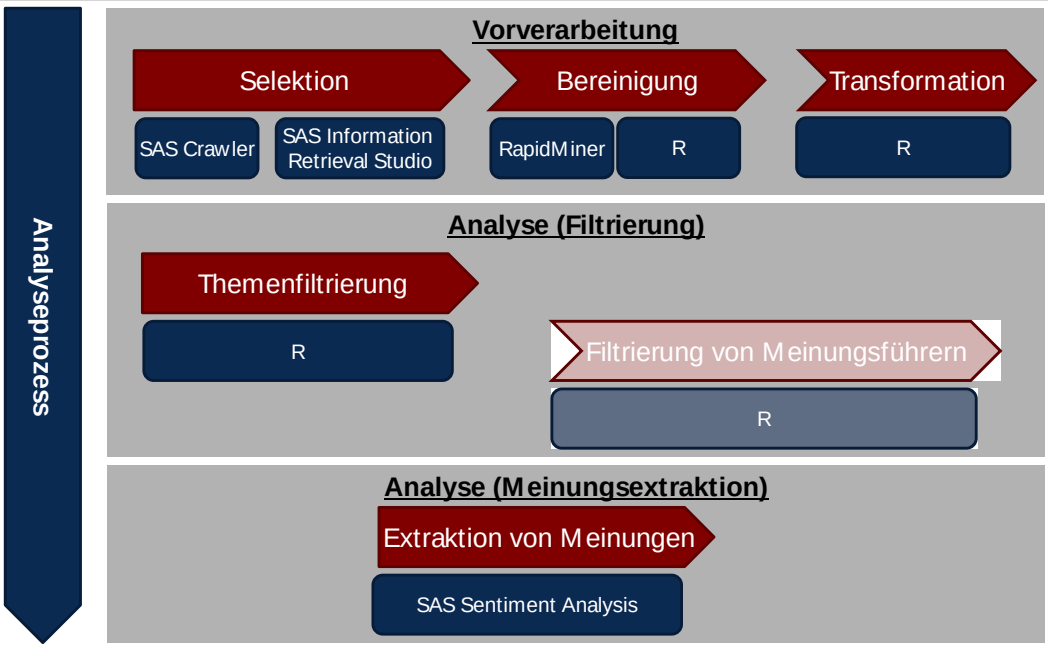

Abbildung 2: Verwendete Software zur Analyse

Die Ergebnisse sind einmal mit vorgeschalteter Meinungsführeranalyse und einmal ohne durchgeführt. Während die Ergebnisse beim „ Apple“ Stichwort sich noch die Waage halten, ist bei dem Twitter Thema aufgrund der stark verminderten Stichprobe kein hinreichendes Datenmaterial mehr vorhanden um eine ausbalancierte Analyse durchzuführen. In den Abbildungen 3.1-4.2 sind die Ergebnisse in Form von Accuracy-Tabellen zusammengefasst.

	positive	negative
pred. positive	32	13
pred. negative	4	41

Accuracy: 80%

Abbildung 3.1: Ergebnis mit Filtrierung zum Stichwort „Apple“

	positive	negative
pred. positive	134	42
pred. negative	57	335

Accuracy: 83%

Abbildung 3.2: Ergebnis ohne Filtrierung zum Stichwort „Apple“

	positive	negative
pred. positive	4	7
pred. negative	4	12

Accuracy: 59%

Abbildung 4.1: Ergebnis mit Filtrierung zum Stichwort „Twitter“

	positive	negative
pred. positive	49	11
pred. negative	19	67

Accuracy:

79%

Abbildung 4.2: Ergebnis ohne Filtrierung zum Stichwort „Twitter“

6 Ausblick und Fazit

Nachdem die Frage „Worüber reden die Kunden?“ mit der explorativen Extraktion von Themen und die Zuordnung entsprechender beschreibender Wörter beantwortet ist (Schieber, Sommer, Heinrich, & Hilbert, 2011 [24]), stellen sich darauf aufbauend die Fragen: „Was denken die Kunden über dieses Thema“ und „Wie verändert sich diese Meinung Laufe der Zeit?“ Diese Arbeit schlägt ein entsprechendes Prozessmodell zur Beantwortung beider Fragen vor, welches auf vier Säulen beruht: Themenextraktion, Filtrierung von Meinungsführern, Extraktion von Themen und schließlich der Abbildung von Meinungstrends im Zeitverlauf.
In zukünftiger Forschungsarbeit sollen auf Basis der Erkenntnisse dieser Arbeit die beiden letzten Säulen ausgestaltet, implementiert und evaluiert werden.

Literatur

[1] Barnes, S., & Böhringer, M. (2009). Continuance Usage Intention in Microblogging Services: The Case of Twitter. *Proceedings of the 17th European Conference on Information Systems*, S. 1-13.

[2] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*.

[3] Blei, D. & Lafferty, J. *Topic Models.* Retrieved September 25, 2011, from http://www.cs.princeton.edu/~blei/papers/BleiLafferty2009.pdf.

[4] Böhringer, M., & Gluchowski, P. (2009). Microblogging. *Informatik-Spektrum*.

[5] Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Seventh International Word-Wide Web Conference,* from http://ilpubs.stanford.edu:8090/361/1/1998-8.pdf.

[6] Brüne, G. (1989). Meinungsführerschaft im Konsumgütermarketing.

[7] Chin, A., & Chignell, M. (2007). Identifying communities in blogs: roles for social network analysis and survey instruments. *International Journal of Web Based Communities, 3*(3), 345–363.

[8] Dave, K., Lawrence, S., & Pennock, D. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In ACM (Ed.), *Proceedings of the 12th international conference on World Wide Web* (pp. 519–528). New York, New York: ACM.

[9] Dey, L., & Haque, M. (2009). Opinion mining from noisy text data. *International Journal on Document Analysis and Recognition, 3*(12), 205–226.

[10] Gluchowski, P. (2001). Business Intelligence - Konzepte, Technologien und Einsatzbereiche. *HMD - Praxis der Wirtschaftsinformatik*, (222), 5–15. Retrieved September 20, 2010.

[11] Heinrich, K. (2011). Influence Potential Framework: Eine Methode zur Bestimmung des Referenzpotenzials in Microblogs. *Tagungsband zum 14. Interuniversitären Doktorandenseminar Wirtschaftsinformatik.* Universitätsverlag Chemnitz.

[12] Hevner, A., March, S., Park, J., & Ram, S. (2004). Design Science in Information Systems Research . *MIS Quarterly*.

[13] Iacobucci, D., & Hopkins, N. (1992). Modeling Dyadic Interactions and Networks in Marketing. *Journal of Marketing Research, 29*(1), 5–17.

[14] Java, A., Finn, T., & Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007.*

[15] Kaiser, C., & Bodendorf, F. (2009). Opinion and Relationship Mining in Online Forums. *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, (1), 128–131.

[16] Kröber-Riel, W., & Weinberg, P. (2003). *Konsumentenverhalten* (3. Auflage). München: Vahlen.

[17] Liu, B. (2007). *Web Data Mining: Exploring hyperlinks, contents, and usage data* (1. Auflage). Berlin, Heidelberg: Springer-Verlag.

[18] Liu, B. (2008). *Opinion Mining.* Retrieved July 25, 2011, from http://www.cs.uic.edu/%7Eliub/FBS/opinion-mining.pdf.

[19] Mei, Q., Ling, X., Wondra, M., Su, H., & Zhai, C. (2007). Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs. *Proceedings of the Sixteenth International World Wide Web Conference (WWW2007).*

[20] O'Connor, B., Balasubramanyan, R., Routledge, B., & Smith, N. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series . *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.*

[21] Pak, A., & Paroubek, P. (2010). 'Twitter as a Corpus for Sentiment Analysis and Opinion Mining',. *Proceedings of the International Conference on Language Resources and Evaluation*, S. 1320-1326.

[22] Pettey, C., & Stevens, H. (2009). Abgerufen am 12 2010 von Gartner's Hype Cycle Special Report for 2009: http://www.gartner.com/it/page.jsp?id=1124212

[23] Richter, A., Koch, M., & Krisch, J. (2007). *Social Commerce - Eine Analyse des Wandels im E-Commerce* . Fakultät Informatik, Universität der Bundeswehr München.

[24] Schieber, A., Sommer, S., Heinrich, K., & Hilbert, A. (2011). Analyzing customer sentiments in microblogs – A topic-model-based approach for Twitter datasets. *Proceedings of the Seventeenth Americas Conference on Information Systems (forthcoming)* .

[25] Xia, L., Gentile, A., Munro, J., & Iria, J. (2009). Improving Patient Opinion Mining through Multi-step Classification. In V. Matoušek & P. Mautner (Eds.), *Lecture notes in computer science. Text, Speech and Dialogue* (pp. 70–76). Springer Berlin / Heidelberg.

[26] Zhang, Y., Callan, J., & Minka, T. (2002). Novelty and Redundancy Detection in Adaptive Filtering. *Proceedings of the 25st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland*.