

Julia Maria Struß

Multilinguales aspekt- basiertes Opinion Mining

**Entwicklung eines ressourcenarmen
Extraktionsverfahrens und Unter-
suchung von Nutzerperspektiven**

vwh

Verlag Werner Hülsbusch
Fachverlag für Medientechnik und -wirtschaft

J. M. Struß: Multilinguales aspektbasiertes Opinion Mining

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet unter <http://d-nb.de> abrufbar.

© Verlag Werner Hülsbusch, Glückstadt, 2017

vwh Verlag Werner Hülsbusch
Fachverlag für Medientechnik und -wirtschaft

www.vwh-verlag.de

Einfache Nutzungsrechte liegen beim Verlag Werner Hülsbusch, Glückstadt. Eine weitere Verwertung im Sinne des Urheberrechtsgesetzes ist nur mit Zustimmung der Autorin möglich.

Markenerklärung: Die in diesem Werk wiedergegebenen Gebrauchsnamen, Handelsnamen, Warenzeichen usw. können auch ohne besondere Kennzeichnung geschützte Marken sein und als solche den gesetzlichen Bestimmungen unterliegen.

Druck und Bindung: SOWA Sp. z o. o., Piaseczno

Printed in Poland

Zugleich: Diss., Univ. Hildesheim, 2016

– Als Typoskript gedruckt –

ISSN: 0938-8710

ISBN: 978-3-86488-123-7

**Multilinguales aspektbasiertes
Opinion Mining:
Entwicklung eines ressourcenarmen
Extraktionsverfahrens und Untersuchung
von Nutzerperspektiven**

Vom Fachbereich III (Sprach- und
Informationswissenschaften) der
Universität Hildesheim

zur Erlangung des Grades
einer Doktorin der Philosophie (Dr. phil.)

angenommene Dissertation von

Julia Maria Struß

geboren am 17. Dezember 1981 in Aachen

1. Gutachter: Prof. Dr. Christa Womser-Hacker
2. Gutachter: Prof. Dr. Thomas Mandl

Tag der Einreichung: 04.04.2016
Tag der mündlichen Prüfung: 27.10.2016

Bibliothekssiegel: Hil 2

Vorwort

Diese Dissertation ist während meiner Tätigkeit an der Universität Hildesheim im Institut für Informationswissenschaft und Sprachtechnologie entstanden. Die Arbeit ist im *Opinion Mining* an der Schnittstelle zwischen dem *Information Retrieval* und der *Computerlinguistik* angesiedelt und befasst sich mit der Entwicklung eines ressourcenarmen Ansatzes, der sich mit geringem Aufwand bei vergleichbaren Ergebnissen für andere Sprachen adaptieren lässt. Dies wird an den Sprachen Englisch, Deutsch und Spanisch demonstriert. Darüber hinaus wird in der Arbeit erstmalig auch die Nutzerperspektive im Opinion Mining betrachtet. Im Rahmen einer breit angelegten Online-Befragung, bei der die Rekrutierung der Teilnehmer über einen Crowdsourcing-Dienst erfolgte, wird untersucht, welche Faktoren Einfluss auf die wahrgenommene Nützlichkeit einzelner Meinungsäußerungen haben. Die so gewonnenen Erkenntnisse bilden die Basis für die Entwicklung von ersten Gestaltungsrichtlinien für die Ergebnispräsentation von *Opinion Mining*-Systemen und liefern Rückschlüsse für Schwerpunktsetzungen bei der automatischen Identifikation von Meinungsäußerungen.

Eine solche Arbeit kann nicht ohne die breite Unterstützung im beruflichen wie privaten Umfeld entstehen. Ich möchte daher an dieser Stelle die Gelegenheit nutzen und mich bei all jenen bedanken, die zum Gelingen dieser Arbeit beigetragen haben. Mein besonderer Dank gilt meinen Betreuern Prof. Dr. Christa Womser-Hacker und Prof. Dr. Thomas Mandl, die immer ansprechbar waren und mir in vielen konstruktiven Gesprächen und kritischen Diskussionen Rat, Anregung und Motivation gaben und mir halfen das Ziel nicht aus den Augen zu verlieren.

Auch bei den übrigen Institutsmitgliedern möchte ich mich für die vielen konstruktiven und inspirierenden Gespräche bedanken, insbesondere bei Daniela Becks und Katrin Werner, mit denen ich jeweils längere Zeit im selben Büro gearbeitet habe. In vielen intensiven Diskussionen entstanden hier Ideen, die in diese Dissertation Eingang gefunden haben. Auch den übrigen Doktoranden im Institut sei an dieser Stelle gedankt. Die regelmäßigen Gesprächsrunden haben mir immer wieder neue Motivation gegeben.

Ganz besonders danken möchte ich überdies meiner Familie, allen voran meinem Ehemann Holger, der mich über die gesamte Zeit immer unterstützt hat und häufig auf mich verzichten musste. Er und meine Tochter Clara haben mir durch ihre Liebe und ihr Vertrauen immer wieder neue Kraft gegeben.

Aber auch bei meiner Mutter Lydia und meinen Schwestern Eva und Petra, die mich unermüdlich motiviert haben und gerade in der letzten Phase häufig Babysitting-Dienste übernommen haben, bedanke ich mich von Herzen für die Unterstützung.

Söhlde, August 2017
Julia Maria Struß

Inhaltsverzeichnis

Abbildungsverzeichnis	15
Tabellenverzeichnis	21
Abkürzungsverzeichnis	27
1 Einleitung	29
2 Aufgaben und Problemstellungen des Opinion Mining und der Sentimentanalyse	35
2.1 Begriffliche Abgrenzungen	35
2.2 Dimensionen und Facetten des Opinion Mining	37
2.2.1 Zielsetzungen und Aufgaben im Opinion Mining	37
2.2.1.1 Erkennen von Subjektivität	37
2.2.1.2 Identifizieren von Meinungsäußerungen	38
2.2.1.3 Ermittlung der Polarität und Intensi-	
tät von Meinungsäußerungen	39
2.2.1.4 Identifizieren des Urhebers einer Mei-	
nungsäußerung	40
2.2.1.5 Identifizieren des Ziels einer Meinungs-	
äußerung	42
2.2.1.6 Anforderungen an ein Opinion Mining	
(OM)-System	43
2.2.2 Textsorten und Anwendungskontexte	43
2.2.3 Granularität der Untersuchungsebene	45
2.2.4 Zusammenfassung	45
2.3 Evaluierung	46
2.3.1 Evaluierung der Identifikation von Meinungsäu-	
ßerungen, Urhebern und Zielen einer Meinungs-	
äußerung	49
2.3.2 Evaluierung der Ermittlung von Polarität	50
2.3.3 Einbeziehen des Nutzers in den Evaluierungspro-	
zess	51

3	Stand der Forschung	55
3.1	Aspektbasiertes Opinion Mining	56
3.1.1	Identifikation von Meinungsäußerung	57
3.1.1.1	Monolinguale Ansätze und Methoden	58
3.1.1.2	Multi- und cross-linguale Ansätze und Methoden	67
3.1.1.3	Zusammenfassung	74
3.1.2	Identifikation der Ziele von Meinungsäußerungen	79
3.1.2.1	Identifikation von Eigennamen und To- pics als Meinungsziele	80
3.1.2.2	Identifikation von Produkteigenschaften in Rezensionen	86
3.1.2.3	Zusammenfassung	104
3.1.3	Identifikation von Relationen zwischen Meinungs- äußerungen und zugehörigen Zielen	110
3.1.3.1	Regelbasierte Ansätze	110
3.1.3.2	Maschinelle Lernverfahren	112
3.1.3.3	Zusammenfassung	122
3.1.4	Abschließende Betrachtung zum aspektbasierten Opinion Mining	126
3.2	Wahrgenommene Nützlichkeit von Meinungsäußerungen	128
3.2.1	Einflussfaktoren auf die wahrgenommene Nütz- lichkeit von Kundenrezensionen	128
3.2.1.1	Modellierung und Vorhersage der wahr- genommenen Nützlichkeit	128
3.2.1.2	Ermittlung von Einflussfaktoren auf die wahrgenommene Nützlichkeit	133
3.2.2	Einflussfaktoren auf die Relevanzbewertungen von Dokumenten und Textpassagen	143
3.2.3	Zusammenfassung	149
4	Forschungsfragen und Methoden	151
4.1	Forschungsfragen und Hypothesen	151
4.2	Methoden	161
5	Entwicklung mehrsprachiger Ressourcen für die Evaluierung feingranularer OM-Systeme	167
5.1	Existierende Korpora	167
5.2	Erstellen eines multilingualen Korpus	177
5.2.1	Annotationsmethodik	179
5.2.2	Werkzeug für den Annotationsprozess	183

5.2.3	Inter-Annotator-Agreement	184
5.3	Erstellen eines Goldstandards für die Evaluierung von Relationen zwischen Meinungsäußerungen und den zugehörigen Zielen	191
5.4	Zusammenfassung	195
6	Extraktion von Produkteigenschaften und Meinungen	197
6.1	Aufbau des Extraktionsprozesses	201
6.2	Linguistische Vorverarbeitung	203
6.2.1	Normalisierung nutzergenerierter Inhalte	203
6.2.2	OpenNLP-Komponenten	204
6.2.3	Konvertieren der morphosyntaktischen Wortklassen	205
6.2.4	Segmentieren von Teilsätzen	206
6.3	Identifizieren von Meinungsäußerungen und zugehörigen Zielen	208
6.3.1	Extraktion von Head-Modifier-Phrasen unter Verwendung von Begrenzerpaaren	209
6.3.2	Regeln für Head-Modifier-Relationen	210
6.3.3	Priorisieren der extrahierten Kandidaten	214
6.3.4	Ermitteln der Meinung sowie des zugehörigen Ziels in extrahierten Phrasen	216
6.3.5	Iterativer Filterprozess für Meinungsäußerungen und Ziele von Meinungsäußerungen	218
6.3.5.1	Erstellen einer initialen Liste meinungstragender Terme für das Englische	219
6.3.5.2	Erstellen einer initialen Liste meinungstragender Terme für das Deutsche	219
6.3.5.3	Erstellen einer initialen Liste meinungstragender Terme für das Spanische	221
6.3.6	Überprüfen von Phrasen niedrigerer Priorität	221
6.3.7	Ermitteln diskontinuierlicher Phrasen	222
6.4	Zusammenfassung	222
7	Systembezogene Evaluierung	225
7.1	Evaluierung meinungstragender Adjektiv-Nomen-Relationen	228
7.2	Evaluierung aller meinungstragenden Head-Modifier-Relationen	238
7.3	Evaluierung der Übertragbarkeit des DOHM-Ansatzes auf weitere Sprachen	241

7.4	Evaluierung der Identifikation impliziter Meinungsäußerungen und -ziele	248
7.5	Zusammenfassung	254
8	Untersuchung der wahrgenommenen Nützlichkeit von Meinungsäußerungen	259
8.1	Definition des Untersuchungsgegenstands	260
8.2	Untersuchungsdesign	260
8.2.1	Variablen	262
8.2.2	Skala zur Messung der wahrgenommenen Nützlichkeit	265
8.2.3	Aufbau des Fragebogens	266
8.2.4	Simulierte Arbeitsaufgabensituationen	268
8.2.5	Annotation der Merkmale der Meinungsäußerungen	272
8.2.6	Kriterien für die Auswahl der im Fragebogen verwendeten Meinungen	274
8.3	Pretest	276
9	Crowdsourcing als Form der Datenerhebung	277
9.1	Crowdsourcing	277
9.2	Qualität von mittels Crowdsourcing erhobenen Daten	278
9.3	Nachteile durch die Verwendung von Crowdsourcing-Plattformen	282
9.4	Rekrutierung der Teilnehmer	284
10	Auswertung	289
10.1	Beschreibung der Daten	289
10.2	Methode	294
10.2.1	Logistische Regression	295
10.2.2	Einbeziehen der subjektspezifischen Varianz in der logistischen Regression	298
10.2.3	Kreuzvalidierung	299
10.2.4	Anpassung des logistischen Regressionsmodells	299
10.3	Ergebnisse	300
10.3.1	Ergebnisse der Untersuchung des Hauptmodells	300
10.3.2	Ergebnisse der Untersuchung des Hauptmodells unter Einbezug der Kontrollvariablen	302
10.3.3	Überprüfen der Annahme auf proportionale Odds	306
10.3.4	Überprüfung von Multikollinearität	311

10.3.5	Haupteffekt für die Nennung der Produkteigenschaft in einer Meinungsäußerung	312
10.3.6	Interaktionseffekte zwischen der Art der Meinungsäußerung, dem kulturellen Hintergrund und den Kontrollvariablen	314
10.3.6.1	Interaktionseffekt zwischen der Art der Meinungsäußerung, dem kulturellen Hintergrund und dem Geschlecht	314
10.3.6.2	Interaktionseffekt zwischen Art der Meinungsäußerung, kulturellem Hintergrund und Polarität	317
10.3.7	Interaktionseffekte zwischen der Nennung von Zusatzinformation, dem kulturellen Hintergrund und den Kontrollvariablen	324
10.3.7.1	Interaktionseffekt zwischen der Nennung von Zusatzinformation, dem kulturellem Hintergrund deutscher und spanischer Teilnehmer und dem Geschlecht der Teilnehmer	326
10.3.7.2	Interaktionseffekt zwischen der Nennung von Zusatzinformation, dem kulturellen Hintergrund deutscher und spanischer Teilnehmer und der Polarität einer Meinungsäußerung	331
10.3.7.3	Interaktionseffekt zwischen der Nennung von Zusatzinformation, dem kulturellen Hintergrund amerikanischer und europäischer Teilnehmer und dem Geschlecht der Teilnehmer	332
10.3.7.4	Interaktionseffekt zwischen der Nennung von Zusatzinformation, dem kulturellen Hintergrund amerikanischer und europäischer Teilnehmer und der Polarität einer Meinungsäußerung	336
10.3.8	Erklärungsgüte des Modells	339
10.3.9	Subjektspezifische Varianz	340
10.3.10	Zusammenfassung	340
11	Fazit	345
	Literatur	353

Anhang	379
A Annotationsrichtlinien	379
B Liste der Penn Treebank PoS-Tags	382
C Detaillierte Übersicht der Evaluierungsergebnisse	383
D Befragung	391
D.1 Deutscher Fragebogen (Smartphones)	391
D.2 Korrekturen der Meinungsäußerungen für die Befragung	401
E Auswertung der Befragung	402
E.1 Histogramme	402
E.2 Bearbeitungsdauer	403
E.3 Überprüfen der Annahme auf proportionale Odds	404