

Abstract zur Diplomarbeit

„Opinion Mining in Weblogs“

von Dipl.-Wirt.-Inf. Joachim Franz

Betreuer: Prof. Dr. Andreas Hilbert

Was denken unsere Kunden über unsere Produkte? Welche Meinung haben Kunden zu den Konkurrenzprodukten? Die Antworten auf diese Fragen sind für jedes Unternehmen von entscheidender Bedeutung. Zum einen, um die Marketingmaßnahmen anzupassen und kritischen Stimmen entgegenzuwirken und zum anderen, um in der Produktentwicklung den Wünschen der Kunden nachzukommen. Traditionell wurden für die Marktforschung hauptsächlich Fragebögen genutzt, um Kundenmeinungen zu ermitteln. Den bekannten Vorteilen von Fragebögen stehen allerdings einige Nachteile gegenüber, wie z.B. hoher Personalaufwand, vorgegebenes Spektrum an Fragen und Vernachlässigung kompletter Kundensegmente. In den letzten Jahren hat das Internet als weitere Quelle zur Ermittlung der Kundenmeinungen stark an Bedeutung gewonnen. Immer mehr Privatpersonen veröffentlichen ihre eigenen Seiten im Internet und/oder nutzen aktiv Foren, Weblogs (Blogs), Newsgroups, aber auch Seiten kommerzieller Anbieter wie z.B. Amazon. Diese neuen Inhalte enthalten inzwischen häufig auch die Meinung der Autoren über Produkte und Dienstleistungen, die sie konsumiert haben. Diese Form des Feedbacks der Kunden ist kostenlos im Internet verfügbar, und es stellt sich deshalb die Frage, ob diese Informationen zur Meinungsforschung genutzt werden können. Ein Vorteil ist auch die freie Formulierung der Kritiken, wodurch diese nicht auf eine bestimmte Fragestellung ausgerichtet sind. Darüber hinaus können neben der Erhebung der Meinungen über die eigenen Produkte auch Auswertungen zu Konkurrenzprodukten erfolgen.

Aufgabe der Diplomarbeit war deshalb die Untersuchung, welche Ansätze für die Auswertung von Blogs existieren, um Opinion Mining zu betreiben. Opinion Mining bezeichnet in diesem Zusammenhang die Auswertung von Texten mit dem Ziel der Identifikation und Bewertung von (Kunden-)Meinungen und den zugehörigen Produkten und deren Merkmalen mit Hilfe von Data-, Text- und Web-Mining-Methoden. Außerdem sollte untersucht werden, inwieweit für diese Analysen klassische Standardsoftware geeignet ist.

Die Arbeit gliederte sich inhaltlich in zwei Abschnitte. In dem ersten Abschnitt wurde eine Auswahl der vorhandenen Literatur der Jahre 2002 bis 2008 analysiert, um die einzelnen Prozessschritte des Opinion Mining zu identifizieren. Die Auswertung orientierte sich an dem CRISP-DM-Prozess. Dabei wurden verschiedene Arten der textuellen Aufbereitung der in den Forschungsarbeiten untersuchten Webseiten und unterschiedliche Verfahren zur Klassifikation der Texte herausgearbeitet. Im zweiten Ab-

schnitt wurde aus diesen Erkenntnissen ein Anforderungskatalog für Opinion-Mining-Systeme erstellt, an dem verschiedene kommerzielle und nicht-kommerzielle Business-Intelligence-Standardsoftwarelösungen auf ihre Anwendbarkeit für das Opinion Mining evaluiert wurden. Auf dieser Basis wurde der RapidMiner für die Implementierung eines Opinion-Mining-Prototyps ausgewählt. Nach der Implementierung erfolgte eine Bewertung der Klassifikationsergebnisse, auch unter Einbeziehung der zuvor betrachteten Forschungsarbeiten.

In der Arbeit konnte herausgearbeitet werden, dass die Meinungen der Weblogautoren aus den einzelnen Einträgen herausgefiltert werden und Weblogs somit als Quelle von Meinungen dienen können. Allerdings besitzen Weblogs für die Bildung einer Kaufentscheidung einen eher untergeordneten Stellenwert im Gegensatz zu Foren und Produktreviewseiten. Zur Analyse der Außenwirkung eines Unternehmens können sie jedoch einen wichtigen Beitrag leisten. Außerdem sind polarisierende Produkte bzw. Hersteller für eine Auswertung besser geeignet, da zu diesen Produkte und Herstellern mehr Weblogeinträge zu finden sind.

Ein softwaregestütztes Opinion-Mining-System sollte das Crawlen der Weblogs und das Speichern der gefundenen Weblogeinträge beinhalten. Weiterhin muss das System die eigentlichen Weblogeinträge isolieren und linguistisch verarbeiten können. Außerdem müssen Klassifikationsverfahren implementiert sein, die eine Einteilung in Klassen erlauben. Schließlich ist es unverzichtbar, dass das System die Visualisierung der Klassen ermöglicht.

Weiterhin wurde festgestellt, dass die kommerziellen Softwareanbieter die gestellten Anforderungen an ein Opinion-Mining-System erfüllen, wenn auch nicht alle auf direktem Weg. Einerseits gibt es Softwareprodukte, die mit Ausnahme des Webcrawlings alle Anforderungen erfüllen. Andererseits existieren BI-Lösungen, die selbst keine Textanalysen anbieten, aber über strategische Kooperationen mit spezialisierten Text-Mining-Anbietern diese Lücke schließen. Unter den Open-Source-Softwarelösungen ist das Text Mining noch nicht flächendeckend vertreten. Mit Ausnahme des RapidMiner beschränken sich die nicht-kommerziellen BI-Anbieter primär auf das Reporting und OLAP-Funktionen. Wie diese Arbeit zeigte, sind aber auch mit einer nicht-traditionellen Standardsoftware die gestellten Anforderungen zu erfüllen.