31

Schriften aus der Fakultät Wirtschaftsinformatik und Angewandte Informatik der Otto-Friedrich-Universität Bamberg

Korpusbasierte Verfahren zur Generierung lexikalischer Ressourcen für das Opinion Mining

Dirk Reinel

University of Bamberg Press

31 Schriften aus der Fakultät Wirtschaftsinformatik und Angewandte Informatik der Otto-Friedrich-Universität Bamberg

Contributions of the Faculty Information Systems and Applied Computer Sciences of the Otto-Friedrich-University Bamberg

Schriften aus der Fakultät Wirtschaftsinformatik und Angewandte Informatik der Otto-Friedrich-Universität Bamberg

Contributions of the Faculty Information Systems and Applied Computer Sciences of the Otto-Friedrich-University Bamberg

Band 31

University of Bamberg Press
2018

Korpusbasierte Verfahren zur Generierung lexikalischer Ressourcen für das Opinion Mining

von Dirk Reinel

University of Bamberg Press
2018

Bibliographische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Informationen sind im Internet über http://dnb.ddb.de/ abrufbar.

Diese Arbeit hat der Fakultät Wirtschaftsinformatik und Angewandte Informatik der Otto-Friedrich-Universität Bamberg als Dissertation mit dem Titel „Korpusbasierte Verfahren zur Generierung lexikalischer Ressourcen für das Opinion Mining – Statistische Ansätze und deren Einsatzmöglichkeiten“ vorgelegen.
1. Gutachter: Prof. Dr. Andreas Henrich
2. Gutachter: Prof. Dr. Jörg Scheidt
Tag der mündlichen Prüfung: 26.04.2018

Herstellung und Druck: docupoint, Magdeburg
Umschlaggestaltung: University of Bamberg Press, Larissa Günther
Umschlagbild: © Dirk Reinel

ISSN: 1867-7401
ISBN: 978-3-86309-593-2 (Druckausgabe)
eISBN: 978-3-86309-594-9 (Online-Ausgabe)
URN: urn:nbn:de:bvb:473-opus4-523492
DOI: http://dx.doi.org/10.20378/irbo-52349

Results! Why, man, I have gotten a lot of results!
I know several thousand things that won't work.

— Thomas A. Edison [16]

DANKSAGUNG

Es war ein weiter Weg von der ersten Promotionsidee über die Konzeption und die Durchführung vieler Versuche bis hin zur fertigen Dissertation, der auch den ein oder anderen Rückschlag beinhaltete. Aus diesem Grund bin sehr dankbar, dass ich in dieser spannenden Zeit von vielen großartigen Menschen – fachlich wie emotional – begleitet wurde, die mich zu jeder Zeit und in jeder Phase unterstützt haben. Diesen Menschen möchte ich meinen Dank aussprechen.

Zuerst möchte ich mich bei meinem Doktorvater, Prof. Dr. Andreas Henrich, für die Betreuung meiner Arbeit, die konstruktiven Gespräche sowie seine umfassende Hilfe und Unterstützung während der Promotion bedanken. Außerdem möchte ich Prof. Dr. Jörg Scheidt meinen ganz besonderen Dank für die hilfreichen fachlichen Diskussionen, die gemeinsam entwickelten Ideen und für viele gute Ratschläge aussprechen, ohne ihn wäre ich nicht so weit gekommen. Außerdem danke ich Prof. Dr. Kai Fischbach für seine Bereitschaft in der Promotionskommission mitzuwirken.

Ich bedanke mich bei meinen früheren sowie den aktuellen Kollegen der Forschungsgruppe „Analytische Informationssysteme" – Johannes Drescher, Florian Wogenstein, Yannic Siebenhaar, Alexander Kern, und Prof. Dr. Sven Rill – für viele hilfreiche Diskussionen, anregende Kaffeerunden, eine stets angenehme Atmosphäre und das Korrekturlesen meiner Dissertation. Ein ganz

besonderer Dank geht an Niko Brucker, der mich hervorragend bei der Implementierung des in dieser Arbeit vorgestellten Systems unterstützt hat sowie Prof. Dr. Richard Göbel, der durch sein Engagement einen großen Anteil am Zustandekommen des Promotionsvorhabens hatte.
Zum Schluss möchte ich mich ganz herzlich bei meiner Familie – vor allem bei meinen Eltern Veronika und Dieter, meiner Schwester Jana, und meiner Cousine Manja – für die uneingeschränkte Unterstützung und den festen Rückhalt bedanken.

Mein besonderer Dank gilt Denise und Jakob für viel Geduld und Verständnis.

Ohne Euch wäre dies alles nicht möglich gewesen.

EIGENE VERÖFFENTLICHUNGEN

Sowohl im Vorfeld als auch bei der Erstellung dieser wissenschaftlichen Arbeit wurden Ergebnisse fortlaufend publiziert bzw. sind in Vorbereitung. Teile dieser Dissertation, die sich auf die entsprechenden Veröffentlichungen beziehen, sind an den jeweiligen Stellen als solche markiert. Nachfolgend eine Liste der Veröffentlichungen in chronologischer Reihenfolge.

1. Sven Rill, Sven Adolph, Johannes Drescher, Dirk Reinel, Jörg Scheidt, Oliver Schütz, Florian Wogenstein, Roberto V Zicari, and Nikolaos Korfiatis. A Phrase-Based Opinion List for the German Language. In *KONVENS*, pages 305–313, 2012. [47]

2. Sven Rill, Jörg Scheidt, Johannes Drescher, Oliver Schütz, Dirk Reinel, and Florian Wogenstein. A Generic Approach to Generate Opinion Lists of Phrases for Opinion Mining Applications. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 7, 2012. [48]

3. Dirk Reinel. Evaluation der Qualität lexikalischer Ressourcen zur Stimmungserkennung in literarischen Texten. In *LWA*, pages 168–172, 2013. [42]

4. Florian Wogenstein, Johannes Drescher, Dirk Reinel, Sven Rill, and Jörg Scheidt. Evaluation of an Algorithm for Aspect-Based Opinion Mining Using a Lexicon-Based Approach. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 5, 2013. [64]

5. Sven Rill, Dirk Reinel, Jörg Scheidt, and Roberto V Zicari. Politwi: Early Detection of Emerging Political Topics on Twitter and the Impact on Concept-Level Sentiment Analysis. *Knowledge-Based Systems*, volume 69, pages 24–33, 2014. [49]

6. Dirk Reinel and Jörg Scheidt. Automatische Auswertung von Kundenmeinungen – Opinion Mining am Beispiel eines Projekts für die Versicherungswirtschaft. In *Dialogmarketing Perspektiven 2014/2015*, pages 129–149, 2015. [43]

7. Dirk Reinel, Jörg Scheidt, Andreas Henrich, and Niko Brucker. Sentiment Phrase Generation Using Statistical Methods. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, 2018. (in Kürze erscheinend). [44]

VORBEMERKUNG

Das Forschungsgebiet *Opinion Mining* ist stark geprägt von englischsprachigen Veröffentlichungen und Fachbegriffen. Da diese nur selten geschickt übersetzt werden können, ohne dass die Übersetzung ins Deutsche „schwerfällig" und bemüht konstruiert wirkt, wird innerhalb der vorliegenden Forschungsarbeit größtenteils auf eine Übersetzung verzichtet und stattdessen die jeweilige englischsprachige Originalbezeichnung verwendet. Da sich die hier verwendeten Begriffe bereits etabliert haben und Teil der Fachtermini sind, ist dieses Vorgehen zudem hilfreich bei der Sichtung und dem Verständnis der referenzierten Fachliteratur.

Da die beiden Begriffe *Opinion Mining*[1] und *Sentiment Analysis*[2] häufig synonym verwendet werden – weitere, wenn auch deutlich seltener verwendete Begriffe für dieses Forschungsgebiet lauten *Sentiment Detection*, *Opinion Analysis*, *Emotion AI*, *Emotion Analysis*, *Opinion Extraction*, *Sentiment Mining*, *Review Mining*, und noch einige mehr –, wird in dieser Arbeit ausschließlich der Fachterminus *Opinion Mining* verwendet. Alle genannten Begriffe bezeichnen den Forschungsbereich, der sich mit der Analyse von Meinungen und Stimmungen in unstrukturierten Texten befasst.

1 Eingeführt in Dave et al. [14]
2 Eingeführt in Nasukawa and Yi [36]

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Listings

Algorithmenverzeichnis

1

EINLEITUNG

Das in dieser Arbeit in Teilen beleuchtete Forschungsgebiet *Opinion Mining* – in der Fachliteratur wird häufig der Begriff *Sentiment Analysis* synonym verwendet – beschäftigt sich mit Methoden zur automatischen bzw. semi-automatischen Identifikation und Extraktion von Meinungen und Stimmungen aus unstrukturierten Textdaten. Im Fokus der Forschungen stehen dabei Texte, die im Umfeld des Web 2.0 von Benutzern generiert werden, wie beispielsweise Kundenrezensionen zu Produkten und Hotels, Forenbeiträge oder Blogs.

Die Notwendigkeit der maschinellen Auswertung textueller Daten und die Zunahme der Forschungsaktivitäten im Umfeld des Opinion Mining in den letzten 15 Jahren hängen unmittelbar mit dem schrittweisen Übergang des World Wide Web (WWW) von einem Medium, bei dem die Nutzer Texte hauptsächlich konsumierten (Web 1.0), zu einem Medium, bei dem sich die Nutzer nun auch in der Rolle von Text-Produzenten wiederfanden (Web 2.0), zusammen. Durch diese Verschiebung der Rollen nahm – und nimmt – die Menge textueller Daten im WWW stark zu, etwa in Form der oben bereits erwähnten Kundenrezensionen.[1] Dadurch wuchsen gleichzeitig auch die Begehrlichkeiten, diese Daten, auch hinsichtlich der darin ausgedrückten Meinungen zu Unternehmen, Produkten, Dienstleistungen, Politik(ern) etc., auszuwerten.

Diese rasant wachsenden Datenmengen im Web 2.0 sind jedoch, hinsichtlich der Analysen der Inhalte, manuell nicht zu be-

1 Selbstverständlich trifft dies auch auf andere, von Usern generierte, Daten zu, wie etwa Videos.

wältigen. Ein gutes Beispiel dafür stellen die Bewertungen zum Smartphone *Samsung Galaxy S5*[2] auf der Plattform *Amazon.de* mit 1.066 Rezensionen dar (Stand: März 2017). Das Lesen aller Bewertungen, um die Meinungen der Nutzer in Erfahrung zu bringen und eine Kaufentscheidung zu treffen, würde einige Stunden in Anspruch nehmen.

An diesem Punkt setzen die Methoden des *Opinion Mining* an, die je nach Grad der Granularität (siehe Kapitel 2.2.2) Tendenzen oder sogar detaillierte – subsumierte – Auswertungen bereitstellen und die Dauer der Analysen stark verkürzen können.

1.1 Motivation und Einordnung der Arbeit

Das Ziel der Forschungsgruppe „Analytische Informationssysteme" des Instituts für Informationssysteme (iisys) der Hochschule Hof, innerhalb derer diese Arbeit in enger Zusammenarbeit mit der Otto-Friedrich-Universität Bamberg entstand, war von Beginn an, das Opinion Mining für die deutsche Sprache zu erweitern und zu verbessern. Neben der Entwicklung neuer Algorithmen zur Identifikation von meinungstragenden Wörtern sowie der Zuordnung der dazugehörigen Aspekte [64], stand dabei auch die Erzeugung neuer lexikalischer Ressourcen für die deutsche Sprache im Fokus [47]. Die Art, der Umfang und die Güte vorhandener Lexika meinungstragender Wörter, die für die von der Forschungsgruppe verwendeten Methoden des *Aspect-based Opinion Mining* benötigt wurden, erwiesen sich zu Beginn der Forschungstätigkeiten als nicht ausreichend. Kapitel 4.1 gibt einen Überblick über die Ideen und Annahmen, die in der Forschungsgruppe entwickelt wurden und die Basis für die Entwicklung einer eigenen lexikalischen Ressource, der *Sentiment Phrase List* (SePL), bildeten. Dieses neuartige Lexikon war die erste

2 https://www.amazon.de/dp/B00IKFB4OS/

deutschsprachige Ressource, welche automatisch durch Verwendung eines speziellen Textkorpus generiert wurde und die, neben meinungstragenden Wörtern, auch Phrasen enthielt (siehe Kapitel 2.3.2.4). Eine detaillierte Beschreibung für die Generierung der deutschsprachigen *Sentiment Phrase List* wird in der eigenen Veröffentlichung [48] gegeben. Die neuen Ideen zur Generierung lexikalischer Ressourcen legten schließlich den Grundstein für die vorliegende Forschungsarbeit.

Nach anfänglicher Prüfung vorhandener Lexika meinungstragender Wörter, mit dem vormaligen Ziel deutschsprachige literarische Texte mit den Methoden des Opinion Mining zu analysieren, wurde deutlich, dass Lexika meinungstragender Wörter und Phrasen zwar durchaus geeignet sind, Texte dieser Art zu analysieren, vorhandene deutschsprachige lexikalische Ressourcen jedoch nicht den speziellen Anforderungen an die Analyse genügten [42]. Anhand der Ergebnisse dieser Untersuchung entstand sukzessive die Idee für eine neue lexikalische Ressource bzw. für einen Algorithmus zur Erstellung dieser, der die Vorteile der *Sentiment Phrase List* übernehmen, jedoch sprachunabhängig[3] und nur auf Basis statistischer Methoden meinungstragende Wörter und Phrasen automatisch aus geeigneten Korpora ableiten sollte. Außerdem sollte der neu entwickelte Ansatz in der Lage sein, meinungstragende Redewendungen, wie z.B. „Es ist nicht alles Gold, was glänzt“, zu extrahieren. Durch diese Redewendungen werden, besonders im Bereich der Rezensionen, Meinungen ausgedrückt. Eine automatische Generierung gestaltet sich jedoch schwierig, da für diese Wortverbindungen keine Muster gebildet werden können und vorhandene, oft Pattern-basierte, Algorithmen somit nicht in der Lage sind, diese zu identifizieren.

3 Das bedeutet, Verzicht auf gängige *Natural language processing* (NLP)-Methoden, wie POS-Tagging, Lemmatisierung etc.

1.2 Problemstellungen und Forschungslücke

Die Legitimation für die Entwicklung einer neuen lexikalischen Ressource bzw. eines entsprechenden Algorithmus zur Erzeugung dieser, ergibt sich aus den folgenden vier Punkten, die die Einschränkungen vorhandener Lexika und Erzeugungsmethoden verdeutlichen sollen:

1. **Lücken für viele Sprachen**: Lexikalische Ressourcen für das Opinion Mining existieren vor allem für die englische Sprache (siehe Tabelle 1 in Kapitel 2.3.2), da ein Großteil der Forschungsarbeit für diese Sprache geleistet wird. Für viele andere Sprachen – auch für das Deutsche – existieren nur wenige oder überhaupt keine Lexika. Die Adaption vorhandener Algorithmen sowie die maschinelle Übersetzung existierender Ressourcen funktionieren nur teilweise und mit schlechten Ergebnissen.

2. **Probleme beim Umgang mit *Valence Shifter Words***: Die meisten verfügbaren lexikalischen Ressourcen – unabhängig von der Sprache – enthalten meinungstragende Worte und eine dazugehörige Klassifikation (z.B. positiv, neutral, negativ). Bei der späteren Anwendung dieser Lexika, d.h. bei der Analyse von Texten mit dem Ziel, Meinungen zu extrahieren, kommt es zu Problemen mit *Valence Shifter Words*. Diese Wörter stehen meist vor den eigentlichen meinungstragenden Begriffen als a) Verstärker b) Abschwächer oder c) Negationen. Beispiele dafür sind „sehr“, „absolut“ und „nicht“. Der Umgang mit diesen Begriffen ist nicht trivial und es existieren keine Muster der Art: *Das Wort „nicht“ invertiert stets den Meinungswert bzw. die Meinungsklasse eines meinungstragenden Wortes*. Deutlich wird dies bei der Betrachtung der beiden Adjektive „gut“ und „perfekt“.

Während „nicht gut“ nahezu das Gegenteil von „gut“ ausdrückt, wird die ausgedrückte Meinung bei Verwendung von „nicht perfekt“ im Gegensatz zu „perfekt“ nur etwas abgeschwächt[4].

3. **Einschränkungen bei automatischen Verfahren**: Vorhandene Verfahren zur Generierung lexikalischer Ressourcen benötigen, je nach Methode, oft einen *Seed* aus meinungstragenden Wörtern in der jeweiligen Sprache und sind, durch Verwendung von *Natural language processing (NLP)*-Methoden, wie *Part-of-speech Tagging* und *Stemming* bzw. *Lemmatisierung*, zudem stark an die jeweilige Sprache gebunden.

4. **Fehlen wichtiger Wortverbindungen, wie etwa Redewendungen**: Durch die Beschränkung auf Einzelwörter und die Einschränkungen durch Pattern-basierte Verfahren, die zwar bestimmte Wortphrasen finden und klassifizieren können, bei komplexeren Wortverbindungen wie Redewendungen jedoch scheitern, fehlen häufig wichtige Phrasen in diesen Lexika.

1.3 Ziele der Arbeit

Das Hauptziel dieser Arbeit ergibt sich aus der Forschungslücke, die in Abschnitt 1.2 aufgezeigt wurde, und adressiert die Entwicklung einer neuen Methode, die es ermöglichen soll, Lexika meinungstragender Wörter und Phrasen sprachunabhängig, das heißt ohne Einsatz von NLP-Methoden, automatisch und ohne Angabe eines Seeds zu erzeugen. Durch den Einsatz verschiedener statistischer Methoden wird angestrebt, die Beschränkung der Sprache durch Wortarten, Wortformen, Satzbau, Pattern etc.

4 Eine detaillierte Untersuchung dieses Sachverhaltes findet sich in der Dissertation von Rill [46, S. 62–64].

aufzuheben und so neue meinungstragende Phrasen zu finden und in die Ressource zu integrieren. Des Weiteren soll der entwickelte Algorithmus prototypisch umgesetzt werden, um damit eine lexikalische Ressource für die deutsche Sprache zu erzeugen und diese zu evaluieren. Voraussetzung für die Generierung einer solchen Ressource ist ein Textkorpus mit bestimmten Eigenschaften, wie er schon bei der Erzeugung der *Sentiment Phrase List* (siehe Kapitel 2.3.2.4) verwendet wurde. Dieser Korpus muss aus mehreren getrennten Texteinheiten (z.B. Sätzen) bestehen, denen eine numerische Bewertung – beispielsweise in Form von „Punkten“ oder „Sternen“, wie häufig in Bewertungsportalen zu finden – zugeordnet wurde. Durch Verwendung eines so aufgebauten Korpus soll die automatische Generierung einer neuen lexikalischen Ressource für das Opinion Mining, durch den Einsatz statistischer Methoden, ermöglicht werden.

Ein weiteres Hauptziel, das sich im Laufe der Forschungsarbeit ergab, adressiert die manuelle Erzeugung eines Referenzlexikons meinungstragender deutscher Adjektive. Da für die Evaluation deutschsprachiger Meinungslexika bislang keine Referenz in Form einer (manuell) überprüften Liste relevanter Adjektive existierte und sich der Vergleich mit vorhandenen lexikalischen Ressourcen, aufgrund des unterschiedlichen Aufbaus (siehe Kapitel 2.3.2) als schwierig erwies, ist diese Ressource essentiell für die Evaluation des neu entwickelten automatischen Verfahrens. Zudem stellt diese ein eigenständiges, sehr gutes und überprüftes Lexikon häufiger deutscher meinungstragender Adjektive dar.

1.4 Aufbau der Arbeit

Die vorliegende Arbeit ist in acht Kapitel unterteilt, welche im Folgenden – exklusive **Kapitel 1** *Einleitung* – kurz beschrieben werden.

Kapitel 2 *Stand der Forschung* beinhaltet den aktuellen Stand der Forschung im Bereich des Opinion Mining und gibt zudem einen detaillierten Überblick über bereits vorhandene lexikalische Ressourcen sowie deren Vor- und Nachteile und beleuchtet die Methoden der Generierung solcher Ressourcen.

Kapitel 3 *Verwendete Maße und Methoden* umfasst einen Überblick über die, in der Arbeit verwendeten, statistischen Maße und Methoden.

Kapitel 4 *Generierung lexikalischer Ressourcen* bildet den Kern der Arbeit. In diesem Kapitel wird der erste Schwerpunkt der Forschungsarbeit beschrieben, der entwickelte Algorithmus zur automatischen Generierung einer neuen lexikalischen Ressource für das Opinion Mining. Dabei werden zunächst die grundlegenden Ideen und Annahmen aufgezeigt. Anschließend folgt der Aufbau des Systems sowie eine detaillierte Beschreibung der Implementierung. Nach Experimenten zu Laufzeitmessung, Speicherbedarf und Schwellenwertbestimmung der benötigten Parameter, werden Ergebnisse der neu erzeugten lexikalischen Ressource für die deutsche Sprache präsentiert.

Kapitel 5 *Manuelle Erzeugung eines Referenzlexikons* adressiert den zweiten Schwerpunkt dieser Arbeit und beinhaltet alle wichti-

gen Schritte beim Aufbau eines Referenzlexikons für die deutsche Sprache.

Kapitel 6 *Evaluation* beschreibt die Methoden zur Evaluation der neu erzeugten lexikalischen Ressource sowie der weiteren deutschen Ressourcen. Im Anschluss werden die Ergebnisse dieser Evaluierungen präsentiert und diskutiert.

Kapitel 7 *Diskussion möglicher Anwendungsszenarien* beinhaltet Diskussionen bezüglich der praktischen Verwendung der neuen lexikalischen Ressource sowie des möglichen Einsatzes des entwickelten Algorithmus in anderen Forschungsgebieten.

Kapitel 8 *Diskussion der Ergebnisse* bildet den Abschluss dieser Arbeit und setzt sich intensiv mit den erzielten Ergebnissen sowie weiteren möglichen Ansatzpunkten auseinander.

2

STAND DER FORSCHUNG

Im folgenden Kapitel wird zunächst ein kurzer Überblick über die verwendete Terminologie (Kapitel 2.1) gegeben, bevor in Kapitel 2.2 der aktuelle Stand der Forschung im Bereich *Opinion Mining* beleuchtet wird. Dabei werden die Methoden des Opinion Mining in Kapitel 2.2.1 sowie die verschiedenen Ebenen, auf denen die Meinungsanalyse durchgeführt werden kann, in Kapitel 2.2.2 näher betrachtet.

Anschließend werden in einem eigenen Abschnitt die Grundlagen und der Stand der Forschung zu lexikalischen Ressourcen, d.h. Lexika meinungstragender Wörter und Phrasen, in diesem Forschungsfeld vorgestellt (Kapitel 2.3). Da die Erzeugung sowie die Evaluation solcher Ressourcen den Kern dieser Dissertation bilden, werden in Kapitel 2.3.1 zunächst die Methoden zur Generierung von Meinungslexika vorgestellt. Anschließend wird in Kapitel 2.3.2 eine detaillierte Übersicht über vorhandene Lexika gegeben.

2.1 Terminologie

Innerhalb dieser Arbeit werden verschiedene Fachbegriffe verwendet, die im Folgenden kurz eingeführt werden. Die Anordnung erfolgt dabei in alphabetischer Reihenfolge.

ASPECT-BASED OPINION MINING Als *Aspect-based Opinion Mining* wird eine feingranulare Methode des *Opinion Mining* bezeichnet, die auf der untersten Textebene (der *Aspekt*ebene) operiert und durch die Erzeugung von *Opinion Quintuples*

sehr exakte Analysen zu den im Text ausgedrückten *Meinungen* zulässt.

ASPEKT *Aspekte* bezeichnen beim *Aspect-based Opinion Mining* Eigenschaften einer *Entität*, die durch entsprechende Wörter ausgedrückt werden. Beispiele sind das Display eines Smartphones oder der Service eines Unternehmens.

ENTITÄT *Entitäten* bezeichnen im Umfeld des *Aspect-based Opinion Mining* die betrachteten Objekte, über die in den zu untersuchenden Texten berichtet wird. Die *Entität* ist damit die höchste Instanz, unter die alle *Aspekte* eingehängt werden. Beispiele für *Entitäten* sind Unternehmen, Produkte oder Personen.

INTENSITÄT (EINER MEINUNG) Als *Intensität* wird die Stärke einer *Meinung*, bestimmt durch das *meinungstragende Wort* bzw. die *meinungstragende Phrase*, bezeichnet.

LEMMATISIERUNG Rückführung eines Wortes auf dessen Grundform. Ein Beispiel dafür ist die Rückführung von „guter“ auf „gut“.

LEXIKALISCHE RESSOURCE Im Kontext des *Opinion Mining* werden alle Listen bzw. Lexika mit *Aspekten*, *Entitäten*, oder *meinungstragenden Wörtern und Phrasen* als *lexikalische Ressourcen* bezeichnet. Im Umfeld dieser Arbeit, die sich primär mit Lexika meinungstragenden Wörter und Phrasen beschäftigt, wird der Begriff *Lexikalische Ressource* jedoch synonym für diese Lexika verwendet.

MEINUNG Im Kontext des *Aspect-based Opinion Mining* wird eine Meinung immer erst durch den Bezug eines *meinungstragenden Wortes* auf einen *Aspekt* oder eine *Entität* ausgedrückt. Dazu wird bei der Analyse pro Meinungsäußerung

ein *Opinion Quintuple* erzeugt, welches alle relevanten Informationen enthält. Die *Polarität* sowie die *Intensität* der Meinung wird dabei durch das jeweilige *meinungstragende Wort* bzw. durch die jeweilige *meinungstragende Phrase* bestimmt.

MEINUNGSTRAGENDE PHRASE Als *meinungstragende Phrasen* werden Gruppen von Wörtern bezeichnet, mit deren Hilfe Meinungen ausgedrückt werden können. Diese enthalten oft Negationen („nicht“), Verstärker („sehr“) oder Abschwächer („fast“), wodurch *Polarität* sowie *Intensität* beeinflusst werden. Durch die Verwendung von *meinungstragenden Phrasen* können *Opinion Mining* Algorithmen verbessert werden, da die (schwierige) Behandlung der oben erwähnten *Valence Shifter Words* entfällt. Beispiele für solche Wortverbindungen sind „sehr gut“, „nicht gut“ und „einfach nur schlecht“, wobei Kombinationen verschiedener Wortformen in Frage kommen.

MEINUNGSTRAGENDES WORT Als *meinungstragendes Wort* wird ein einzelnes subjektives Wort bezeichnet, mit dessen Hilfe eine Meinung ausgedrückt werden kann. Beispiele für solche Wörter sind die Adjektive „gut“ und „schlecht“, wobei auch andere Wortformen in Frage kommen.

n-GRAMM Im Kontext dieser Arbeit werden *n*-Gramme als eine Aneinanderreihung von *n* einzelnen Wörtern bezeichnet. Dementsprechend bezeichnen Unigramme einzelne Wörter, Bigramme 2-Wort-Kombinationen, Trigramme 3-Wort-Kombinationen etc.

OPINION MINING Ein Forschungsgebiet, das alle Algorithmen und Methoden zur Identifikation und Extraktion von *Meinungen* aus unstrukturierten Texten umfasst (siehe Kapitel 2.2).

OPINION QUINTUPLE *Opinion Quintuples* werden beim *Aspect-based Opinion Mining* eingesetzt und bilden als kleinste Einheit jeweils die Meinung eines Autors ab. Ein *Opinion Quintuple* besteht dabei aus *Entität*, *Aspekt*, Meinungswert bzw. Meinungsklasse (siehe *Sentiment Value*), Autor, und dem Zeitpunkt der Meinungsäußerung.

PART-OF-SPEECH TAGGING Dient der Bestimmung von Wortarten in Texten und bildet dabei beispielsweise die Grundlage für die *Lemmatisierung*.

POLARITÄT (EINER MEINUNG) Als *Polarität* wird die Orientierung einer Meinung, bestimmt durch das *meinungstragende Wort* bzw. die *meinungstragende Phrase*, bezeichnet. Mögliche Ausprägungen sind „positiv“, „neutral“ und „negativ“.

SENTIMENT VALUE Als *Sentiment Value* – alternativ: *Opinion Value* – wird der Meinungswert eines *meinungstragende Wortes* bzw. einer *meinungstragenden Phrase* bezeichnet. Dieser bestimmt, welche *Polarität* und *Stärke* der meinungstragende Begriff besitzt. Im Kontext dieser Arbeit liegen die Meinungswerte auf einer kontinuierliche Skala $[-1, +1]$.

VALENCE SHIFTER WORDS *Valence Shifter Words* oder *Meinungsshifter* sind Wörter, die sowohl die *Polarität* als auch die *Intensität* eines *meinungstragenden Wortes* beeinflussen können. Arten und Beispiele für solche Wörter sind Negationen („nicht“), Verstärker („sehr“) und Abschwächer („fast“).

2.2 Opinion Mining

Das Forschungsgebiet *Opinion Mining* umfasst im Wesentlichen alle Algorithmen und Methoden, die der Identifikation und Extraktion von Meinungen aus unstrukturierten Texten, mit dem

Ziel der Meinungsanalyse, dienen. Hinter dieser trivial erscheinenden Aufgabe verbirgt sich jedoch eine Vielzahl komplexer Teilaufgaben, für die in den vergangen 15 Jahren bereits wichtige Forschungsarbeit – vor allem für die englische Sprache – geleistet wurde und bis heute geleistet wird. Einige wichtige Teilaufgaben des Opinion Mining werden im Folgenden aufgeführt:

1. Identifikation und Bewertung meinungstragender Begriffe
2. Identifikation und Extraktion von Aspekten und Entitäten
3. Zuordnung meinungstragender Begriffe zu Aspekten
4. Identifikation und Behandlung verschiedener Typen von Meinungen, wie z.B. normale (*regular*) Meinungen im Gegensatz zu vergleichenden (*comparative*) Meinungen oder subjektive (*subjective*) Meinungen im Gegensatz zu faktologischen (*fact-implied*) Meinungen
5. Auflösung von Koreferenzen

Hinzu kommen sprachliche Charakteristika, die von besonderer Bedeutung sind, sobald der englische Sprachraum verlassen wird und Texte in anderen Sprachen analysiert werden sollen. Für die englische Sprache existieren bereits gut funktionierende Opinion Mining Algorithmen sowie lexikalische Ressourcen. Zwar lassen sich viele Ergebnisse übertragen, jedoch gibt es in jeder Sprache Besonderheiten, die berücksichtigt werden müssen, um optimale Analyseergebnisse zu erzielen. Im Deutschen sind solche Besonderheiten beispielsweise die Verwendung von Nominalkomposita, die nicht nur das Auffinden von Aspekten („Worauf bezieht sich eine Meinungsäußerung?“) sondern auch von Meinungsäußerungen erschwert, was am Beispiel des zusammengesetzten Nomens „Empfangsproblem“ deutlich wird, bei

dem sowohl Aspekt („Empfang“) als auch die (faktologische) Meinungsäußerung („Problem“) in einem Wort ausgedrückt wird. Eine weitere Besonderheit der deutschen Sprache ist die Verwendung von verschachtelten Nebensätzen, durch die eine automatische Textanalyse erschwert wird.

Obwohl ein genereller Überblick über die verschiedenen Teilaufgaben und Herausforderungen des Opinion Mining wichtig für die Einordnung dieser Arbeit innerhalb des umfassenden Themengebietes ist, liegt der Fokus auf der Erzeugung neuer lexikalischen Ressourcen, d.h. Listen mit meinungstragenden Begriffen, die für Lexikon-basierte Analyseverfahren benötigt werden. Diese Arbeit beschäftigt sich somit ausschließlich mit der oben genannten Teilaufgabe *Identifikation und Bewertung meinungstragender Begriffe*.

Bei der Meinungsanalyse von Texten wird sowohl zwischen dem Detaillierungsgrad bzw. der Granularität der Analyse, die auf Dokumentenebene, Satzebene oder Aspektebene stattfinden kann, als auch der Art der Analyse unterschieden. Die grundlegenden Methoden sind hierbei die Klassifikation von Meinungen mittels Machine-Learning Algorithmen sowie die Klassifikation durch Lexikon-basierte Verfahren. Einen guten Überblick über die gesamte Thematik gibt Liu [31].

In den folgenden Abschnitten werden die entsprechenden Methoden (Abschnitt 2.2.1) sowie die genannten Ebenen des Opinion Mining (Abschnitt 2.2.2) näher betrachtet.

2.2.1 *Methoden des Opinion Mining*

Es existieren verschiedene Methoden und Ansätze zur Klassifikation von Meinungen in Texten, einen guten Überblick dazu geben Cambria et al. [8] und Liu [31]. Die beiden grundlegenden Me-

thoden, welche in Forschung und Praxis zum Einsatz kommen, werden im Folgenden kurz vorgestellt.

2.2.1.1 *Verfahren des Maschinellen Lernens*

Verfahren des *Maschinellen Lernens* basieren auf Systemen, die darauf trainiert werden, Klassifikationsprobleme zu lösen. Nach der Trainingsphase sind diese Systeme in der Lage, neue Daten und Objekte automatisch den trainierten Klassen (z.B. positiv und negativ) zuzuordnen. Dabei werden – je nach Sprache und Granularität – oft hohe Genauigkeiten erreicht.

Diese Methode des Opinion Mining wird häufig beim *Document-Level Opinion Mining* (siehe Abschnitt 2.2.2.1) eingesetzt. Das System lernt dabei aus Erfahrungen (Training) und setzt das erlernte Wissen ein, um neue Informationen zu extrahieren. Für die Lernphase solcher Systeme werden allerdings Trainingsdaten benötigt, die manuell erstellt werden müssen und die meist nur für spezifische Domänen verwendbar sind. Bezogen auf das Forschungsgebiet des Opinion Mining wird ein so trainiertes System eingesetzt, um beispielsweise Dokumente [38] oder Sätze [62] in vorher definierte Klassen (z.B. positiv, negativ) einzuteilen. Pang et al. [38] nutzen diese Methode dabei erstmals im Bereich des Opinion Mining, um Filmbewertungen des Internet Movie Database (IMDb) Archivs automatisch den Klassen *positiv* und *negativ* zuzuordnen.

Die eigentliche Umsetzung der zuvor beschriebenen Vorgehensweise maschinenlernender Verfahren geschieht mittels verschiedener Algorithmen. Wichtige überwachte Lernverfahren sind dabei *Naive Bayes*, der beispielsweise im Bereich Spamfilterung zum Einsatz kommt (Bayes-Klassifikator), *Maximum Entropy* sowie *Support Vector Machines*, die häufig für Klassifikationsaufgaben im Bereich des Opinion Mining eingesetzt werden.

2.2.1.2 *Lexikon-basierte Verfahren*

Im Gegensatz zu den Verfahren des *Maschinellen Lernens* verwenden Lexikon-basierte Verfahren lexikalische Ressourcen, d.h. Listen mit meinungstragenden Begriffen, um Meinungen zu identifizieren und die entsprechenden Texte danach zu klassifizieren. Diese Ressourcen existieren für verschiedene Sprachen, wobei ein Großteil jedoch für die englische Sprache zur Verfügung steht.
Die Lexikon-basierte Methode wird häufig für die Meinungsanalyse mit dem höchsten Detaillierungsgrad, dem *Aspect-based Opinion Mining* (siehe Abschnitt 2.2.2.3), verwendet.

Lexikalische Ressourcen als wichtigste Voraussetzung für diesen Ansatz müssen in der jeweiligen Sprache vorliegen, in denen die zu analysierenden Texte verfasst wurden. Ein Problem bei der Verwendung solcher Ressourcen, die meist nur einzelne meinungstragende Wörter enthalten, ist die Behandlung von Wörtern, die Meinungsäußerungen verstärken, abschwächen oder negieren. Verdeutlicht wurde dies bereits in Kapitel 1.2 (Punkt 2: *Probleme beim Umgang mit Valence Shifter Words*) an dem Negationswort „nicht“, welches, je nach Polarität und Intensität eines meinungstragenden Wortes, Meinungsäußerungen sowohl invertieren als auch lediglich etwas abschwächen kann.

Ein weitere Herausforderung dieses Ansatzes ist die Erkennung und Behandlung von meinungstragenden Begriffen, die je nach Domäne unterschiedliche Meinungen ausdrücken. Ein Beispiel für einen solchen Begriff ist das Adjektiv „gruselig“, das in der Domäne „Bücher“ fast ausschließlich für positive Meinungsäußerungen verwendet wird („Das neue Buch von Stephen King war ja wieder richtig gruselig!“). In der Domäne „Elektronik“ wird dasselbe Adjektiv jedoch dazu genutzt, um negative Meinungen auszudrücken („Das Design des neuen MacBook ist wirklich gruselig.“).

2.2.2 *Ebenen des Opinion Mining*

In diesem Abschnitt werden die drei verschiedenen Ebenen vorgestellt, auf den Opinion Mining Methoden ansetzen können. Die jeweilige Granularität der Analyse und die Aussagekraft der Ergebnisse wird dabei anhand eines durchgehenden Beispiels verdeutlicht.

2.2.2.1 *Dokumentenebene*

Beim *Document-Level Opinion Mining*, der Meinungsanalyse auf Dokumentenebene, werden komplette Dokumente nach deren Tonalität, beispielsweise in *positiv*, *negativ* und *neutral*, eingeteilt. Solche Dokumente können z.B. Rezensionen, E-Mails, Beiträge aus Foren oder Nachrichtenartikel sein. Abbildung 1 soll die Analyse eines Textes auf dieser Ebene verdeutlichen. Der Text wird dabei einfach als Ansammlung von Wörtern (*Bag-of-Words*) betrachtet, wobei zunächst mittels einer geeigneten lexikalischen Ressource *positive* (grün) und *negative* (rot) Wörter und Phrasen extrahiert werden. Anschließend werden diese meinungstragenden Begriffe gezählt. Enthält der Text mehr positive als negative Meinungsäußerungen wird er insgesamt als *positiv* klassifiziert und umgekehrt. Im Beispiel werden die vier Begriffe „einfach toll“, „einfache“, „starke“ und „große“ als *positive* Meinungsäußerungen erkannt, die zwei Begriffe „unverschämt“ und „sehr unfreundlich“ als *negative*. Der Text wird in seiner Gesamtheit somit *positiv* klassifiziert.

Dieses Verfahren stößt jedoch schnell an seine Grenzen. Sobald Texte komplexer werden und Aussagen zu den nicht relevanten Aspekten überwiegen, liefert die Analyse auf dieser Ebene falsche Ergebnisse.

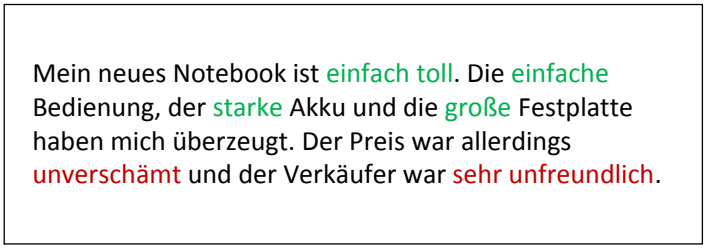


Abbildung 1: Opinion Mining auf Dokumentenebene und Satzebene

2.2.2.2 *Satzebene*

Für eine granularere Betrachtung kann Opinion Mining auch auf Satzebene stattfinden, es wird dann vom *Sentence-Level Opinion Mining* gesprochen. Die Vorgehensweise bei der Analyse ist analog zur Dokumentenebene, mit dem Unterschied, dass der Text nicht mehr als Ansammlung von Wörtern ohne jede Struktur, sondern von Sätzen und ggf. von Abschnitten, betrachtet wird. Für diese Art der Analyse wird der Text, beispielsweise durch Verwendung eines *Sentence Tokenizers*, zunächst in Sätze zerlegt und diese anschließend analysiert. Am Beispiel des Textes in Abbildung 1 werden die drei Sätze „Mein neues Notebook ist einfach toll.“, „Die einfache Bedienung, der starke Akku und die große Festplatte haben mich überzeugt.“ und „Der Preis war allerdings unverschämt und der Verkäufer war sehr unfreundlich.“ extrahiert und entsprechend der positiven und negativen Begriffe klassifiziert. Dabei werden die beiden ersten Sätze als *positiv* und der letzte Satz als *negativ* eingestuft.

Analog zum *Document-Level Opinion Mining* erreicht dieses Verfahren seine Grenzen, sobald Texte komplexer werden und in einzelnen Sätzen Meinungen gegensätzlicher Polaritäten zum Ausdruck gebracht werden. Auch Äußerungen zu nicht relevanten Aspekten führen hier wieder zu Problemen.

2.2.2.3 *Aspektebene*

Der höchste Detaillierungsgrad bei der Analyse von Meinungen in Texten wird durch die Methode des *Aspect-based Opinion Mining* erreicht. Beschrieben wurde diese Methode erstmals von Hu and Liu [25] unter dem Begriff *Feature-based Opinion Mining*. Bei diesem Verfahren werden Meinungen in Form von *Opinion Quadruples* [25] bzw. *Opinion Quintuples* [29, 30, 31] aus Texten extrahiert und können so separat bzw. aggregiert betrachtet werden. Abbildung 2 illustriert den Ablauf beim *Aspect-based Opinion Mining*, bei welchem aus einem gegebenen Dokument nacheinander Entitäten, Aspekte, meinungstragende Wörter und Phrasen und ggf. Autor(en) sowie der Zeitpunkt der Meinungsäußerung extrahiert werden. Anschließend werden mittels eines weiteren Algorithmus die meinungstragenden Wörter und Phrasen den jeweiligen Aspekten und Entitäten zugeordnet und die Ergebnisse in Form von *Opinion Quintuples* ausgegeben.

Opinion Quintuples sind die am häufigsten verwendete Form bei der Meinungsanalyse und kommen auch in praktischen Anwendungen zum Einsatz. Ein *Opinion Quintuple (e, a, s, h, t)* besteht dabei aus den folgenden Elementen:

- Entität **e**: Als *Entität* wird das betrachtete Objekt des zu untersuchenden Textes bezeichnet. Üblicherweise bilden Entitäten Unternehmen, Produkte oder Personen ab.

- Aspekt **a** der Entität e: *Aspekte* sind Komponenten oder Eigenschaften der Entität, wie beispielsweise der Akku (des Produktes X) oder die Kundenfreundlichkeit (des Unternehmens Y).

- Meinung **s** zum Aspekt a der Entität e: Der Meinungswert bzw. die Meinungsklasse des meinungstragenden Begriffs, der sich auf den Aspekt der Entität bezieht. Dieser wird, je nach Algorithmus und Lexikon, entweder als Klasse (z.B.

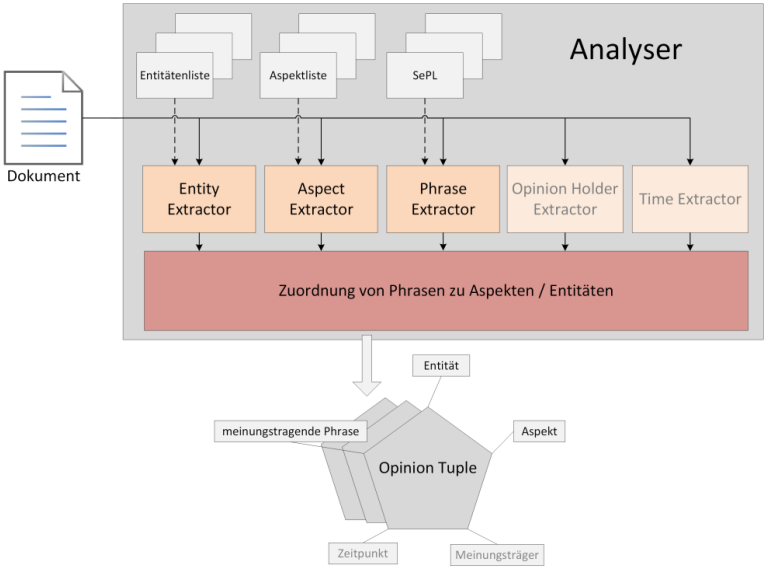


Abbildung 2: Analyseprozess beim *Aspect-based Opinion Mining*

„positv“, „negativ“) oder auf einer kontinuierlichen Skala (z.B. $[-1,0;+1,0]$) angegeben. Ein Beispiel ist die positive (Meinungsklasse) Bewertung des Akkus (Aspekt) des iPhone (Entität).

- Autor **h** der Meinungsäußerung: Dieses Element beinhaltet den Namen des Autors – Username, Pseudonym, Hash, oder Klarname – der Meinungsäußerung, der beispielsweise von einem entsprechenden Metatag des Textes stammt oder direkt aus dem Text extrahiert wurde.

- Zeitpunkt **t** der Meinungsäußerung: Datum und, falls vorhanden, die Uhrzeit der Meinungsäußerung. Diese Information wird ebenfalls aus den Metadaten des Textes oder direkt aus diesem extrahiert.

Abbildung 3 verdeutlicht den hohen Detaillierungsgrad dieser Analysemethode. Im Vergleich zu den vorhergehenden Ver-

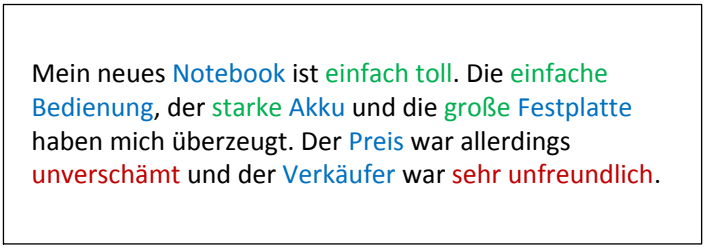


Abbildung 3: Opinion Mining auf Aspektebene

fahren wird deutlich, wie exakt die Analyse in diesem Beispiel funktioniert und welche Auswertungen unter optimalen Bedingungen möglich sind. Allerdings sind solche Bedingungen in der Praxis selten. Vielmehr müssen vor der eigentlichen Analyse oft entsprechende Vorverarbeitungsschritte bezüglich fehlerhafter Rechtschreibung, fehlender Satz- und Leerzeichen etc. durchgeführt werden, um optimale Ergebnisse zu erzielen.

Die folgenden sechs Meinungen, bestehend aus Entität bzw. Aspekt (in Abbildung 3 *blau* dargestellt) sowie der dazugehörigen Meinungsklasse der Meinungsäußerungen – *positiv* (in Abbildung 3 *grün*) und *negativ* (in Abbildung 3 *rot*) –, ausgedrückt durch meinungstragenden Wörter und Phrasen, können nun aus dem Text extrahiert werden:[1]

- Notebook → positiv
- Bedienung → positiv
- Akku → positiv
- Festplatte → positiv
- Preis → negativ
- Verkäufer → negativ

1 Aus Gründen der Übersichtlichkeit wird in diesem Beispiel auf vollständige Quintuple verzichtet, d.h. Autor und Zeitpunkt werden nicht berücksichtigt.

2.3 Lexikalische Ressourcen

Im Umfeld des *Opinion Mining* wird der Begriff *lexikalische Ressource*, neben Textressourcen für Aspekte und Entitäten beim *Aspect-based Opinion Mining*, vor allem für Listen bzw. Lexika verwendet, in denen meinungstragende Wörter und – seltener – Phrasen zusammengefasst werden. Beispiele für solche Begriffe geben die Adjektive „gut" und „schlecht" sowie die 2-Wort-Phrase „nicht perfekt". Durch diese Wörter und Phrasen können positive, negative und neutrale Meinungen ausgedrückt werden[2]. Lexikalische Ressourcen kommen bei fast allen *Opinion Mining* Methoden zum Einsatz und bilden damit eine wichtige Grundlage der Meinungsextraktion. Allerdings ist eine gute lexikalische Ressource allein kein Garant für zufriedenstellende Ergebnisse bei der Extraktion von Meinungen. Auch die dahinter liegenden Algorithmen müssen auf das Lexikon und den zu analysierenden Text abgestimmt sein. Vorhandene Ressourcen unterscheiden sich – neben der Sprache – vor allem durch die Anzahl der Einträge, die Verwendung von Einzelwörtern und Phrasen sowie durch den Wertebereich der Polarität. Auch wurden verschiedene Verfahren bei der Erzeugung solcher Listen angewandt.

In den folgenden Abschnitten werden zunächst die drei wichtigsten Methoden für die Generierung lexikalischer Ressourcen beschrieben. Anschließend werden vorhandene Ressourcen für verschiedene Sprachen vorgestellt. Dabei werden die deutschen Meinungslexika im Detail betrachtet, da diese für die in der Arbeit beschriebene Ressource (siehe Kapitel 4) besondere Relevanz besitzen und zusammen mit dieser in Kapitel 6 evaluiert werden.

2 Eine Meinung wird durch den Bezug eines meinungstragenden Wortes bzw. Phrase auf einen Aspekt oder eine Entität gebildet (siehe *Opinion Quintuple*).

2.3.1 *Methoden der Generierung*

Für die Erzeugung von Meinungslexika – als wichtigste lexikalische Ressource des Opinion Mining – existieren drei verschiedene Methoden, beschrieben von Liu [31]. In diesem Abschnitt werden diese Methoden vorgestellt.

2.3.1.1 *Manuelle Erzeugung*

Die manuelle Erzeugung von Lexika meinungstragender Begriffe ist extrem arbeits- und zeitintensiv. Dazu wird eine geeignete Anzahl von Wörtern von Annotatoren in verschiedene Meinungsklassen eingeteilt. Dabei werden häufig 3-Klassen-Systeme (positiv, neutral, negativ) oder 5-Klassen-Systeme (stark positiv, schwach positiv, neutral, schwach negativ, stark negativ) verwendet. Da die Erstellung solcher Listen sehr aufwendig ist, werden diese üblicherweise a) als Seed für die anderen beiden Verfahren (kleines Set) oder b) für die Evaluation automatischer Methoden verwendet. Dieses Verfahren wird in Tabelle 1 auf Seite 27, innerhalb derer eine Übersicht erzeugter Ressourcen enthalten ist, mit „ma“ (*manual approach*) abgekürzt.

2.3.1.2 *Wörterbuch-basierte Erzeugung*

Die grundlegende Idee des Wörterbuch-basierten Ansatzes ist die Verwendung eines Seeds, bestehend aus meinungstragenden Wörtern, mit dessen Hilfe semantisch verwandte Begriffe aus Wörterbüchern bzw. Wortnetzen extrahiert werden. Dazu werden zu allen Wörtern des Seeds iterativ Synonyme und ggf. Antonyme gesucht und diesen der gleiche Meinungswert, die gleiche Meinungsklasse oder Polarität (je nach Seed), bzw. bei Antonymen das Gegenteil, zugeordnet. Für viele vorhandene lexikalische Ressourcen wurde als Grundlage dazu das *WordNet* [32] verwendet. Eine Übersicht über aktuelle Ressourcen, die mit diesem

Ansatz erzeugt wurden, findet sich in Tabelle 1. Das beschriebene Verfahren wird dort mit „dba“ (*dictionary-based approach*) abgekürzt.

Eine Sonderform des Wörterbuch-basierten Ansatzes ist das Übersetzungsverfahren. Dazu wird eine vorhandene lexikalische Ressource einer fremden Sprache, mittels eines Übersetzungstools[3], automatisch oder semi-automatisch in die benötigte Sprache übersetzt. Bei diesem Verfahren kommt es allerdings häufig zu Problemen mit Ambiguitäten und der automatischen Auswahl der korrekten Übersetzung. Aus diesem Grund kommt dieses Verfahren meist bei kleineren Lexika überwacht zum Einsatz und ist dann Basis (Seed) für Erweiterungen durch das oben erläuterte bzw. das im Folgenden beschriebene Korpus-basierte Verfahren.

2.3.1.3 *Korpus-basierte Erzeugung*

Beim klassischen Korpus-basierten Verfahren werden ebenfalls Seeds mit meinungstragenden Begriffen verwendet. Allerdings wird für die Extraktion weiterer meinungstragender Wörter bei dieser Methode ein ausreichend großer Textkorpus, wie beispielsweise ein Korpus bestehend aus News-Artikeln, benötigt. Innerhalb dieses Korpus wird nun versucht weitere meinungstragende Wörter zu identifizieren. Dazu werden verschiedene Methoden verwendet, die sich die Syntax der jeweiligen Sprache zu Nutze machen. Hatzivassiloglou and McKeown [23] verwendeten beispielsweise einen Seed zufällig ausgewählter Adjektive mit Polarität, um anschließend im *1987 Wall Street Journal corpus* mittels der Konjunktionen „and“, „or“, „but“, „either-or“, und „neither-nor“ weitere meinungstragende Begriffe zu finden. Remus et al. [45] verwendeten unter anderem signifikante Satzkookkurrenzen, um Begriffe zu identifizieren, die häufig gemeinsam mit Seed-

3 Wie etwa Google Translate (`https://translate.google.com/`) oder das Leo Wörterbuch (`https://dict.leo.org/`).

Wörtern auftraten. Diesen wurde anschließend die gleiche Polarität wie dem Seed-Wort zugeordnet.

Eine Übersicht über aktuelle Ressourcen, die mit diesen Ansatz erzeugt wurden, findet sich ebenfalls in Tabelle 1. Das beschriebene Verfahren wird dort mit „cba“ (*corpus-based approach*) abgekürzt.

Die in der Forschungsgruppe „Analytische Informationssysteme“ entwickelten Methoden zur Generierung einer lexikalischen Ressource inklusive des in dieser Arbeit vorgeschlagenen Algorithmus zählen ebenfalls zu den Korpus-basierten Verfahren. Dennoch stellen die beschriebenen Methoden einen Sonderfall dar, insofern kein initialer Seed, sondern ein Korpus mit bestimmten Eigenschaften für die Erzeugung benötigt wird. Dieser Korpus muss aus mehreren Textabschnitten bestehen, denen jeweils eine numerische Bewertung zugeordnet wurde. Solche Korpora sind vor allem im Bereich der *Kundenbewertungen* zu finden und enthalten oft Rezensionstitel und -texte sowie eine numerische Bewertung in Form von „Sternen“. Durch die Korrelation zwischen Bewertungstext und numerischer Bewertung können so meinungstragende Begriffe extrahiert sowie die dazugehörigen Meinungswerte gebildet werden (siehe Kapitel 4.1 und 4.3.3.3).

2.3.2 *Übersicht vorhandener Ressourcen*

In den letzten Jahren wurden verschiedene lexikalische Ressourcen mit unterschiedlichen Methoden (siehe Abschnitt 2.3.1), für verschiedene Domänen und mit verschiedenen Wertebereichen bzw. Skalen für die Polarität und Stärke der Meinung erstellt. Die meisten dieser Ressourcen haben jedoch gemein, dass sie ausschließlich in englischer Sprache existieren. Wie bereits in Kapitel 2.3.1.2 dargestellt, ist eine vollständig automatische Übersetzung solcher Listen – beispielsweise mittels eines einfachen First-Best-

Verfahrens, bei welchem stets das erste von mehreren Übersetzungsergebnissen eines Wortes verwendet wird – jedoch nur bedingt geeignet, um daraus eine Ressource für andere Sprachen zu generieren.

Tabelle 1 gibt einen Überblick über vorhandene lexikalische Ressourcen für die folgenden Sprachen: Deutsch (de), Englisch (en), Spanisch (es) und Französisch (fr). Die Spalte „Verfahren" gibt die Art der Erzeugung an, wie in Abschnitt 2.3.1 beschrieben. Zusätzlich dazu wird in der Spalte „Phrasen" vermerkt, ob die Ressource ausschließlich aus einzelnen Wörtern (z.B. „gut") besteht oder zudem auch Phrasen (z.B. „sehr gut") listet.

In den folgenden Abschnitten werden verfügbare lexikalische Ressourcen für die deutsche Sprache im Detail betrachtet. Dabei sollen Besonderheiten der Listen sowie Aufbau und Beispiele für Einträge aufgezeigt werden.

2.3.2.1 *SentiWS*

SentiWS – die Kurzform für *SentimentWortschatz* – war die erste deutschsprachige Ressource für das Opinion Mining und wurde 2010 von Remus et al. [45] entwickelt. Das Lexikon wurde in drei Stufen und durch Verwendung von drei verschiedenen Ressourcen generiert. In Stufe 1 wurde zunächst die erste Ressource, das *General Inquirer Lexicon* von Stone et al. [52], semi-automatisch mittels Google Translate[4] übersetzt und anschließend manuell überarbeitet. Danach wurde dieses Basis-Set durch ausgewählte Begriffe aus dem Finanzsektor ergänzt. In Stufe 2 wurden durch eine Kookkurrenzanalyse mit dem Log-Likelihood Maß [15] innerhalb der zweiten Ressource, einem Set von 10.200 annotierten Produktrezensionen[5], weitere Kandidaten, nach manueller Überprüfung, hinzugefügt. In Stufe 3 wurde schließlich die drit-

4 `http://translate.google.com`

5 Diese annotierten Rezensionen wurden von einem nicht weiter genannten Geschäftspartner zur Verfügung gestellt.

Name	Autor(en)	Jahr	Sprache	Verfahren	Umfang	Phrasen
SentiWS	Remus et al. [45]	2010	de	cba	3.468	nein
GermanPolarityClues	Waltinger [60]	2010	de	dba	10.000	vereinzelt (290 Bigramme)
GermanLex	Clematide and Klenner [11]	2010	de	cba	8.000	nein
Sentiment Phrase List	Rill et al. [47]	2012	de	cba	2.833	ja
SentiMerge	Emerson and Declerck [17]	2014	de	merge[a]	99.701	vereinzelt (290 Bigramme)
General Inquirer Lexicon	Stone et al. [52]	1966	en	ma	4.206	nein
-	Hu and Liu [25]	2004	en	dba	6.789	nein
WordNet-Affect	Strapparava et al. [53]	2004	en	dba	4.787	nein
Subjectivity Clues	Wilson et al. [63]	2005	en	dba	8.000	nein
SentiSpin	Takamura et al. [54]	2005	en	cba	88.015	nein
SentiWordNet	Esuli and Sebastiani [18]	2006	en	dba	100.000	nein
Polarity Enhancement	Waltinger [59]	2009	en	dba	137.088	?
Web GP	Velikovich et al. [58]	2010	en	cba	178.104	ja
SentiWordNet 3.0	Baccianella et al. [3]	2010	en	dba	120.000	nein
SenticNet	Cambria et al. [6]	2010	en	dba + cba	5.700	nein
NRC Emotion Lexicon	Mohammad and Turney [35]	2010	en	ma[b]	14.182	nein
SenticNet 2	Cambria et al. [7]	2012	en	dba + cba	14.200	nein
SenticNet 3	Cambria et al. [9]	2014	en	dba + cba	13.700	nein
-	Brooke et al. [5]	2009	es	dba + ma	4.660	nein
FrenchLex	Universität Zürich[c]	2012	fr	dba	7.108	nein

Tabelle 1: Übersicht lexikalischer Ressourcen

a Kombination der vier deutschen Ressourcen *GermanLex*, *SentiWS*, *German SentiSpin* und *GermanPolarityClues*
b Erzeugt durch den Einsatz von *Amazon Mechanical Turk* (`https://www.mturk.com/`)
c `http://bics.sentimental.li/index.php/downloads/`

te Ressource, das Wörterbuch der Kollokationen im Deutschen – German Collocation Dictionary [40] – verwendet, um weitere Kandidaten für das Lexikon zu bestimmen. Dazu wurden die in den Stufen 1 und 2 generierten Begriffe genutzt, um semantische Gruppen innerhalb der Ressource zu identifizieren, die „im Zusammenhang mit Meinungen" – *"related to sentiment"* – stehen. Aus diesen Gruppen wurden dann weitere meinungstragende Wörter extrahiert.

Um Gewichte – ein Pendant zu *Sentiment Values* – für die meinungstragenden Begriffe zu berechnen, kam die Methode *Pointwise Mutual Information (PMI)* [10], angepasst für die Bedürfnisse des Opinion Mining [55, 56], zum Einsatz. Diese Methode basiert auf der Annahme, dass eine semantische Beziehung zwischen Wörtern auch auf eine gleiche Ausrichtung (positiv, negativ) hindeutet. Für die Berechnung wurden zwei Sets mit positiven und negativen deutschen Adjektiven erstellt, sowie ein Textkorpus, bestehend aus 100 Millionen deutschen Sätzen, verwendet. Die berechneten Gewichte wurden schließlich noch auf ein Intervall im Bereich $[-1,0;1,0]$ skaliert. $+1$ steht für absolut positiv, -1 für absolut negativ.

Tabelle 2 zeigt den Aufbau der Ressource *SentiWS* und enthält typische Beispiele für deutschsprachige meinungstragende Adjektive. Die Part-of-speech (POS) Tags werden dabei in Form des *Stuttgart-Tübingen-Tagsets (STTS)* [50] angegeben. Die aufgeführten Flexionen stammen aus einer internen Datenbank der Autoren und wurden, sofern vorhanden, für jede Grundform hinzugefügt.

Die aktuelle Version 1.8c der lexikalischen Ressource besteht aus zwei Listen mit 1.650 positiven (davon 784 Adjektive) bzw. 1.818 negativen (davon 698 Adjektive) Wörtern. Enthalten sind neben Adjektiven auch Adverbien, Nomen und Verben. Durch

Word	POS Tag	Weight	Inflections
perfekt	ADJX	0,7299	perfekterer, perfektest, perfekteren, perfektes, perfekter, perfekterem, perfektester, perfektestes, perfektem, perfekten, perfektesten, perfekteres, perfekteste, perfektestem, perfektere, perfekte
gut	ADJX	0,3716	gutere, guten, gutem, gutst, gutstem, guteres, guterer, gutster, gutstes, gute, guteren, gutsten, guterem, gutste, gutes, guter
befriedigend	ADJX	0,2152	befriedigendst, befriedigendes, befriedigendstem, befriedigendsten, befriedigender, befriedigendstes, befriedigenden, befriedigendster, befriedigendem, befriedigendste, befriedigende, befriedigendere, befriedigenderem, befriedigenderen, befriedigenderer, befriedigenderes
schlecht	ADJX	-0,7706	schlechtem, schlechten, schlechteste, schlechtes, schlechtest, schlechte, schlechter, schlechteren, schlechterem, schlechtesten, schlechtestem, schlechtere, schlechtester, schlechteres, schlechterer, schlechtestes
miserabel	ADJX	-0,2004	miserablen, miserabler, miserables, miserablerem, miserableren, miserablere, miserable, miserabelst, miserabelster, miserableres, miserabelste, miserabelstes, miserablerer, miserabelsten, miserabelstem, miserablem

Tabelle 2: Aufbau SentiWS mit realen Beispielen

die Hinzunahme der Flexionen erhöht sich der Umfang der Ressource auf 15.649 positive sowie 15.632 negative Wörter.

2.3.2.2 *GermanPolarityClues*

GermanPolarityClues wurde 2010 von Waltinger [60] entwickelt. Dazu wurden zunächst die wichtigsten englischsprachigen Ressourcen, *Subjectivity Clues* [63], *SentiSpin* [54], *SentiWordNet* [18] und *Polarity Enhancement* [59], mittels eines Experiments zur Identifikation der Polarität auf Dokumentenebene evaluiert [61]. Anschließend wurden die beiden ausgewählten Ressourcen *Subjec-*

tivity Clues und *SentiSpin* durch einen automatischen Ansatz ins Deutsche übersetzt.[6] Dabei wurden bis zu maximal drei mögliche Übersetzungen übernommen – das Gewicht der Polarität wurde vererbt –, was den Umfang der erzeugten Lexika stark erhöhte. Bei der Überprüfung der beiden automatisch generierten Lexika *German Subjectivity Clues* und *German SentiSpin* stellten die Autoren allerdings Probleme mit der Mehrdeutigkeit von Begriffen fest.

Aufgrund der, durch den automatischen Übersetzungsansatz entstandenen, Ambiguitäten, wurde die Ressource *GermanPolarityClues* schließlich durch manuelle Überprüfung aller Einträge der vorher erzeugten *German Subjectivity Clues* generiert. Zudem wurden der Ressource 290 Negationsphrasen (z.B. „nicht schlecht“) sowie die häufigsten Synonyme aller vorhandenen Einträge (durch Verwendung von Wiktionary[7]) hinzugefügt. In einer späteren Version der Ressource wurden zudem die Einträge von *SentiWS* [45] mit einbezogen.

Die beschriebene Version der lexikalischen Ressource besteht aus 10.141 meinungstragenden Wörtern. Davon sind 25,7% Adjektive, 43,5% Nomen und 26,9% Verben, alle in der jeweiligen Grundform.

Tabelle 3 zeigt den Aufbau der lexikalischen Ressource und enthält wiederum reale Beispiele meinungstragender Adjektive. Da diese Daten aus der aktuellen Ressource stammen, ist nicht verwunderlich, dass die Werte der fünf Adjektive exakt denen aus *SentiWS* in Tabelle 2 gleichen.

2.3.2.3 *GermanLex*

Das Polarity Lexicon for German, kurz *GermanLex*, wurde 2010 von Clematide and Klenner [11] entwickelt. Dazu wurde zunächst ein Seed, bestehend aus 2.899 meinungstragenden Adjektiven

6 Dazu wurde der Online-Service `https://dict.leo.org/` verwendet.
7 `https://de.wiktionary.org/`

Feature	Lemmata	POS	Rating	Pos	Neg	Neu
perfekt	perfekt	AD	positive	0,7299	-	-
gut	gut	AD	positive	0,3716	-	-
befriedigend	befriedigend	AD	positive	0,2152	-	-
schlecht	schlecht	AD	negative	-	-0,7706	-
miserabel	miserabel	AD	negative	-	-0,2004	-

Tabelle 3: Aufbau GermanPolarityClues mit realen Beispielen

inklusive deren Polarität, verwendet, um mit diesem über den Service des *Wortschatz Leipzig*[8] weitere Flexionen zu generieren. 23.761 Wortformen konnten auf diese Weise ermittelt werden. Zusätzlich dazu wurden zu jeder Wortform mehrere Beispielsätze geladen, insgesamt 2.039.175. Mittels Frequenzanalyse untersuchten die Autoren Paare direkt aufeinander folgender Adjektive und extrahierten jene Paare, in denen eine Wortform aus dem erweiterten Seed vorkam. Die häufigsten Partner solcher Adjektive wurden schließlich in die lexikalische Ressource übernommen und erbten deren Polarität.

Tabelle 4 illustriert den Aufbau der Ressource und enthält fünf Beispiele meinungstragender Adjektive.

Die aktuelle Version der lexikalischen Ressource besteht aus ca. 8.000 meinungstragenden Wörtern. Enthalten sind Adjektive, Nomen und Verben, alle in der jeweiligen Grundform. Zudem enthält die Liste verschiedene Meinungsshifter. Zu beachten ist, dass diese Ressource über lediglich 3 Gewichtsangaben pro Polarität verfügt, nämlich 1, 0,7 und 0.

2.3.2.4 *Sentiment Phrase List*

Die Sentiment Phrase List [47], kurz *SePL*, wurde 2012 in der Forschungsgruppe „Analytische Informationssysteme“ entwickelt. Bei der Entwicklung kam ein vollkommen neuer Ansatz für die Generierung lexikalischer Ressourcen zum Einsatz. Grundlage

8 http://wortschatz.uni-leipzig.de

Word	Polarity	Strength	POS
perfekt	POS	1	-
gut	POS	1	-
befriedigend	POS	1	-
schlecht	NEG	1	-
miserabel	NEG	1	-

Tabelle 4: Aufbau GermanLex mit realen Beispielen

für die Erzeugung dieser Ressource waren deutschsprachige Produktrezensionen von Amazon[9]. Dazu wurden zum einen Pattern definiert, um meinungstragende Wörter und Phrasen aus Titeln zu extrahieren. Der Fokus lag dabei auf (meinungstragenden) Adjektiven und Nomen, beispielsweise „sehr gut“ oder „absoluter Mist“. Zum anderen wurde die Korrelation zwischen Titel und Sternebewertungen (1 bis 5) der Rezensionen genutzt, um aus der großen Anzahl von Sternebewertungen pro Eintrag automatisch einen Meinungswert (Opinion Value) zu berechnen.[10] Abschließend wurden jene Meinungswerte korrigiert, die durch die sogenannte „J-shaped“-Verteilung [27, 26] der Rezensionen[11] falsch gebildet wurden. Der Opinion Value (OV) der Wörter und Phrasen dieser Ressource liegt auf einer kontinuierlichen Skala im Bereich $[-1,0;1,0]$, wobei -1 für sehr negative und 1 für sehr positive Einträge steht.

Tabelle 5 zeigt den Aufbau der lexikalischen Ressource, inklusive der realen Beispiele meinungstragender Adjektive. Zusätzlich zur meinungstragenden Phrase und dem Opinion Value (OV), listet die Ressource die dazugehörige Standardabweichung (SD) sowie Standardfehler (SE), Phrasentyp – wobei „a“ für Adjektive steht – und, in einer neueren Version, eine Kennzeichnung die

9 `https://www.amazon.de/`

10 Der genaue Ablauf sowie die Formel zur Berechnung von *Sentiment Values* bzw. *Opinion Values* ist Teil dieser Arbeit und kann in Kapitel 4.3.3.3 nachgeschlagen werden.

11 Die Verteilungen der Rezensionen auf der originären Skala $[+1, +5]$ zeigen die Form einer Parabel mit einem Minumum bei 2. Diese Verteilung erinnert an ein „J“.

Phrase	*OV*	*SD*	*SE*	**Phrase Type**	**Manuel Correction**
perfekt	0,948	0,191	0,002	a	-
gut	0,632	0,427	0,003	a	-
befriedigend	-0,035	0,267	0,037	a	-
schlecht	-0,677	0,595	0,011	a	-
miserabel	-0,821	0,399	0,029	a	-

Tabelle 5: Aufbau Sentiment Phrase List mit realen Beispielen

angibt, ob der Opinion Value bei einem nachgelagerten Korrekturprozess manuell korrigiert wurde.

Die beschriebene Liste enthält 3.210 Einträge. Davon 1.277 Adjektive, 938 Adjektivphrasen, 502 Nomen und 493 Nomenphrasen. In der aktuellen Version 1.1 (2014) wurde durch diverse Erweiterungen – u.a. wurden Adjektivphrasen nun auch im Text der Rezension gesucht und Verben in Titeln zugelassen – der Umfang der Ressource auf ca. 14.400 Einträge erhöht, davon 77,5% Adjektive und Adjektivphrasen, 20,7% Nomen und Nomenphrasen und 1,8% Verben und Verbphrasen.

2.3.2.5 *SentiMerge*

SentiMerge ist die jüngste verfügbare Ressource für die deutsche Sprache und wurde 2014 von Emerson and Declerck [17] entwickelt. Im Prinzip handelt es sich nicht um ein eigenständiges Lexikon, sondern vielmehr um die Kombination der vier deutschen Ressourcen *GermanLex*, *SentiWS*, *German SentiSpin*[12] und *GermanPolarityClues*. Die Autoren stellten dazu ein Bayessches Wahrscheinlichkeitsmodell vor, um die vier Listen zu vereinen. Dafür formulierten sie die Annahme, dass für jeden Begriff ein (verborgener) echter Meinungswert existiert und jede der vier lexikalischen Ressourcen eine Beobachtung dieses Wertes enthält

12 *German SentiSpin* war ein Nebenprodukt der Arbeiten von Waltinger [60] bei dem die englischsprachige Ressource *SentiSpin* automatisch übersetzt wurde. Aufgrund der Größe und vieler Ambiguitäten wurde diese Liste später jedoch nicht dazu genutzt, um *GermanPolarityClues* zu generieren.

Phrase	POS	Sentiment[13]	Weight
perfekt	AJ	1,07	11,98
gut	ADJA	0,28	6,69
gut	AJ	0,91	11,98
befriedigend	ADJA	0,83	3,53
befriedigend	AJ	0,61	11,98
schlecht	ADJA	-0,49	6,69
schlecht	AJ	-1,35	11,98
miserabel	AJ	-0,77	11,98

Tabelle 6: Aufbau SentiMerge mit realen Beispielen

sowie etwas „Hintergrundrauschen“. Nach der Normalisierung der Meinungswerte aller Lexika, wurden diese durch das Bayessche Verfahren kombiniert. Aufgrund der Normalisierung liegen die Meinungswerte der Ressource *SentiMerge* in einem Intervall $[-1,628961; 1,521099]$. Begriffe im Bereich $[-0,23; 0,23]$ wurden dabei von den Autoren als „neutral“ definiert. Diese Definition trifft auf 81,1% aller Einträge zu.

Tabelle 6 zeigt den Aufbau der lexikalischen Ressource, inklusive der Beispiele für meinungstragende Adjektive. Auffällig sind hierbei doppelte Einträge von Adjektiven mit verschiedenen POS Tags („ADJA“ und „AJ“). Aus der Veröffentlichung geht die Bedeutung dieser Trennung nicht hervor. Zudem wird zu jedem Eintrag ein Gewicht („Weight“) angegeben, welches anzeigt, in wie vielen Listen der entsprechende Eintrag vorhanden war. Ein höheres Gewicht steht dabei für einen zuverlässigeren Meinungswert.

Die aktuelle Ressource enthält 99.701, größtenteils lemmatisierte, Einträge. Davon 14,4% Adjektive, 70,0% Nomen und 11,5% Verben. Zudem enthält die Liste 290 Bigramm-Phrasen auf Basis der Verneinungen „nicht“ und „kein“ (z.B. „nicht gut“), die jedoch teils unplausible Meinungswerte besitzen.

13 Aus Gründen der Übersichtlichkeit wurden die, bis zu 16 Nachkommastellen langen, Werte für *Sentiment* und *Weight* auf zwei Nachkommastellen gerundet.

3

VERWENDETE MASSE UND METHODEN

Dieses Kapitel gibt einen Überblick über die in dieser Arbeit verwendeten statistischen Maße und Methoden. Dabei sollen vor allem deren Funktionsweise und Einsatzzweck innerhalb der Dissertation aufgezeigt werden.

3.1 Interrater-Reliabilität

Sowohl für die Evaluation automatisch erstellter lexikalischer Ressourcen, als auch für die Optimierung der dazu benötigten Algorithmen, werden manuell annotierte Datensets als Referenz benötigt. Diese, durch verschiedene Personen (sogenannte Rater) annotierten und erstellten, Sets müssen vor dem Einsatz als Referenzset auf Plausibilität überprüft werden. Für diese Überprüfung existieren unter dem Begriff *Interrater-Reliabilität* verschiedene Methoden, mit denen der Grad der Übereinstimmung zwischen zwei oder mehreren Personen, unter Berücksichtigung der zufälligen Übereinstimmung, bestimmt werden kann. In dieser Arbeit werden dazu die bekannten Kappa-Statistiken verwendet.

Zur Messung der Übereinstimmung zwischen zwei Ratern wird dazu das statistische Maß *Cohens Kappa* κ [12] verwendet. Bei der Überprüfung von Übereinstimmung bzw. Nichtübereinstimmung (Nominalskala) der Bewertungen von zwei Ratern, kann nach Aufstellung einer $z \times z$ Kontingenztabelle der Beurteilungsergebnisse z der beiden Rater, in welcher die Urteilshäufigkeiten

h aufgetragen werden, κ wie folgt berechnet werden, wobei N die Anzahl aller Beurteilungen darstellt:

$$\kappa = \frac{p_0 - p_c}{1 - p_c} \tag{1}$$

mit:

$$p_0 = \frac{\sum_{i=1}^{z} h_{ii}}{N} \tag{2}$$

und:

$$p_c = \frac{\sum_{i=1}^{z} h_{i\cdot} \cdot h_{\cdot i}}{N^2} \tag{3}$$

p_0 wird dabei als gemessene Übereinstimmung (Hauptdiagonale) und p_c als erwartete zufällige Übereinstimmung (Produkte der Randsummen) bezeichnet. In Tabelle 7 wird dies anhand eines kurzen Beispiels verdeutlicht, bei dem die Übereinstimmungshäufigkeiten h der beiden Rater 1 und 2 für die Kategorien „positiv“ (pos), „neutral“ (neu) und „negativ“ (neg) in einer 3×3 Kontingenztafel gelistet werden.

	Rater 2			
Rater 1	pos	neu	neg	Summe
pos	h_{11}	h_{12}	h_{13}	$h_{1\cdot}$
neu	h_{21}	h_{22}	h_{23}	$h_{2\cdot}$
neg	h_{31}	h_{32}	h_{33}	$h_{3\cdot}$
Summe	$h_{\cdot 1}$	$h_{\cdot 2}$	$h_{\cdot 3}$	N

Tabelle 7: Kontingenztabelle als Basis für die Bestimmung von Cohens Kappa

Sollen die Rater bei der Beurteilung eine Einstufung in Form einer mehrstufigen, ordinalen Skala vornehmen – beispielsweise

zwischen 1 und 5 Sternen –, kann die Berechnung entsprechend angepasst und ggf. Gewichte hinzugefügt werden [13, 20]. Somit kommt auch der Abstand der Beurteilungen stärker zum Tragen, d.h. bei unterschiedlichen Meinungen bezüglich einer Einstufung wird der Fall „Rater 1: 1 Stern, Rater 2: 5 Sterne“ stärker *bestraft*, als der Fall „Rater 1: 1 Stern, Rater 2: 2 Sterne“.

Bei der Überprüfung der Übereinstimmung von mehr als zwei Ratern werden die oben beschriebenen Formeln nach Fleiss et al. [19] erweitert. Diese Erweiterung wird dabei als *Fleiss' Kappa* bezeichnet.

Für die anschließende Beurteilung des errechneten κ-Wertes, schlagen Landis and Koch [28] folgende Einteilung vor:

- $\kappa < 0$: Less than chance agreement (schlechte Übereinstimmung)
- $0,01 \leq \kappa \leq 0,20$: Slight agreement (etwas Übereinstimmung)
- $0,21 \leq \kappa \leq 0,40$: Fair agreement (faire Übereinstimmung)
- $0,41 \leq \kappa \leq 0,60$: Moderate agreement (moderate Übereinstimmung)
- $0,61 \leq \kappa \leq 0,80$: Substantial agreement (beachtliche Übereinstimmung)
- $0,81 \leq \kappa \leq 0,99$: Almost perfect agreement (fast perfekte Übereinstimmung)

3.2 Precision und Recall

Für die Beurteilung einer automatischen Klassifikation, werden die Güteparameter *Precision* (Genauigkeit), *Recall* (Vollständigkeit) und der F_1-*Wert*, als das harmonische Mittel der beiden Parameter, verwendet.

Anhand einer beispielhaften Kontingenztabelle (siehe Tabelle 8), soll die Berechnung der Parameter verdeutlicht werden. Die Ergebnisse einer Klassifikation können dabei in vier mögliche Felder eingeordnet werden:

- *true positive (tp)*: Klasse „positiv“ richtig als „positiv“ klassifiziert
- *false positive (fp)*: Klasse „negativ“ falsch als „positiv“ klassifiziert
- *false negative (fn)*: Klasse „positiv“ falsch als „negativ“ klassifiziert
- *true negative (tn)*: Klasse „negativ“ richtig als „negativ“ klassifiziert

Klasse / **Klassifikation**	positiv	negativ
positiv	tp	fp
negativ	fn	tn

Tabelle 8: Kontingenztabelle für die Berechnung von Precision und Recall

Mit dieser Einteilung können die Güteparameter *Precision* und *Recall* berechnet werden. Der Wert für *Precision*, der das Verhältnis der richtig als „positiv“ klassifizierten Elemente zu allen „positiv“ klassifizierten Elementen beschreibt, wird folgendermaßen berechnet:

$$Precision = \frac{tp}{tp + fp} \tag{4}$$

Der Wert für den *Recall*, der das Verhältnis zwischen richtig als „positiv“ klassifizierten Elementen und allen tatsächlich „positiven“ Elementen widerspiegelt, wird wie folgt berechnet:

$$Recall = \frac{tp}{tp + fn} \tag{5}$$

Der F_1-*Wert*, als harmonisches Mittel von *Precision* und *Recall*, wird folgendermaßen bestimmt:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{6}$$

3.3 Korrelationskoeffizienten

Zur Messung der Korrelation, d.h. des linearen Zusammenhangs zwischen zwei Variablen, können Korrelationskoeffizienten berechnet werden. In dieser Arbeit werden der Korrelationskoeffizient nach Pearson (r) sowie der Spearman-Korrelationskoeffizient (ρ), aus der Gruppe der Rangkorrelationskoeffizienten, verwendet.

Für die Berechnung von r werden alle Beobachtungspaare x_i und y_i der einzelnen i Messungen verwendet, siehe Formel 7. Dabei bezeichnen $\bar{x}$ und $\bar{y}$ jeweils das arithmetische Mittel der x-Werte bzw. der y-Werte.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{7}$$

Da der Korrelationskoeffizient nach Pearson sehr empfindlich gegenüber Ausreißern ist, wurde u.a. der Spearman-Korrelationskoeffizient (ρ) entwickelt. Die Berechnung von ρ ist analog zu

Pearson, allerdings werden statt der Originalwerte von x_i und y_i deren Ränge $rang(x_i)$ und $rang(y_i)$ verwendet, siehe Formel 8:

$$\rho = \frac{\sum_{i=1}^{n}(rang(x_i) - \overline{rang(x)})(rang(y_i) - \overline{rang(y)})}{\sqrt{\sum_{i=1}^{n}(rang(x_i) - \overline{rang(x)})^2} \cdot \sqrt{\sum_{i=1}^{n}(rang(y_i) - \overline{rang(y)})^2}} \tag{8}$$

Sowohl Pearson's r als auch Spearman's ρ können Werte zwischen -1 und $+1$ annehmen. Voraussetzung für die Bestimmung von r ist dabei die (annähernde) Normalverteilung der beiden Variablen.
Interpretiert werden können die Werte der Koeffizienten folgendermaßen:

- $r = -1$ bzw. $\rho = -1$: perfekte (negative) Korrelation
- $r = 0$ bzw. $\rho = 0$: keine Korrelation vorhanden
- $r = +1$ bzw. $\rho = +1$: perfekte (positive) Korrelation

3.4 Weitere Maße

Für die Erzeugung einer neuen lexikalischen Ressource werden weitere statistische Maße verwendet, die in Kapitel 4.3.3 näher beschrieben werden. Bei diesen Maßen handelt es sich um die klassische *Frequenz* (Kapitel 4.3.3.1), ein angepasstes Maß für *signifikante Kookkurrenzen* (Kapitel 4.3.3.2) sowie eine Formel für die Berechnung von *Meinungswerten* für meinungstragende Begriffe (Kapitel 4.3.3.3).

4

GENERIERUNG LEXIKALISCHER RESSOURCEN

Der Inhalt dieses Kapitels basiert zu Teilen auf der eingereichten und angenommenen Veröffentlichung *Sentiment Phrase Generation Using Statistical Methods* [44], die während der Forschungsarbeiten zum Thema „Generierung lexikalischer Ressourcen für das Opinion Mining" entstand und das Kernthema dieser Dissertation bildet.

Im Folgenden wird ein neu entwickelter Algorithmus vorgestellt, der es ermöglicht, aus einem Textkorpus mit bestimmten Eigenschaften (siehe Abschnitt 4.3.2) – hier Produktrezensionen der Plattform *Amazon* – automatisch ein Lexikon mit meinungstragenden Wörtern und Phrasen zu erzeugen. Dabei verzichtet das beschriebene Verfahren auf gängige NLP-Methoden, wie POS-Tagging oder Lemmatizing, und stützt sich ausschließlich auf statistische Methoden, wie Worthäufigkeiten oder signifikante Kookkurrenzen. Aus diesem Grund kann das Verfahren vergleichsweise einfach für die Erzeugung lexikalischer Ressourcen für verschiedene Sprachen und Domänen angepasst werden und die damit erzeugten Lexika können für verschiedene Opinion Mining Anwendungen verwendet werden. In dieser Arbeit wird eine solche Ressource mittels des beschriebenen Verfahrens exemplarisch für die deutsche Sprache, unter Verwendung eines Korpus, bestehend aus deutschsprachigen Kundenrezensionen von Amazon, generiert.

Wie bereits erwähnt, basieren Teile der neuen Methode auf den Ideen der Forschungsgruppe „Analytische Informationssysteme" zur Erzeugung lexikalischer Ressourcen, die bereits veröffentlicht wurden [48].

Dieses Kapitel ist wie folgt gegliedert. In Abschnitt 4.1 werden die grundlegenden Ideen und Annahmen des neuen Verfahrens zur Generierung von Listen meinungstragender Wörter und Phrasen beschrieben. Des Weiteren werden in Abschnitt 4.2 das dafür entwickelte System sowie die Datenstruktur der eingesetzten Graphdatenbank vorgestellt. In Abschnitt 4.3 folgt schließlich die ausführliche Beschreibung der Implementierung einschließlich einer Übersicht der verwendeten Quellen und Kenngrößen. Zudem werden in diesem Abschnitt die Datenvorverarbeitung (4.3.4), der Algorithmus zum Aufbau des Lexikons (4.3.5) sowie die anschließende Bereinigung der Daten (4.3.6) detailliert beschrieben. In Abschnitt 4.4 werden schließlich Experimente zum Laufzeitverhalten und in Abschnitt 4.5 Experimente zur Schwellenwertbestimmung der definierten Parameter vorgestellt. Abschließend werden die Ergebnisse der Listengenerierung in Abschnitt 4.6 präsentiert.

4.1 Grundlegende Ideen und Annahmen

Die grundlegende Idee des neuen Ansatzes ist es, die vorhandene Korrelation zwischen dem Titel – meist Wortgruppen oder kurze Sätze – und der numerischen Bewertung einer Produktrezension zu nutzen [48]. Diese Korrelation wurde in der Dissertation von Rill [46] untersucht und letztendlich nachgewiesen. Ein Beispiel für die vorhandene Korrelation zeigt der Titel der Amazon Produktrezension „Super Gerät, eierlegende Wollmilchsau“, siehe Abbildung 4. Der Rezensent ordnet einer so betitelten Rezension normalerweise eine hohe, d.h. eine sehr gute, Sternebewertung zu. Das grundlegende Prinzip, das sich dabei zeigt, ist, dass sowohl Titel als auch Sternebewertung vereinfachte Zusammenfassungen des gesamten Rezensionstextes sind.

Abbildung 4: Beispiel einer Amazon Kundenrezension – Sternebewertung und Titel (Quelle: [33])

Frühere Experimente zeigten außerdem, dass die Verwendung des gesamten Textes der Rezensionen – zusätzlich zum Titel – zwar die Anzahl extrahierter meinungstragender Wörter und Phrasen erhöht, dadurch aber im selben Zuge eine große Anzahl ungewollter Begriffe extrahiert wird. Zudem verlängert sich durch die Hinzunahme des Textes die Berechnungsdauer des Verfahrens.

Bisherige Verfahren zur Generierung lexikalischer Ressourcen verwenden NLP-Methoden, die auf die jeweiligen Sprachen (meist Englisch) angepasst wurden und dementsprechend gut funktionieren. Eine Anpassung für andere Sprachen ist mit diesen Verfahren jedoch nicht, oder nur durch großen Aufwand, möglich. Diese Einschränkung soll durch das in dieser Arbeit vorgeschlagene Verfahren aufgehoben werden, indem vollständig auf NLP-Methoden verzichtet und stattdessen auf statistische Methoden zurückgegriffen wird. Außerdem soll für Aufbau und Datenhaltung – aufgrund der Struktur der Daten – eine Graphdatenbank eingesetzt werden. Ein Beispiel für einen solchen Graph zeigt Abbildung 5.

4.2 Systemarchitektur

Das gesamte System wurde so konzipiert und aufgebaut, dass die Listengenerierung auf einem handelsüblichen Standard-PC betrieben werden kann. Einzige Voraussetzung ist, neben dem ob-

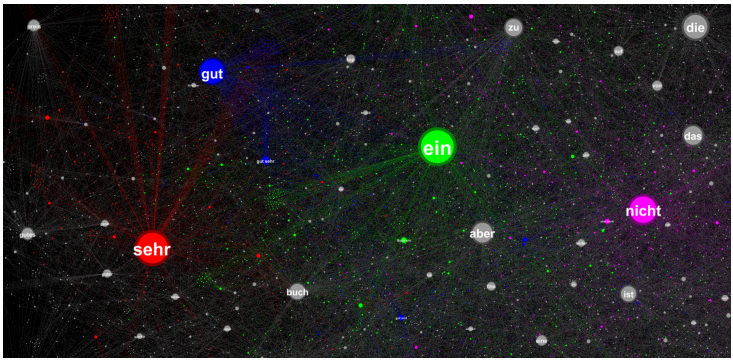


Abbildung 5: Beispiel einer Graphenstruktur für meinungstragende Phrasen

ligatorischen Textkorpus, das aktuelle *Java Development Kit (JDK)*[1], welches das ebenfalls benötigte *Java Runtime Environment (JRE)* enthält. Je nach verfügbarer Hardware – dabei sind vor allem die Geschwindigkeit des Prozessors, die Größe des Arbeitsspeichers und die Art des Speichermediums von Bedeutung – erhöht bzw. verkürzt sich die Laufzeit des Generierungsprozesses bei gleichem Ausgangskorpus, siehe dazu Kapitel 4.4.

Im Folgenden werden das Datenmodell der verwendeten Graphdatenbank sowie die gewählten Input- sowie Output-Formate näher beschrieben.

4.2.1 *Datenmodell der Graphdatenbank*

Für die Berechnung der lexikalischen Ressource wurde, wie bereits erwähnt, als Hilfsmittel für die Umsetzung eine Graphdatenbank (Neo4J[2]) verwendet. Prinzipiell wären zwar auch klassische Datenstrukturen möglich gewesen, für die Art der Daten und deren Verbindungen war der Einsatz von Graphenstruktu-

1 http://www.oracle.com/technetwork/java/javase/downloads/index.html
2 https://neo4j.com/

ren allerdings naheliegend. Im Folgenden wird der Aufbau des Graphen näher beschrieben.
Die grundlegende Struktur des Graphen besteht aus drei verschiedenen Knotentypen:

- *Sentence nodes* repräsentieren die Sätze des jeweiligen Textkorpus. Diese Knoten enthalten eine eindeutige ID, den Text des Titels bzw. Satzes sowie die dazugehörende Bewertung.

- *Phrase nodes for unigrams* – auch *unigram nodes* genannt – enthalten das meinungstragende Unigramm, die Frequenz (Berechnung siehe Abschnitt 4.3.3.1), den Sentiment Value SV_i sowie die dazugehörige Standardabweichung σ_{SV_i} (Berechnung siehe Abschnitt 4.3.3.3).

- *Phrase nodes for n-grams* mit $n \geq 2$ – auch *n-gram nodes* genannt – enthalten die meinungstragende Phrase, den Signifikanzwert $sig(A_1, \ldots, A_n)$ (Berechnung siehe Abschnitt 4.3.3.2), die Frequenz, den Sentiment Value SV_i und die dazugehörige Standardabweichung σ_{SV_i}.

Diese Knoten werden durch folgende Kanten verbunden:

- *Occurrence edges* verbinden *Phrase nodes* mit *Sentence nodes*, in denen diese Phrasen vorkommen. In den Kanten werden die entsprechenden Frequenzen des Vorkommens gespeichert.

- *Sub-phrase edges* verbinden *n-gram nodes* mit allen *Phrase nodes*, die Subsets dieser Knoten mit $n - 1$ Wörtern darstellen. Ein Beispiel dafür gibt das Bigramm „gut sehr“ (siehe Abbildung 6). Der entsprechende *n-gram node* ist durch die gestrichelte Kante (*Sub-phrase edge*) mit den beiden *Unigram nodes* „gut“ und „sehr“ verbunden.

Zum besseren Verständnis wird in Abbildung 6 ein Beispielgraph dargestellt, der aus drei *Sentence nodes* und den dazugehörigen, auf meinungstragende Phrasen reduzierte, *Phrase nodes* besteht.

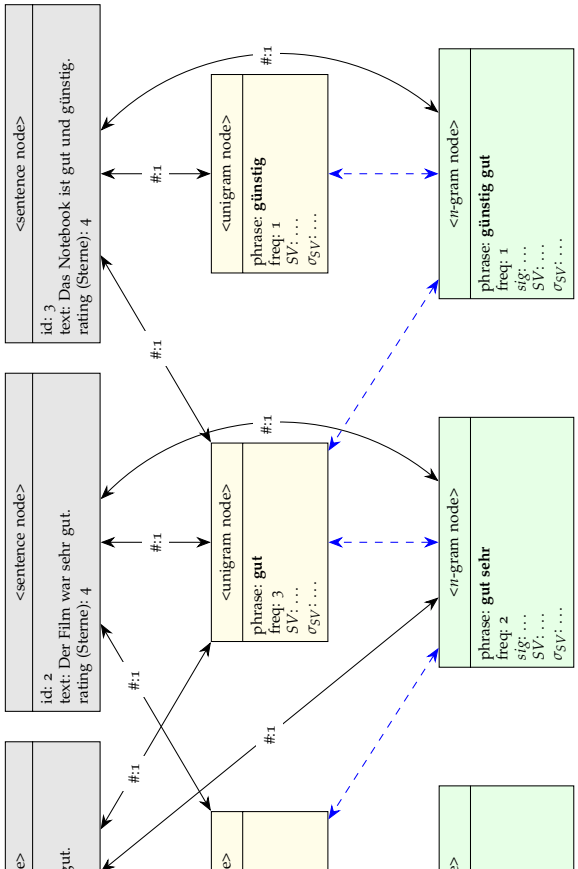


Abbildung 6: Beispielgraph nach dem Einfügen von Bigrammen

4.2.2 *Input- und Output-Format*

Als Eingabe für den Algorithmus kommen alle Korpora mit Texten (in Sätzen) und einer entsprechenden Bewertung in Frage. Das Eingabeformat muss dabei in Listenform vorliegen, siehe Listing 1.

```
1 ID      Bewertung       Text
2 1       5               "Dieses Buch ist einfach klasse."
3 2       3               "Die Ausstattung des PCs ist ok."
4 3       1               "Der Film war wirklich furchtbar."
```

Listing 1: Input-Format

Nach Berechnung der meinungstragenden Wörter und Phrasen durch das in Kapitel 4.3 beschriebene statistische Verfahren, wird die finale lexikalische Ressource durch Übertragung aller Phrasenknoten des Graphen in eine Lexikon-Datei erzeugt. Diese Datei ist wie folgt aufgebaut, siehe Listing 2. *n* steht dabei für die Anzahl der Wörter pro Phrase, *Phrase* für den Text, *Freq* für die Frequenz der Phrase, *Sig* für den Signifikanzwert, *SV* für den Sentiment Value und *SD* für die Standardabweichung.

```
1 n       Phrase          Freq    Sig     SV      SD
2 1       gut             4955    0,00    0,69    0,36
3 2       gut sehr        6933    786,01  0,98    0,09
4 2       finger weg      1356    660,98 -1,00    0,00
```

Listing 2: Output-Format

4.3 Implementierung

In diesem Abschnitt werden alle Schritte zur Erzeugung lexikalischer Ressourcen, mittels des vorgeschlagenen Verfahrens, detailliert beschrieben.

Nach einer generellen Systemübersicht in Abschnitt 4.3.1, wird in Abschnitt 4.3.2 der Korpus, der bei den späteren Experimenten zum Einsatz kam, näher beschrieben. Abschnitt 4.3.3 gibt einen Überblick über die statistischen Maße, die eingesetzt wurden, sowie deren Berechnung. In Abschnitt 4.3.4 werden notwendige Vorverarbeitungs- und Filterschritte beschrieben. Abschnitt 4.3.5 beinhaltet die Beschreibung des Aufbaus der Graphdatenbank, die dazu genutzt wird, um einen Graph mit meinungstragenden Wörtern und Phrasen zu erzeugen, aus dem später die lexikalische Ressource generiert wird. In Abschnitt 4.3.6 werden schließlich die notwendigen Nachbearbeitungsschritte beschrieben.

4.3.1 *Systemübersicht*

Abbildung 7 zeigt eine schematische Darstellung des Prozesses zur Generierung von Listen meinungstragender Wörter und Phrasen. Das System wurde in die fünf Bereiche *Data Retrieval* (Abschnitt 4.3.2), *Value Calculation* (Abschnitt 4.3.3), *Preprocessing* (Abschnitt 4.3.4), *Graph Construction* (Abschnitt 4.3.5) und *Postprocessing* (Abschnitt 4.3.6) unterteilt, auf die in den folgenden Abschnitten näher eingegangen wird.

4.3.2 *Verwendete Datenquellen*

Die Basis des automatischen Verfahrens sowie der dazugehörigen Experimente bildet ein Textkorpus, der bestimmte Eigenschaften besitzen muss. Dies ist zum einen das Vorkommen von (kurzen) Textabschnitten, wie etwa Wortgruppen oder Sätzen, und zum anderen die jeweilige Zuordnung einer numerischen Bewertung zu diesen lexikalischen Einheiten. Dies trifft insbesondere auf Kundenrezensionen in Bewertungsportalen zu. Dabei spielt

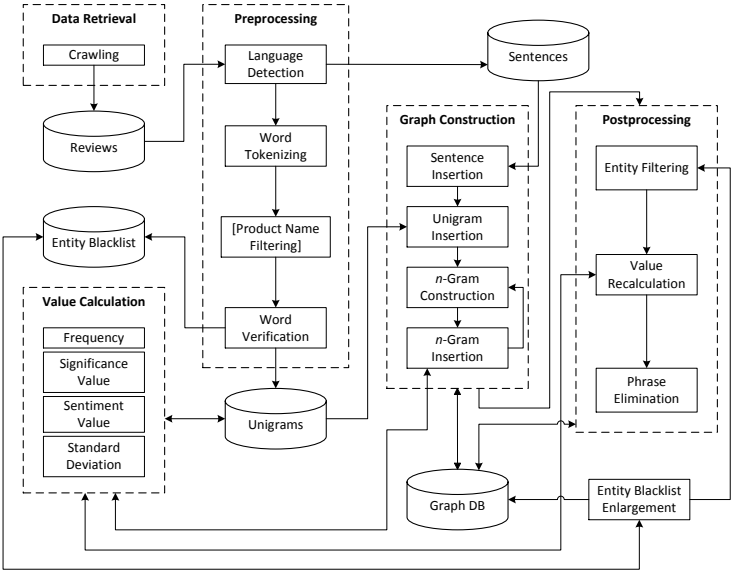


Abbildung 7: Ablauf des Listengenerierungsprozesses

die Art des Bewertungsgegenstands – Produkte, Filme, Hotels, Unternehmen, Services etc. – keine Rolle.

Für die Versuche in dieser Arbeit wurde exemplarisch ein Korpus, bestehend aus deutschen Kundenrezensionen der Website *Amazon.de*, ausgewählt, der durch einen selbst entwickelten Scraper aufgebaut wurde. Neben der vergleichsweise einfachen Datengewinnung war insbesondere die große Anzahl an Kundenrezensionen ein Grund für die Verwendung dieser Ressource. Der Korpus enthält ca. 1.500.000 deutsche Rezensionen von etwa 62.000 verschiedenen Produkten aus 40 unterschiedlichen Kategorien. Für die automatische Generierung des Lexikons meinungstragender Wörter und Phrasen wurden ausschließlich der Titel und die Sternebewertung der jeweiligen Rezensionen verwendet. Die Idee des Ansatzes wurde bereits in Abschnitt 4.1 formuliert, ebenso die Annahme, dass jeder Rezensionstitel als ei-

genständiger Satz betrachtet wurde. Der Datensatz des Amazon-Korpus besteht aus ca. 1,5 Millionen Sätzen mit einer durchschnittlichen Satzlänge von 3,91 Wörtern. Abbildung 8 zeigt die Verteilung der Satzlängen, d.h. die Anzahl der Wörter pro Satz, im Korpus.

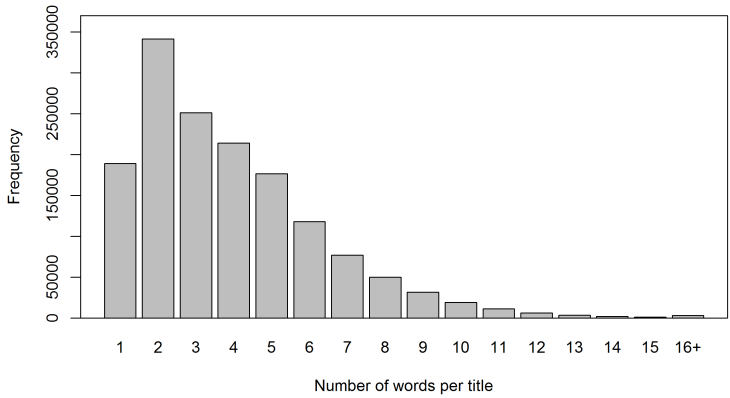


Abbildung 8: Verteilung der Satzlängen im Amazon-Korpus

4.3.3 *Verwendete Maße und deren Berechnung*

Für die Erzeugung einer lexikalischen Ressource mit statistischen Methoden werden verschiedene Maße benötigt, die angeben, ob bzw. wie relevant Wörter oder Phrasen als Kandidaten für das endgültige Lexikon sind und welche Meinungswerte diese erhalten. Die verwendeten Maße sowie deren Berechnung und Verwendung werden im Folgenden detailliert beschrieben.

4.3.3.1 *Frequenz*

Das Maß *Frequenz* wird verwendet, um die Anzahl der Sätze anzugeben, die das jeweilige n-Gramm mindestens einmal enthalten. Der dazugehörige Schwellenwert der Frequenz $\hat{f}_{n-gramm}$ wird im Algorithmus dazu genutzt, um zu entscheiden, ob ein n-Gramm als Kandidat für das Lexikon in Frage kommt. Experimente zur Bestimmung des optimalen Schwellenwerts für diesen Parameter werden in Abschnitt 4.5 beschrieben. Zudem wird im vorgestellten Algorithmus die *within sentence frequency* verwendet, um die Häufigkeit eines n-Gramms innerhalb eines Satzes anzugeben.

Bei der Angabe von Frequenzen höherer n-Gramme ($n > 1$), wird jedes n-gram nur einmal gezählt, d.h. es gibt keine Überlappungen. Im Beispiel „Das Buch war sehr sehr gut." wird das Bigramm „sehr gut" also nur einmal gezählt, im Beispiel „Verpackung: sehr gut, Inhalt: sehr gut" dagegen zweimal.

4.3.3.2 *Signifikanzwert der Kookkurrenz*

Im Forschungsgebiet der Sprachstatistik sind Kookkurrenzmaße sehr bekannt und dienen zur Bestimmung der Signifikanz von gemeinsam auftretenden Wörtern. Im Allgemeinen wird zwischen Satz- und Nachbarschaftskookkurrenz unterschieden. Satzkookkurrenzen bezeichnen das gemeinsame Auftreten von lexikalischen Einheiten – meist Wörtern – in Sätzen, wohingegen bei Nachbarschaftskookkurrenzen nur benachbarte lexikalischen Einheiten betrachtet werden. Ein anschauliches Beispiel für signifikante Kookkurrenzen gibt das Bigramm „Harry Potter", das bei der Analyse von Buchrezensionen auftrat. In diesem Fall trat das Unigramm „Potter" in allen Sätzen fast ausschließlich zusammen mit dem Unigram „Harry" auf.

Für die Bestimmung der Signifikanz von Kookkurrenzen existieren verschiedene Arten von Signifikanzmaßen, wie etwa das

Log-Likelihood Maß von Dunning [15]. Die Grundlage für das hier vorgestellte Verfahren bildet das Signifikanzmaß von Quasthoff and Wolff [41], welches in der Veröffentlichung von Heyer et al. [24] detailliert beschrieben wurde. Mit Hilfe dieses Maßes konnten bereits im Vorfeld plausible Ergebnisse mit dem verwendeten Korpus erzielt werden.

Gleichung 9 ist die aus [24] entnommene Originalformel. N ist dabei definiert als die gesamte Anzahl an Sätzen. $\lambda = \frac{f(a) \cdot f(b)}{N}$, wobei $f(x)$ als die Anzahl der Sätze definiert ist, in denen Wort x vorkommt. λ ist somit der Erwartungswert für das gemeinsame Auftreten von a und b, bei Annahme der stochastischen Unabhängigkeit des Vorkommens von a und b. Der Parameter k wird schließlich definiert als die Anzahl von Sätzen, die sowohl a als auch b enthalten.

$$sig(a,b) \approx \frac{\lambda - k \cdot \ln \lambda + \ln k!}{\ln N} \tag{9}$$

Durch Verwendung der bekannten Stirling-Formel kann für $k!$ ein Näherungswert mit $k! \approx \sqrt{2\pi k} \cdot \left(\frac{k}{e}\right)^k$ bestimmt werden. Für große Werte von k ist $ln(k!)$ somit näherungsweise $ln(k!) \approx k \cdot ln(k) - k$. Diese Näherung führt zu Formel 10, die aus Gründen der besseren Performance im Verfahren verwendet wurde.

$$sig(a,b) \approx \frac{\lambda + k \cdot (\ln k - \ln \lambda - 1)}{\ln N} \tag{10}$$

Um diese Formel für Mehrwortvorkommen mit $n > 2$ zu verwenden, schlagen Quasthoff and Wolff [41] folgende Anpassung vor: Um einen Signifikanzwert für eine Phrase mit s Wörtern $a_1, \ldots, a_s$ mit den Wahrscheinlichkeiten $P_1 = \frac{f(a_1)}{N}, \ \ldots \ , P_s = \frac{f(a_s)}{N}$ zu bestimmen, muss λ wie folgt berechnet werden: $\lambda = N \cdot P_1 \cdot P_2 \cdot \cdots \cdot P_s$. Dies führt schließlich zu Formel 11.

$$sig(a_1, ..., a_s) \approx \frac{\lambda + k \cdot (\ln k - \ln \lambda - 1)}{(s-1) \ln N} \quad (11)$$

Das Maß für signifikante Kookkurrenzen ist ein wesentlicher Baustein des vorgestellten Verfahrens. Zum einen wird es zur Identifikation von *n*-Grammen verwendet, die als Kandidaten für die lexikalische Ressource in Frage kommen – dazu wird der Schwellenwert $\hat{sig}_c$ verwendet – und zum anderen dient es der Bestimmung von unerwünschten *n*-Grammen für die *Entity Blacklist*. Diese Blacklist enthält Einträge, die nicht im Lexikon vorkommen sollen. Dazu wird der Schwellenwert $\hat{sig}_p$ verwendet (siehe Abschnitt 4.3.6.1).

4.3.3.3 *Meinungswert (SV) mit Standardabweichung*

In den durch das neue Verfahren erzeugten lexikalischen Ressourcen wird der Meinungswert eines *n*-Gramms, wie bei anderen Lexika allgemein üblich (siehe Kapitel 2.3.2), durch eine kontinuierliche Skala $[-1, +1]$ ausgedrückt. -1 steht dabei für maximal negativ und $+1$ für maximal positiv. Da die originalen Bewertungen von Amazon und anderen Plattformen meist auf einer diskreten, gleich verteilten Skala[3] $[r_{min}, r_{max}]$ liegen, werden zur Umrechnung die folgenden Formeln definiert und verwendet. Für jedes, in den Rezensionstiteln vorkommende, *n*-Gramm i wird die durchschnittliche Bewertung R_i der Bewertungen r_i^j aller m_i Rezensionstitel berechnet (siehe Gleichung 12).

$$R_i = \frac{\sum_{j=1}^{m_i} r_i^j}{m_i} \quad (12)$$

3 Um das für die Berechnung der Meinungswerte benötigte arithmetische Mittel bestimmen zu können, wird Äquidistanz – also gleiche Abstände zwischen den einzelnen Skalenpunkten – angenommen. Das bedeutet, dass der Unterschied zwischen 1 Stern und 2 Sternen genau dem zwischen 4 Sternen und 5 Sternen entspricht.

Anschließend wird der Sentiment Value $SV_i \in [-1, +1]$ durch die Transformation der Durchschnittsbewertung R_i auf die Skala $[-1, +1]$ bestimmt (siehe Gleichung 13). Dafür wird $\bar{r}$ als das arithmetische Mittel von r_{min} und r_{max} definiert.

$$SV_i = \frac{R_i - \bar{r}}{\frac{r_{max} - r_{min}}{2}} \tag{13}$$

Für die Experimente in dieser Arbeit muss schließlich die originale Skala von Amazon $[+1, +5]$, die auf Sternebewertungen beruht, auf die hier verwendete Skala $[-1, +1]$ konvertiert werden. Unter Anwendung von Formel 13 führt dies zu Formel 14, die im Folgenden verwendet wird.

$$SV_i = \frac{R_i - 3}{2} \tag{14}$$

Zusätzlich zum Sentiment Value SV wird die Standardabweichung σ_{SV} des Sentiment Values berechnet. Dieses Maß gibt an, wie stark die Bewertungen der jeweiligen Meinungsphrasen streuen und damit, wie plausibel der SV ist.
Die Formel zu Berechnung von Sentiment Values wurde in einem eigenen Konferenzbeitrag [48] bereits veröffentlicht.

4.3.4 *Datenvorverarbeitung*

Die Datenvorverarbeitung ist ein wichtiger Schritt auf dem Weg zu einer möglichst sauberen lexikalischen Ressource. Zuerst wurden die Titel sowie die dazugehörigen Sternebewertungen der Amazon Rezensionen aus dem Textkorpus extrahiert. Dabei wurden alle fremdsprachigen Titel durch eine auf statistischen Methoden basierende Spracherkennung [51] identifiziert und entfernt.

In einem nächsten Schritt wurden alle Titel bzw. Sätze in einzelne Token, d.h. in *Unigramme*, zerlegt. Dabei erfolgte die Trennung anhand einfacher Regeln unter Einbeziehung von Leer- und Satzzeichen, um Sprachunabhängigkeit zu gewährleisten. Nach dieser Zerlegung wurden alle Duplikate entfernt. Außerdem wurden alle Unigramme in Kleinbuchstaben umgewandelt, um die Analyse zu vereinfachen und die verschiedenen Schreibweisen der Rezensenten – „Super", „SUPER", „super" etc. – sowie großgeschriebene Wörter am Satzanfang, zu vereinheitlichen. Ein, an dieser Stelle, optionaler Schritt ist das Herausfiltern von bekannten Produktnamen, auf welche sich die Rezensionen beziehen. Allerdings kam dieser Korpus-spezifische Filterschritt nicht zum Einsatz, da er zum einen eine Abhängigkeit des generischen Verfahrens schafft und zum anderen den wichtigeren Schritt zur Filterung von Eigennamen (siehe Kapitel 4.3.6) nur erweitert.

Im nachfolgenden Schritt wurden alle Unigramme entfernt, die andere Zeichen als a-z, ä, ö, ü, ß, é, Zahlen, Apostrophe, Bindestriche oder Schrägestriche enthielten. Des Weiteren wurden alle Unigramme entfernt, die, mit Ausnahme von Zahlen, nur aus einzelnen Zeichen bestanden. Außerdem wurden alle Unigramme, die aus mindestens zwei Zeichen bestanden und mindestens eine Zahl oder ausschließlich Konsonanten enthielten, wie beispielsweise „ssd", in die Blacklist aufgenommen. Die so sukzessiv entstandene Blacklist wurde bei der späteren Bereinigung der Liste eingesetzt (siehe Abschnitt 4.3.6). Ein weiterer optionaler Schritt ist die Aufnahme von meinungstragenden Begriffen, die die oben genannten Kriterien für ein Unigramm nicht erfüllen, wie beispielsweise „1a", in eine Whitelist. Diese Option wurde für die Generierung der lexikalischen Ressource allerdings nicht verwendet.

Der letzte Schritt der Datenvorverarbeitung ist die Berechnung der Frequenz aller Unigramme im Textkorpus.

Algorithmus 1 fasst alle Schritte der Datenvorverarbeitung nochmals in Pseudocode zusammen.

Algorithmus 1 : Datenvorverarbeitung

```
Input : list of all review titles rl, blacklist bl, whitelist wl
Result : list of cleaned review titles cl, list of unigrams ul
for each review title in rl do
    if language is correct then
        add review to cl;
    end
end
for each title in cl do
    split title into single words and eliminate duplicates;
    for each word do
        convert to lower case;
        if word only contains allowed characters and word length ≥ 2 then
            if word already in ul then
                increase frequency of word by 1;
            else
                add word to ul with frequency of 1;
            end
        else
            if word consists of ≥ 2 characters and
               (≥ 1 digit or exclusively consonants) and
               word not in wl [optional] then
                add word to bl;
            end
        end
    end
end
```

4.3.5 *Aufbau des Lexikons durch Einsatz einer Graphdatenbank*

Nach Abschluss der Datenvorverarbeitung wurden Sätze sowie entstandene Unigramme genutzt, um den in Kapitel 4.2.1 beschriebenen Graph aufzubauen.

Da die klassische Brute-Force-Methode, bei der alle möglichen Kombination gebildet würden, für die Erzeugung des Graphen sehr zeitaufwendig wäre und den Graphen zudem unnötig vergrößern würde, wurde eine elegantere Methode, basierend auf

dem AprioriGen Algorithmus [2], gewählt, welche später in Kapitel 4.3.5.2 detailliert beschrieben wird.

Der Algorithmus zum Aufbau des Graphen beginnt mit dem Einfügen von Sätzen und Unigrammen und fügt dann sukzessive höhere n-Gramme ein. Algorithmus 2 fasst den komplexen Vorgang in Pseudocode zusammen.

Algorithmus 2 : Aufbau des Graphen

Input : list of sentences *sl*, list of unigrams *ul*
Data : list of n-grams *nl*
Result : complete graph database *gdb*

for *each sentence in sl* **do**
  insert sentence node in *gdb*;
end
for *each unigram in ul* **do**
  insert unigram node in *gdb*;
  link unigram node to each sentence node containing the unigram by creating an occurrence edge and set value for the within sentence frequency;
  calculate SV_i and σ_{SV_i} and update unigram node;
  calculate J-shaped-corrected SV_i and update unigram node;
  insert unigram into *nl*;
end
set n to 2;
repeat
  create n-grams C by using AprioriGen based on the current list *nl*;
  cache the generating $(n-1)$-grams A_C and B_C according to algorithm 3 for each n-gram C;
  overwrite *nl* with the list of all created n-grams C;
  for *all C in nl* **do**
    choose among A_C and B_C the one with the lowest frequency as L;
    determine the set *ts* with all target sentences of occurrence edges for L;
    delete all sentences from *ts* not containing C;
    calculate the within sentence frequency $f_{ws}(s, C)$ for each $s \in ts$;
    calculate the frequency $f(C)$ as the cardinality of *ts*;
    calculate the significance value $sig(C)$;
    if $f(C) > \hat{f}_{n-gram}$ *and* $sig(C) > \hat{sig}_c$ **then**
      insert a phrase node for C into *gdb*;
      create occurrence edges for C to each sentence s in *ts*;
      annotate these edges with $f_{ws}(s, C)$;
      determine all $(n-1)$-grams to be linked as parent nodes to C as follows:
      begin
        take an arbitrary sentence s out of *ts*;
        select $(n-1)$-gram nodes assoc. with s by an occurrence edge in a set *ags*;
        delete $(n-1)$-grams from *ags* containing words not present in C;
      end
      create sub-phrase edges connecting C with all elements of *ags*;
      calculate SV_C and σ_{SV_C} and update C;
    else
      delete C from *nl*;
    end
  end
  increment n by 1;
until *nl contains less than two elements*;

4.3.5.1 *Einfügen von Sätzen und Unigrammen*

Zunächst wurden die Sätze als *Sentence nodes* in den Graph eingefügt. Für jeden *Sentence node* wurden dazu der Satz als Text, die Anzahl an zugeordneten Sternen (Bewertung) sowie eine eindeutige ID gespeichert. Die Zeilen 1 bis 3 in Algorithmus 2 zeigen diese Vorgänge.

Nach dem Einfügen der *Sentence nodes*, wurden die im Datenvorverarbeitungsschritt erzeugten Unigramme als *Unigram nodes* eingefügt. Jeder *Unigram node* wurde mit Kanten vom Typ *Occurrence edge* mit jedem *Sentence node* verbunden, welcher das entsprechende Unigramm enthielt. Die Kanten wurden mit der entsprechenden Frequenz versehen, siehe Zeile 6 in Algorithmus 2. Während des Einfügevorgangs wurden ebenfalls vorläufige Sentiment Values SV_i mit der dazugehörigen Standardabweichung σ_{SV_i} berechnet. Außerdem wurde an dieser Stelle die Korpusspezifische Verteilung der Bewertungen korrigiert. Diese „J-shaped"-Verteilung tritt üblicherweise bei Kundenrezensionen mit einer 5-stufigen Skala auf und muss mittels der in [47] vorgeschlagenen Formel korrigiert werden. Ohne diese Korrektur erhalten neutrale Unigramme, die gleich verteilt in allen Rezensionen vorkommen, wie beispielsweise die Farbe „blau", einen zu hohen, und damit nicht korrekten, Sentiment Value ($SV \approx 0,6$). Durch diesen Vorgang wurden ca. 31% der Unigramme korrigiert und erhielten einen Sentiment Value $SV \approx 0$.

4.3.5.2 *Generierung höherer n-Gramme mit AprioriGen*

Im nächsten Schritt wurden alle eingefügten Unigramme miteinander kombiniert um Bigramme zu generieren, was zu einer quadratischen Komplexität führte. Dabei waren auch Kombinationen eines Unigramms mit sich selbst erlaubt. Um die Konsistenz bei der Anordnung der einzelnen Wörter bei Bi- bzw. n-Grammen im Allgemeinen zu gewährleisten, wurden die Wörter lexikogra-

phisch angeordnet. Das heißt, ein n-Gramm A wird durch eine Liste von Wörtern $\langle a_1 \dots a_n \rangle$ mit $a_i \leq_{lex} a_{i+1}$ für $i \in \{1, \dots, n-1\}$ repräsentiert. Durch die zugelassene Gleichstellung von Wörtern werden auch Kombinationen, wie beispielsweise $\langle very, very \rangle$, zugelassen.

Die Generierung höherer n-Gramme, durch die Kombination aller $(n-1)$-Gramme, ist durch die immense Anzahl erzeugter n-Gramme sehr zeitaufwendig. Bei Versuchen zeigte sich, dass die AprioriGen[4] Subroutine [1], ein Algorithmus, der für die Identifikation häufiger Item-Mengen im Rahmen von Warenkorbanalysen eingesetzt wird, einen guten Ansatz zur Lösung dieses Problems darstellt. Die AprioriGen Funktion verwendet ein Set eingefügter $(n-1)$-Gramme und generiert daraus ein Set häufiger n-Gramme, die oft genug vorkommen, um in den Graph aufgenommen zu werden. Dazu sind zwei Schritte notwendig. Im ersten Schritt, *Join* genannt, werden alle $(n-1)$-Gramme miteinander kombiniert. Für jedes Paar A, B mit $A = \langle a_1, \dots, a_{n-1} \rangle$ und $B = \langle b_1, \dots, b_{n-1} \rangle$ mit $a_i = b_i$ für $i \in \{1, \dots, n-2\}$ und $a_{n-1} \leq_{lex} b_{n-1}$ wird das n-Gram $\langle a_1, \dots, a_{n-1}, b_{n-1} \rangle$ in ein Set potentiell relevanter n-Gramme aufgenommen. Da auch jedes n-Gramm mit sich selbst kombiniert wird, führt dies auch zum n-Gramm $\langle a_1, \dots, a_{n-1}, a_{n-1} \rangle$.

Im zweiten Schritt, *Prune* genannt, werden alle n-Gramme aus dem vorher erzeugten Kandidatenset entfernt, die mindestens ein $(n-1)$-Subset enthalten, für welches kein entsprechender Knoten im Graph existiert. Da ein solches $(n-1)$-Subset aufgrund zu geringer Häufigkeit nicht im Graph vertreten ist, kann auch das darauf aufbauende n-Gramm nicht häufig genug sein.

Algorithmus 3 fasst den AprioriGen Algorithmus in Pseudocode zusammen. AprioriGen wird in Algorithmus 2 in Zeile 13 aufgerufen.

4 Die AprioriGen Funktion ist Teil des Apriori Algorithmus, der hauptsächlich bei der Warenkorbanalyse einsetzt wird.

Algorithmus 3 : AprioriGen

Input : input set of $(n-1)$-grams I
Result : set of relevant n-grams C_n
initialize C_n (list of candidate n-grams) as an empty set;
for *each* $A = \langle a_1, \ldots, a_{n-1} \rangle \in I$ **do**
  for *each* $B = \langle b_1, \ldots, b_{n-1} \rangle \in I$ **do**
    if $a_i = b_i$ *for* $i \in \{1, \ldots, n-2\}$ *and* $a_{n-1} \leq_{lex} b_{n-1}$ **then**
      add $\langle a_1, \ldots, a_{n-1}, b_{n-1} \rangle$ to C_n
    end
  end
end
for *each n-gram* $C = \langle c_1 \ldots c_n \rangle$ *in* C_n **do**
  for *each subset* $D \subset C$ *with* $(n-1)$ *elements* **do**
    if $D \notin I$ **then**
      delete C from C_n;
    end
  end
end
return C_n;

4.3.5.3 *Einfügen höherer n-Gramme*

Die durch die AprioriGen Subroutine erzeugten n-Gramme wurden anschließend in den Graphen eingefügt, falls deren jeweilige Häufigkeit dem Schwellenwert $\hat{f}_{n-gram}$ entsprach oder diesen überstieg.

Dazu wurden alle n-Gramme in einer Schleife – siehe Zeile 16 in Algorithmus 2 – wie folgt bearbeitet. Um zu überprüfen, ob ein n-Gramm C häufig genug in den Ausgangssätzen vorkam, musste zunächst die Frequenz des betrachteten n-Gramms berechnet werden. Um dies zu vereinfachen und die Laufzeit minimal zu halten, wurde die Tatsache genutzt, dass alle Sätze, die C enthalten per Definition auch die, durch AprioriGen generierten, $(n-1)$-Gramme A und B enthalten müssen. In Algorithmus 2 werden diese $(n-1)$-Gramme als A_C und B_C bezeichnet.

Die dazugehörigen *Phrase nodes* für A_C und B_C wurden bei der Generierung der n-Gramme C, durch die AprioriGen Subroutine, zwischengespeichert. Um möglichst wenige Sätze betrachten zu müssen, wurde das $(n-1)$-Gramm mit der niedrigsten Häufig-

keit (L) als Startpunkt ausgewählt. Von diesem Punkt aus wurde über die jeweiligen *Occurrence edges* zu allen *Sentence nodes* navigiert und überprüft, ob diese Sätze das entsprechende n-Gramm C enthielten. Sätze die C nicht enthielten wurden aus dem entsprechenden Set ts gelöscht. Die Zeilen 17 bis 19 in Algorithmus 2 zeigen diesen Vorgang.

Zudem wurde die Häufigkeit jedes n-Gramms C innerhalb eines Satzes bestimmt (siehe Abschnitt 4.3.3.1), da diese Information für die Beschriftung der jeweiligen *Occurrence edges* benötigt wurde (Zeile 20 in Algorithmus 2). Die Anzahl der Sätze, die C enthalten, wurde schließlich von der Kardinalität von ts abgeleitet und für die Überprüfung gegen den Schwellenwert $\hat{f}_{n-gram}$ genutzt (Zeile 21). Anschließend wurde der Signifikanzwert $sig(C)$ nach der Formel in Abschnitt 4.3.3.2 berechnet. Basis für diese Berechnung waren die $(n-1)$-Gramme, die via *Sub-phrase edges* mit dem entsprechenden n-Gramm verbunden waren (Zeile 22).

Schließlich wurden für alle n-Gramme C, die die Schwellenwerte passierten, *Phrase nodes* in den Graphen eingefügt. Die entsprechenden *Occurrence edges* zu den Satzknoten in ts wurden erzeugt und die *Sub-phrase edges* der entsprechenden $(n-1)$-Gramme hinzugefügt. Zum Schluss wurden die jeweiligen Sentiment Values (SV) und dazugehörigen Standardabweichungen (σ_{SV}) für alle n-Gramme C berechnet (Zeile 24 bis 34).

4.3.6 *Bereinigungen der Daten*

Bei der anschließenden Bereinigung der Daten wurden die vorläufigen Zwischenergebnisse in drei Schritten weiter verfeinert. Im ersten Schritt (Abschnitt 4.3.6.1) wurden, durch die Verwendung eines Bootstrapping-Algorithmus, Eigennamen entfernt. Im zweiten Schritt (Abschnitt 4.3.6.2) wurden alle Werte der n-Gram-

me im Graph neu berechnet, um unvollständige Phrasen zu identifizieren und anschließend zu entfernen. Im dritten Schritt (Abschnitt 4.3.6.3) wurden schließlich alle Phrasen mit einer hohen Standardabweichung aus dem Lexikon entfernt. Alle Schritte zur Bereinigung der Daten werden im Folgenden detailliert beschrieben.

4.3.6.1 *Entfernung von Eigennamen*

Der erzeugte Graph enthielt direkt nach der Erzeugung noch Eigennamen. Da solche Eigennamen in den meisten Fällen nicht als meinungstragende Phrasen für das Opinion Mining verwendet werden[5], mussten diese entfernt werden. Dazu wurden jedoch umfassende Listen benötigt, die die verschiedenen Typen von Eigennamen – Firmen- bzw. Produktnamen und Personennamen – enthielten. Als Ausgangsliste für Firmen- bzw. Produktnamen wurde die in der Datenvorverarbeitungsphase (siehe Abschnitt 4.3.4) erzeugte Blacklist verwendet. Für Personennamen kam eine Liste mit Vornamen zum Einsatz, die von dem Datenportal *Offene Daten Berlin* [37] bezogen wurde. Für beide Typen von Eigennamen wurde schließlich ein Bootstrapping-Algorithmus (siehe Algorithmus 4 auf Seite 67) verwendet, um die initialen Listen sukzessive zu erweitern.

Die Funktionsweise des eingesetzten Bootstrapping-Algorithmus beruht auf der Tatsache, dass Namen häufig miteinander zusammen auftreten. Beispiele dafür sind "Samsung" und „S3" für Firmen- und Produktnamen sowie „Stephen" und „King" für Vor- und Nachnamen.

Kandidaten für Eigennamen, die zu den Listen hinzugefügt werden sollten, waren Unigramme, die

5 Es existieren jedoch Ausnahmen, wie im Beispiel „Dieser Film verdient einen *Oscar*".

1. eine hohe Wahrscheinlichkeit bezüglich gemeinsam auftretender Unigramme, die bereits als Namen identifiziert wurden, aufwiesen und

2. mindestens einen hohen Signifikanzwert für das gemeinsame auftreten mit einem anderen Unigramm, das bereits als Name identifiziert wurde, hatten.

Beginnend mit der initialen Liste l_0 generierte der eingesetzte Bootstrapping-Algorithmus schließlich eine Serie von Unigramm-Listen l_i. Die Frequenz der Bigramme, die Unigramm a enthielten, wurde wie folgt berechnet:

$$f_{bi}(a) = \sum_{C \in \{\langle a,b\rangle \in gdb\} \cup \{\langle b,a\rangle \in gdb\}} f(C) \tag{15}$$

In Schritt i des Algorithmus wurde die Frequenz der Bigramme, die Unigramm a zusammen mit einem Eigennamen b, der bereits in der Liste enthalten war ($b \in l_{i-1}$), wie folgt berechnet:

$$f_{bi,l}(a) = \sum_{C \in \{\langle a,b\rangle \in gdb \mid b \in l_{i-1}\} \cup \{\langle b,a\rangle \in gdb \mid b \in l_{i-1}\}} f(C) \tag{16}$$

Durch die Verwendung der beiden Frequenzen konnte die Wahrscheinlichkeit, dass das gemeinsam auftretende Unigramm Teil der Eigennamen-Liste war, wie folgt berechnet werden:

$$P_{bi,l}(a) = \frac{f_{bi,l}(a)}{f_{bi}(a)} \tag{17}$$

Zudem wurde der maximale Signifikanzwert aller Bigramme, die Unigramm a sowie einen Namen b, der bereits in der Liste mit Eigennamen vorhanden war, wie folgt bestimmt:

$$sig_{bi}^{\max}(a) = \max_{C \in \{\langle a,b\rangle \in gdb \mid b \in l_{i-1}\} \cup \{\langle b,a\rangle \in gdb \mid b \in l_{i-1}\}} sig(C) \tag{18}$$

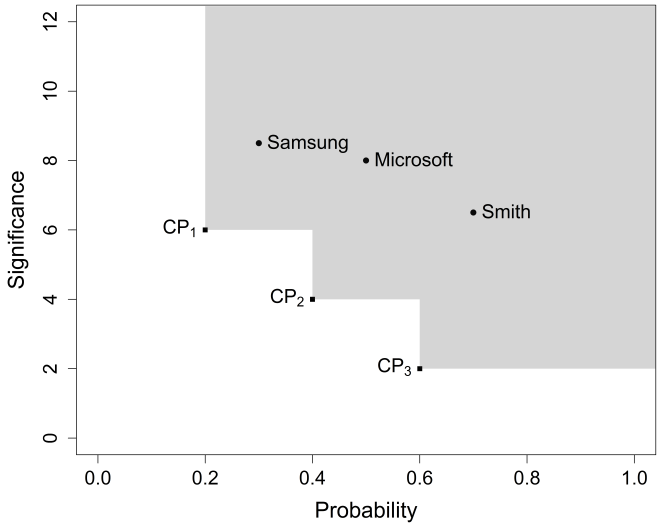


Abbildung 9: Beispiele von Eigennamen in einem durch die Schwellenwert-Eckpunkte CP_1, CP_2 und CP_3 definierten Bereich, gemäß des verwendeten Datensatzes

Durch die Verwendung von $P_{bi,l}(a)$ und $sig^{\max}_{\mathcal{G}bi}(a)$ konnte, wie in Abbildung 9 dargestellt, ein Koordinatenraum aufgespannt werden, bei dem Eigennamen normalerweise in den oberen rechten Bereich fallen. Der exakte Bereich wurde dabei mittels eines heuristischen Verfahrens bestimmt und die Schwellenwert-Eckpunkte des Koordinatensystems so festgelegt. Für den verwendeten Datensatz (siehe Abschnitt 4.3.2) wurden Werte für $P_{bi,l}(a)$ und $sig^{\max}_{\mathcal{G}bi}(a)$ anhand von Unigrammen gewählt, die als Eigennamen klassifiziert bzw. nicht klassifiziert werden sollten. Davon ausgehend wurden die Paare $CP_1(\hat{P}_1, \hat{sig}_{p1})$, $CP_2(\hat{P}_2, \hat{sig}_{p2})$ und $CP_3(\hat{P}_3, \hat{sig}_{p3})$ bestimmt, die in Abbildung 9 dargestellt werden.

Ein verbesserter Ansatz zur automatischen Bestimmung der benötigten Schwellenwerteckpunkte für verschiedene Datensätze ist dabei ein Anknüpfungspunkt für zukünftige Arbeiten.

Algorithmus 4 : Bootstrapping-Algorithmus für die Identifikation von Eigennamen

Input : graph database *gdb*, initial list l_0, set of threshold corner points *cps*
Result : enlarged list l_i
set i to 0;
repeat
  increment i by 1;
  initialize l_i by l_{i-1};
  for *each unigram a in gdb* **do**
    determine $P_{bi,l}(a)$ and $sig_{bi}^{\max}(a)$;
    for *each pair* $CP = (\hat{P}, \hat{sig}_p)$ *in cps* **do**
      if $P_{bi,l}(a) > \hat{P}$ *and* $sig_{bi}^{\max}(a) > \hat{sig}_p$ **then**
        add a to l_i;
      end
    end
  end
until $l_i = l_{i-1}$;

Durch einige weitere Anpassungen, kann die Genauigkeit der erzeugten Listen noch weiter verbessert werden. Um weitere Eigennamen in der Liste der Firmen- und Produktnamen zu identifizieren, könnte die finale Liste noch durch zusammengesetzte Wörter ergänzt werden.

Zum Schluss wurden alle identifizierten Eigennamen aus dem Graphen entfernt. Dazu wurden alle entsprechenden *Unigram nodes* sowie alle *n-gram nodes* entfernt, die diese Namen enthielten.

4.3.6.2 *Neuberechnung der n-Gramm-Werte*

An dieser Stelle enthielt der Graph, bedingt durch die Systematik des Aufbaus, noch unvollständige Phrasen wie „weder noch Fleisch“, als Teil der vollständigen Phrase „weder Fisch noch Fleisch“. Um diese überflüssigen Phrasen zu entfernen, wurde die Tatsache genutzt, dass die Frequenzen der unvollständigen Phrasen entweder gleich hoch oder nur wenig höher waren, als die Frequenzen der entsprechenden vollständigen Phrasen. Das reale Beispiel „weder Fisch noch Fleisch“ kam 65 mal in Sätzen vor, während „weder noch Fleisch“ 67 mal vorkam.

Die grundlegende Idee bei der Entfernung unvollständiger Phrasen war es, Sätze, die vollständige Phrasen enthalten, bei der Berechnung der Frequenzen nicht zu berücksichtigen. Durch dieses Vorgehen und durch die Anwendung des Schwellenwerts für die Frequenz $\hat{f}_{n-gram}$ konnten unvollständige Phrasen schließlich wie folgt entfernt werden.

Diese Methode basiert auf der Neuberechnung der Frequenzen mittels eines reduzierten Graphen, bei dem *Occurrence edges* nur für jene Phrasen betrachtet werden, die eine Meinung unabhängig von zusätzlichen Wörtern tragen. Algorithmus 5 zeigt die Umsetzung der Neuberechnungen in Pseudocode.

Die entsprechenden n-Gramme des Graphen wurden mit zunehmender Länge in der umschließenden While-Schleife bearbeitet. In den Zeilen 3 bis 8 wurden für alle n-Gramme die Teilphrasen ($(n-1)$-Gramme) ermittelt und anschließend alle *Occurrence edges* von diesen entfernt, die auf dieselben Sätze, wie das verbundene n-Gramm, verwiesen.

Um die Änderungen der Kanten im Graphen auch auf die jeweiligen Werte in den Phrasenknoten zu übertragen, mussten diese Werte neu berechnet werden (Zeilen 9 bis 19). Durch diese Neuberechnung fielen einige Phrasenknoten unter die Schwellenwerte und mussten aus dem Graph entfernt werden. Dies deckte sich mit der Annahme, dass solche Knoten oft nur unvollständige Phrasen darstellten – vollständige Phrasen verblieben dagegen im Graphen.

4.3.6.3 *Entfernung von Phrasen mit hoher Standardabweichung*

Im letzten Schritt der Datenbereinigung wurde die Standardabweichung der Sentiment Values σ_{SV_i} für jeden Knoten im Graph untersucht. Ein hoher Wert für σ_{SV_i} ist typisch für neutrale bzw. objektive n-Gramme, da diese häufig in allen Bewertungsklassen vorkommen. Bei einer angenommenen Gleichverteilung der

Algorithmus 5 : Bestimmung vollständiger Phrasen

```
   Input : graph database gdb
 1 set n to 2;
 2 while there are n-gram nodes in gdb do
 3 |   for each n-gram node C in gdb do
 4 |   |   determine all (n − 1)-grams D in gdb representing subsets of C;
 5 |   |   for each (n − 1)-gram D do
 6 |   |   |   delete all occurrence edges of D incident to sentences also
   |   |   |     incident to C via occurrence edges;
 7 |   |   end
 8 |   end
 9 |   for each (n − 1)-gram node D in gdb do
10 |   |   recalculate frequency f(D) based on the remaining occurrence
   |   |     edges;
11 |   |   if f(D) ≤ f̂_{n−gram} then
12 |   |   |   delete D from gdb together with all originating edges;
13 |   |   else
14 |   |   |   if n ≥ 2 then
15 |   |   |   |   recalculate sig(D);
16 |   |   |   end
17 |   |   |   recalculate SV_D and σ_{SV_D} and update D;
18 |   |   end
19 |   end
20 |   increment n by 1;
21 end
```

Werte im verwendeten 5-Sterne-Bewertungssystem, transformiert nach Formel 14, wäre $\sigma_{SV_i} = \frac{\sqrt{2}}{2}$. Ausgehend von dieser Theorie sollte der Schwellenwert für den Parameter $\hat{\sigma}$ im Bereich um $0,7$ gewählt und Phrasen, die diesen übersteigen, gelöscht werden.

Die Experimente zur endgültigen Bestimmung des Schwellenwerts werden in Kapitel 4.5 beschrieben.

Am Ende der Datenbereinigung wurde das finale Lexikon meinungstragender Wörter und Phrasen schließlich erzeugt. Dazu wurden die Einträge aller Phrasenknoten des Graphen, d.h. das entsprechende n-Gramm (Text) mit Frequenz, Sentiment Value SV_i sowie Standardabweichung σ_{SV_i}, in einer Lexikon-Datei abgespeichert.

4.4 Laufzeitmessung und Speicherbedarf

Die Laufzeit des beschriebenen Verfahrens zur Erzeugung lexikalischer Ressourcen hängt sehr stark von der Wahl des Schwellenwerts für die Mindesthäufigkeit der im Textkorpus vorkommenden Phrasen ($\hat{f}_{n-gram}$) ab. Tabelle 9 gibt einen Überblick über die Zeit in Minuten, die der Algorithmus in Abhängigkeit vom Schwellenwert benötigt, um den kompletten Graph mit meinungstragenden Wörtern und Phrasen aufzubauen. Zusätzlich enthält die Tabelle auch die absolute Anzahl der n-Gramme ($1 \leq n \leq 5$), die als Kandidaten für die lexikalische Ressource berücksichtigt werden. Außerdem ist zu beachten, dass die Experimente zur Laufzeit auf einem Standard-PC[6] durchgeführt wurden.

Schwellenwert	Zeit (min)	1-Gramme	2-Gramme	3-Gramme	4-Gramme	5-Gramme
1.000	116	675	228.150	1.676	4	-
500	174	1.205	726.615	6.117	35	1
400	184	1.460	1.066.530	9.280	93	6
300	236	1.847	1.706.628	14.177	245	20
200	293	2.647	3.504.628	28.404	651	78
100	484	4.700	11.047.350	83.598	3.031	426
75	631	5.954	17.728.035	129.162	5.563	607
50	834	8.137	33.109.453	236.513	12.936	1.062
30	1.235	11.888	70.668.216	497.685	34.076	2.735
25	1.553	13.534	91.591.345	645.949	47.499	3.959
15	2.724	19.500	190.134.750	1.340.425	116.979	10.698
10	5.601	26.057	339.496.653	2.372.280	232.993	22.681

Tabelle 9: Laufzeit des Algorithmus zur Erzeugung lexikalischer Ressourcen in Minuten in Abhängigkeit vom Schwellenwert für Parameter $\hat{f}_{n-gram}$

In Abbildung 10 wird zudem der Zusammenhang zwischen der Wahl des Schwellenwertes für $\hat{f}_{n-gram}$ und der Zeit, die das

6 CPU: AMD Phenom II X4 955 3,20 Ghz, RAM: 24,0 GB DDR3, Festplatte: Samsung SSD 830

Programm benötigte, um den kompletten Graph aufzubauen, dargestellt.

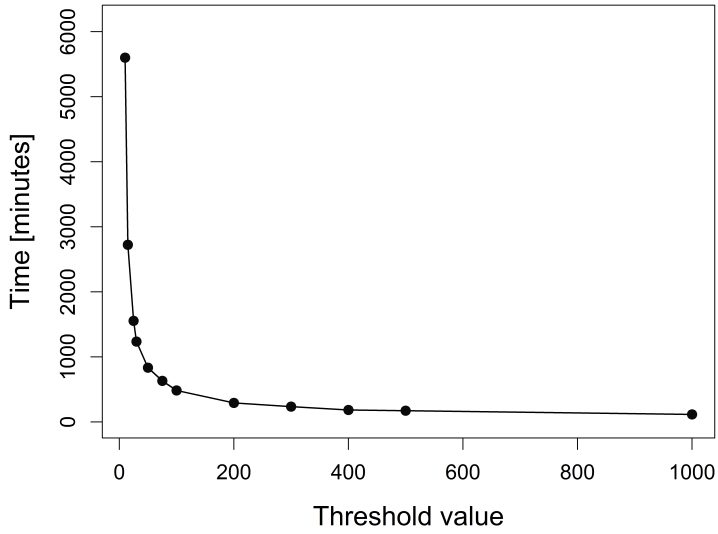


Abbildung 10: Verteilung der gemessenen Werte: Schwellenwert für $\hat{f}_{n-gram}$ und die benötigte Zeit zur Berechnung in Minuten

4.5 Schwellenwertbestimmung der Parameter

In diesem Abschnitt werden die Experimente zur Bestimmung der optimalen Schwellenwerte für die im Algorithmus verwendeten Parameter vorgestellt.

4.5.1 *Minimale n-Gramm Häufigkeit*

Wie in Kapitel 4.4 gezeigt werden konnte, hat der Parameter für die minimale Häufigkeit von n-Grammen $\hat{f}_{n-gram}$ sowohl große

Auswirkungen auf die Anzahl der *n*-Gramme, die in den Graph aufgenommen werden als auch auf die Laufzeit des Algorithmus.

Um den optimalen Schwellenwert für diesen Parameter zu finden, wurde ein Referenzset, bestehend aus manuell annotierten relevanten und nicht relevanten Unigrammen, erzeugt. Als Basis für dieses Set wurden alle Unigramme mit einer absoluten Häufigkeit kleiner oder gleich des niedrigsten gemessenen Schwellenwertes $\hat{f}_{n-gram} = 10$ und einem *Sentiment Value* $SV \geq 0.5$ verwendet. Durch diese Filterschritte wurden zunächst 13.057 Unigramme selektiert, von denen schließlich 10% zufällig ausgewählt wurden. Diese 1.305 Unigramme wurden durch zwei Annotatoren hinsichtlich der Relevanz für das Opinion Mining in *relevant* und *nicht relevant* klassifiziert. Die Interrater-Reliabilität (Cohens Kappa) der beiden Rater lag dabei bei $\kappa = 0.85$. Das so erzeugte Referenzset wurde anschließend genutzt, um den optimalen Schwellenwert für $\hat{f}_{n-gram}$ auf Basis von *Precision* und *Recall* zu bestimmmen. Dazu wurde für verschiedene Schwellenwerte die Anzahl der vorhandenen relevanten sowie nicht-relevanten Unigramme innerhalb der zurückgelieferten Ergebnismenge ermittelt. Dieser Vorgang wurde so lange mit wechselnden Schwellenwerten wiederholt, bis ein Optimum erreicht wurde.
Tabelle 10 gibt einen Überblick über die Ergebnisse der Versuchsreihe. Das F_1 Maß, als relevantes Kriterium, zeigt einen optimalen Wert bei $\hat{f}_{n-gram} = 29$.

4.5.2 *Maximale Standardabweichung von n-Grammen*

Wie bereits in Abschnitt 4.3.6.3 kurz diskutiert, sollten neutrale bzw. objektive Phrasen aus dem erzeugten Lexikon entfernt werden. Dafür musste zunächst ein geeigneter Schwellenwert für Parameter $\hat{\sigma}$ ermittelt werden. Dazu wurde eine Versuchsreihe mit einem manuell annotierten Referenzset durchgeführt. Das ur-

Schwellen-wert	Precision	Recall	F_1	relevante Unigramme	nicht relevante Unigramme
1.000	0,5263	0,0341	0,0641	10	9
500	0,5385	0,0717	0,1265	21	18
400	0,5217	0,0819	0,1416	24	22
300	0,5397	0,1160	0,1910	34	29
200	0,5106	0,1638	0,2481	48	46
100	0,4571	0,2730	0,3419	80	95
75	0,4208	0,3174	0,3619	93	128
50	0,4104	0,4300	0,4200	126	181
35	0,3583	0,5222	0,4250	153	274
30	0,3525	0,6075	0,4461	178	327
29	0,3532	0,6280	**0,4521**	184	337
28	0,3482	0,6382	0,4506	187	350
25	0,3250	0,6621	0,4360	194	403
15	0,2723	0,8328	0,4104	244	652
10	0,2245	1,0000	0,3667	293	1.012

Tabelle 10: Ergebnisse der Versuchsreihe zur Bestimmung des optimalen Parameters für $\hat{f}_{n-gram}$

sprüngliche Datenset für diesen Annotationsprozess wurde durch die Kombination aus 1) 10% zufällig ausgewählter Unigramme mit $\hat{f}_{n-gram} = 29$ und $SV \geq 0,5$ und 2) 10% zufällig ausgewählter Unigramme mit $\hat{f}_{n-gram} = 29$ und $SV \leq -0,5$ erzeugt. Zwei Annotatoren bekamen anschließend die Aufgabe, in diesem Ausgangsset die, für das Opinion Mining (d.h. die meinungstragenden), relevanten Unigramme zu markieren und somit ein neues Set für die Experimente mit verschiedenen Schwellenwerten für $\hat{\sigma}$ zu erzeugen. Die Interrater-Reliabilität der beiden Annotatoren lag bei $\kappa = 0,76$.

Tabelle 11 zeigt die Ergebnisse der anschließend durchgeführten Versuchsreihe. Wie erwartet lagen die im Vorfeld als sinnvoll erachteten Werte für einen geeigneten Schwellenwert der Standardabweichung um 0,7. Da für die weiteren Experimente ein

höherer *Recall* wichtiger war als eine gute *Precision*, wurde der endgültige Wert für den Parameter $\hat{\sigma}$ auf $0,7$ festgesetzt.

Schwellen-wert	Precision	Recall	F_1	relevante Unigramme	nicht relevante Unigramme
0,1	1,0000	0,0039	0,0078	1	0
0,2	0,6250	0,0394	0,0741	10	6
0,3	0,6196	0,2244	0,3295	57	35
0,4	0,4477	0,4213	0,4341	107	132
0,5	0,3527	0,6457	**0,4562**	164	301
0,6	0,3010	0,8071	0,4385	205	476
0,7	0,2757	0,9173	0,4240	233	612
0,8	0,2649	0,9803	0,4171	249	691
0,9	0,2613	1,0000	0,4144	254	718
1,0	0,2608	1,0000	0,4137	254	720

Tabelle 11: Ergebnisse der Versuchsreihe zur Bestimmung des optimalen Parameters für $\hat{\sigma}$

4.6 Ergebnisse

Durch die Verwendung des vorher beschriebenen Verfahrens zur Erzeugung einer neuen lexikalischen Ressource konnten, aus einem Amazon-Korpus mit 1,5 Millionen Rezensionen, 53.312 meinungstragende Wörter und Phrasen extrahiert werden. Tabelle 12 zeigt die Verteilung der extrahierten Wörter und Phrasen nach deren Phrasenlänge. Hierbei wird deutlich, dass der neue Algorithmus vor allem 1-Wort, 2-Wort, 3-Wort und 4-Wort Phrasen extrahiert. Der Grund dafür liegt zum einen an der Länge der durchschnittlichen Rezensionstitel und zum anderen an der selteneren Verwendung langer meinungstragender Phrasen.

	Phrasenlänge								
	1	2	3	4	5	6	7	8	Summe
#	8,517	25,927	15,098	2,962	604	126	53	25	**53,312**

Tabelle 12: Verteilung der extrahierten Wörter und Phrasen nach Phrasenlänge

4.6.1 *Auswahl generierter Wörter und Phrasen*

Tabelle 13 zeigt eine Auswahl meinungstragender Wörter und Phrasen aus der generierten lexikalischen Ressource. Die Tabelle enthält außerdem den *Sentiment Value* (SV), inklusive Standardabweichung (σ_{SV}), sowie die Häufigkeit der Phrasen.

Phrase	SV	*Frequency*	σ_{SV}
einfach fantastisch	1.00	213	0.00
meisterwerk	0.94	971	0.28
überdurchschnittlich	0.67	53	0.29
für den preis nicht schlecht	0.39	35	0.26
unrealistisch	-0.42	99	0.68
nicht zufriedenstellend	-0.53	91	0.32
bitter enttäuscht	-0.81	39	0.24
einfach nur schlecht	-1.00	500	0.00

Tabelle 13: Eine Auswahl meinungstragender Wörter und Phrasen, geordnet nach *Sentiment Value* (SV)

4.6.2 *Auswahl generierter Redewendungen*

Ein weiterer Aspekt dieses neuen Verfahrens ist die Erkennung von Redewendungen. Diese Wortverbindungen werden üblicherweise genutzt, um Produkte sowie deren Eigenschaften umgangssprachlich zu bewerten. Ein Beispiel gibt der Satz „Dieses Smartphone ist eine *eierlegende Wollmilchsau*.“. Insofern sind Redewendungen wichtige Kandidaten für die lexikalische Ressource. Wie bereits erwähnt, ist es für Unigramm-basierte Verfahren bzw. Ver-

fahren, die auf musterbasierten Methoden beruhen, nahezu unmöglich diese Arten von Phrasen zu erkennen.

Tabelle 14 zeigt reale Beispiele für Redewendungen, die im neu erzeugten Lexikon vorkommen – einschließlich deren Sentiment Values (SV) und Frequenzen. Die errechneten Sentiment Values der identifizierten Redewendungen erscheinen plausibel, sodass diese für die Verwendung in der lexikalische Ressource in Frage kommen.

Phrase	SV	*Frequency*	σ_{SV}
aber klein oho	0,90	494	0,20
aber fein klein	0,88	797	0,21
eierlegende wollmilchsau	0,84	82	0,25
alles glänzt gold ist nicht was	-0,07	42	0,38
fisch fleisch noch weder	-0,20	65	0,24
außen hui innen pfui	-0,54	139	0,4
bei bleib deinen leisten schuster	-0.75	43	0,34
gottes um willen	-0,88	35	0,21
finger weg	-1,00	1.356	0,00

Tabelle 14: Eine Auswahl gefundener Redewendungen, geordnet nach *Sentiment Value* (SV)

4.7 Zusammenfassung

Die ersten Experimente mit den deutschsprachigen Texten des Amazon-Korpus als Basis, lieferten zufriedenstellende Ergebnisse. Es wurden relevante meinungstragende Wörter und Phrasen extrahiert und diesen schlüssige Sentiment Values zugeordnet. Besonders die Identifikation (meinungstragender) Redewendungen des sprachunabhängigen Ansatzes konnte überzeugen.

Allerdings zeigten sich bei diesem Verfahren auch Probleme und Einschränkungen, die im Folgenden näher beschrieben werden und Anknüpfungspunkte für zukünftige Arbeiten sind.

1. Fehlende Reihenfolge der Wörter einer Phrase
   Die fehlende Reihenfolge der Wörter bei Phrasen kann unter Umständen zu fehlerhaften Berechnungen der Sentiment Values führen. Die Phrase „riesig aber wunderschön“ trägt bei anderer Wortreihenfolge („wunderschön aber riesig“) eine ganz andere Meinung. Zudem kann beim Einsatz in einem Analysesystem nicht erkannt werden, welche Kombination korrekt ist.

2. Auftreten unpassender Phrasen und Entitäten
   Durch die ausschließliche Verwendung statistischer Verfahren, erscheinen auch immer unpassende Phrasen und Entitäten, wie z.B. Produktnamen, in der lexikalischen Ressource. Aus diesem Grund sollte das Lexikon vor dem Einsatz in Opinion Mining Systemen nochmals manuell bereinigt werden.

3. Weitere Experimente für optimale Parametereinstellungen
   Viele der im Algorithmus verwendeten Parameter wurden durch einfache Experimente festgelegt. Die Änderung eines Parameters kann dabei sowohl Laufzeit, als auch Ergebnisse drastisch verändern. Daher sollten weitere Experimente zur optimalen Parametereinstellung durchgeführt werden, u.a. auch eine Sensitivitätsanalyse.

4. Experimente mit anderen Korpora
   Die durch den beschriebenen Algorithmus erzeugte lexikalische Ressource basiert auf Titeln der Rezensionen von Amazon Deutschland. Das heißt, dass wichtige Phrasen, die nicht häufig genug in diesem Korpus vorkommen, nicht Teil des erzeugten Lexikons sind. Daher sollten in einem nächsten Schritt auch Experimente mit anderen Korpora durchgeführt werden.

5. Fehlende Lemmatisierung der Wörter
   Die Vorteile des sprachunabhängigen Ansatzes werden an dieser Stelle zu einem Problem. Durch die fehlende Lemmatisierung von Wörtern können verschiedene Formen eines Wortes verschiedene Sentiment Values erhalten. Im Beispiel des Adjektivs „gut“, existieren im Lexikon drei verschiedene Formen: „gut“ – $SV = 0,69$, „guter“ – $SV = 0,70$ und „gutes“ – $SV = 0,71$. Zudem kann die fehlende Lemmatisierung dazu führen, dass verschiedene Formen eines Wortes nicht häufig genug auftreten, um als Kandidaten für die lexikalische Ressource berücksichtigt zu werden und dieses Wort (inklusive aller Formen) nicht Teil des Lexikons wird. Ein weiteres Problem dabei betrifft die Anwendung, bei der Wörter und Phrasen im zu analysierenden Text genauso vorkommen müssen wie in der lexikalischen Ressource abgebildet. Um dieses Problem zu lösen, müssten entweder alle fehlenden Wortformen dem Lexikon hinzugefügt, oder alle Einträge nachträglich lemmatisiert werden.

Trotz der aufgeführten Probleme wird durch den beschriebenen Algorithmus eine gute Möglichkeit geboten, Lexika meinungstragender Wörter und Phrasen einfach zu erstellen. Durch die ausschließliche Verwendung statistischer Methoden, statt der üblichen NLP-Methoden, können auch Lexika für andere Sprachen erzeugt werden, ein entsprechender Textkorpus vorausgesetzt.

5

MANUELLE ERZEUGUNG EINES REFERENZLEXIKONS

In diesem Kapitel wird die manuelle Erstellung eines Referenzlexikons mit meinungstragenden Adjektiven für das Opinion Mining für die deutsche Sprache beschrieben. Die Notwendigkeit für die Erstellung einer solchen Ressource wurde in Kapitel 1.3 deutlich gemacht und betrifft die Abwesenheit einer geeigneten Referenz als Baseline zur Evaluation bestehender deutscher Listen.

5.1 Vorüberlegungen und Konzeption

Um einen Überblick über die notwendigen Arbeitsschritte und Abläufe zu erhalten, wurde vorab ein Konzept entwickelt, in welchem die grundlegende Idee, mögliche Quellen sowie der Ablauf bei der Generierung des Referenzlexikons skizziert wurde. Außerdem wurden Annotationsregeln definiert und ein Vorschlag für eine Annotationsanleitung erarbeitet. Relevante Teile aus diesem Konzept sowie sich ergebende Abweichungen werden im Folgenden vorgestellt.

5.1.1 *Grundlegende Idee*

Wie in Kapitel 2.3.1 beschrieben, ist die manuelle Erzeugung einer lexikalischen Ressource ein zeitintensiver Vorgang. Dennoch sind so erzeugte Ressourcen, insbesondere für die Evaluation automatischer Verfahren, oft sinnvoll. Aus diesem Grund wurde für die Evaluation der Güte sowie der Vollständigkeit der in dieser

Arbeit vorgestellten deutschsprachigen lexikalischen Ressourcen (siehe Kapitel 2.3.2), inklusive des automatisch generierten Lexikons, das mittels des beschriebenen Algorithmus in Kapitel 4 im Rahmen dieser Arbeit erzeugt wurde, eine solche Referenzressource erstellt.

Die grundlegende Idee beim Aufbau dieser Ressource war vergleichsweise einfach. Ausgehend von einem Set, welches die wichtigsten – in diesem Fall gleichzusetzen mit *häufigsten* – meinungstragenden Adjektive[1] enthielt, sollten Versuchspersonen diesen Adjektiven Meinungsklassen zuordnen. Als Werkzeug für diese Annotation kam das Tabellenkalkulationsprogramm *Microsoft Excel* zum Einsatz.

Nach Abschluss des Experiments mussten die Ergebnisse der Versuchspersonen auf Übereinstimmung getestet und, bei genügend hoher Übereinstimmung, schließlich Meinungswerte für die Adjektive nach der in Kapitel 4.3.3.3 vorgestellten Formel berechnet werden.

Die so entstandene Ressource konnte anschließend zur Evaluation vorhandener deutschsprachiger Lexika verwendet werden (siehe Kapitel 6).

5.1.2 *Betrachtete und verwendete Quellen*

Als Basis für relevante deutsche Adjektive wurden drei Quellen betrachtet, auf die im Folgenden näher eingegangen werden soll:

1. Wikipedia (Deutsch)[2]
2. Wiktionary (Deutsch)[3]
3. Duden online[4]

1 Ursprünglich geplant waren Adjektive, Nomen und Verben. Aufgrund der Menge der Daten wurde diese Idee jedoch vorerst verworfen.

2 https://de.wikipedia.org/

3 https://de.wiktionary.org/

4 https://www.duden.de/

Als Basis für die Untersuchungen der *Wikipedia* Daten, wurde der deutsche Wikipedia-Korpus „diwiki“[5] (Stand: 10.10.2016) verwendet, welcher aus ca. 2,5 Millionen Artikeln mit ca. 46 Millionen Sätzen und ca. 10 Millionen verschiedenen Wörtern besteht. Die Texte dieses Korpus wurden mit Hilfe des Parsers „DizzyLogic“[6] extrahiert und als Klartext in einer eigenen Datenbank abgespeichert.
Die anschließende Analyse der häufigsten deutschsprachigen Adjektive zeigte nur einen geringen Anteil meinungstragender Begriffe. Da *Wikipedia* als Enzyklopädie hauptsächlich Fakten enthält, war dieses Ergebnis jedoch plausibel. Die häufigsten Adjektive, die durch die Analyse identifiziert wurden, waren Herkunftsbezeichnungen wie „amerikanisch“ oder „englisch“. Aufgrund dieser Ergebnisse wurde *Wikipedia* als Quelle für ein Referenzlexikon ausgeschlossen.

Bei der Untersuchung der *Wiktionary* Website wurde schnell deutlich, dass sich diese Quelle aufgrund fehlender Häufigkeitsangaben ebenfalls nicht als Basis für ein Referenzlexikon eignet. Die Relevanz von Adjektiven kann auf Grundlage dieser Quelle nicht ermittelt werden. Das freie Online-Wörterbuch kann allerdings als ergänzende Quelle für Adjektive verwendet werden, für die keine Flexionstabellen im *Duden online* (siehe folgenden Abschnitt) existieren.

Duden online listet, nach eigenen Untersuchungen, ca. 230.000 Wörter in lemmatisierter Form, davon sind 13,8% Adjektive. Abbildung 11 illustriert die Verteilung der Wortarten. Außerdem enthält die Website weitere wichtige Informationen zu jedem Wort, wie die Wortart und die Häufigkeit des Wortes in der deutschen Sprache. Die Angabe der Häufigkeit wird dabei in Form von „Häufigkeitsklassen“ angegeben, die in Abbildung 12 näher beschrieben werden.

5 https://dumps.wikimedia.org/dewiki/
6 Dieser Parser ist aktuell nicht mehr verfügbar.

Neben den oben genannten Informationen ist die Verfügbarkeit von Flexionstabellen für die meisten Wörter ein weiterer Vorteil des *Duden online*. Diese Tabellen haben Relevanz für die spätere Evaluation von lexikalischen Ressourcen, die unlemmatisierte Einträge enthalten. Durch die beschriebenen Vorzüge wurde diese Quelle schließlich als Basis für das Referenzlexikon ausgewählt.

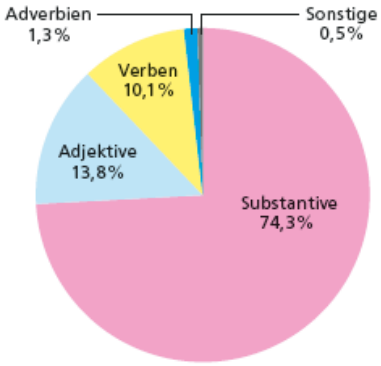
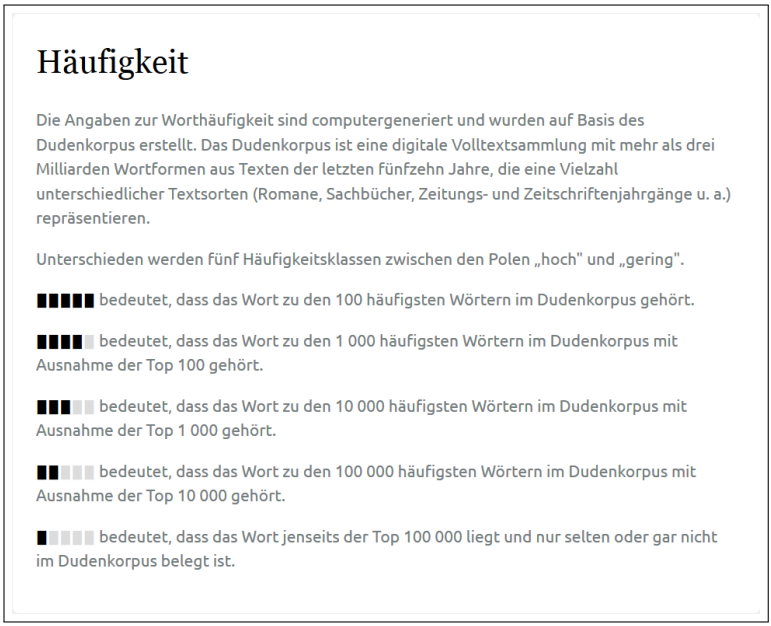


Abbildung 11: Verteilung der Wortarten im Duden-Korpus (Quelle: [21])

5.1.3 *Geplanter Ablauf bei der Erzeugung*

Abbildung 13 illustriert den geplanten Ablauf bei der Erzeugung des Referenzlexikons.

Die benötigten Daten mussten zunächst aus der gewählten Quelle extrahiert werden. Anschließend sollten aus den extrahierten Wörtern alle Adjektive herausgefiltert werden. Beim Duden-Korpus konnte dieser Filterschritt, durch die Verwendung der angegebenen Wortarten, sehr einfach durchgeführt werden.
Im nächsten Schritt wurde ein weiterer Filter verwendet, der ausschließlich häufige Adjektive als Basis für das Experiment zuließ. Dieser Schritt war von entscheidender Bedeutung, da eine Anno-



Häufigkeit

Die Angaben zur Worthäufigkeit sind computergeneriert und wurden auf Basis des Dudenkorpus erstellt. Das Dudenkorpus ist eine digitale Volltextsammlung mit mehr als drei Milliarden Wortformen aus Texten der letzten fünfzehn Jahre, die eine Vielzahl unterschiedlicher Textsorten (Romane, Sachbücher, Zeitungs- und Zeitschriftenjahrgänge u. a.) repräsentieren.

Unterschieden werden fünf Häufigkeitsklassen zwischen den Polen „hoch" und „gering".

■■■■■ bedeutet, dass das Wort zu den 100 häufigsten Wörtern im Dudenkorpus gehört.

■■■■ bedeutet, dass das Wort zu den 1 000 häufigsten Wörtern im Dudenkorpus mit Ausnahme der Top 100 gehört.

■■■ bedeutet, dass das Wort zu den 10 000 häufigsten Wörtern im Dudenkorpus mit Ausnahme der Top 1 000 gehört.

■■ bedeutet, dass das Wort zu den 100 000 häufigsten Wörtern im Dudenkorpus mit Ausnahme der Top 10 000 gehört.

■ bedeutet, dass das Wort jenseits der Top 100 000 liegt und nur selten oder gar nicht im Dudenkorpus belegt ist.

Abbildung 12: Häufigkeitsklassen – Worthäufigkeit im Duden-Korpus (Quelle: [22])

tation aller extrahierten Adjektive (nach ersten Schätzungen ca. 40.000) mit den vorhandenen Ressourcen nicht möglich gewesen wäre. Als Basis für das Experiment wurden schließlich die, von *Duden online* vorgegebenen, Häufigkeitsklassen 5, 4, und 3 gewählt. Durch diese Auswahl wurde das Set auf jene Adjektive beschränkt, welche zu den häufigsten 11.100 deutschen Wörtern gehören (ca. 1.500). Nach der grundlegenden Filterung wurde jedes Adjektiv aus dem Set der häufigen Adjektive in „Stufe 1" eines zweistufigen Verfahrens von mindestens 3 Personen hinsichtlich der Relevanz für das Opinion Mining klassifiziert. Durch diesen Vorgang sollten objektive, d.h. nicht meinungstragende, Adjektive herausgefiltert werden. Das, durch diesen Vorgang, verkleinerte Set subjektiver Wörter wurde in einem weiteren Schritt für die Annotation in „Stufe 2" vorbereitet. Dazu wurde für jedes

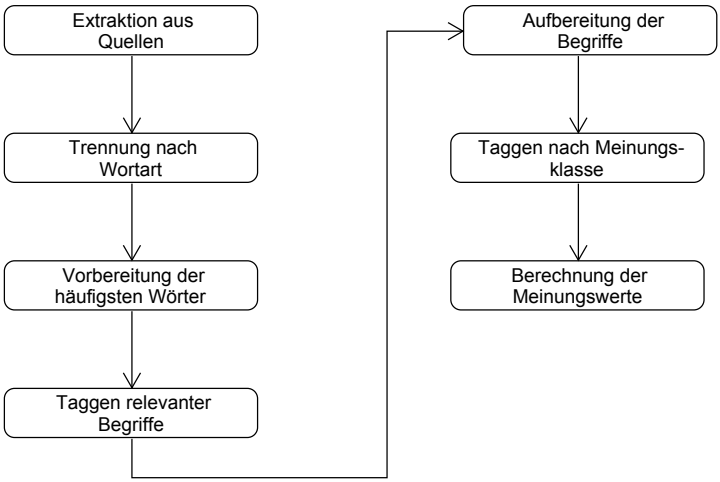


Abbildung 13: Ablauf bei der Erzeugung eines Referenzlexikons

gelistete Adjektiv in der Liste eine weitere Zeile mit der entsprechenden Verneinung hinzugefügt, beispielsweise wurde für „gut“ eine weitere Zeile eingefügt, die „nicht gut“ enthielt. Durch diesen Vorgang sollten Phrasen in das Referenzlexikon aufgenommen werden. In der nächsten Stufe des Annotationsprozesses („Stufe 2“) sollte eine weitere größere Gruppe von Personen allen Wörtern und Phrasen Meinungsklassen zuordnen. Folgende Meinungsklassen wurden vorgegeben: stark positiv, schwach positiv, neutral, schwach negativ und stark negativ. Im letzten Schritt wurde für jeden Eintrag der Referenzliste ein Meinungswert berechnet, wobei ein 5-Punkte-Bewertungssystem, wie nachfolgend beschrieben, verwendet wurde.

- stark positiv: 5 Punkte
- schwach positiv: 4 Punkte
- neutral: 3 Punkte

- schwach negativ: 2 Punkte
- stark negativ: 1 Punkt

Diese Art der Einteilung ermöglicht es, jedem Wort bzw. jeder Phrase des Referenzlexikons einen Meinungswert im Bereich $[-1, +1]$ zuzuordnen. Wie bereits in Kapitel 4.3.3.3 erwähnt, gilt auch bei dieser Skala die Annahme von Äquidistanz der fünf Gruppen, um Mittelwerte pro Wort bilden zu können. Die Berechnung wurde bereits ausführlich in erwähntem Kapitel beschrieben.

5.2 Vortest

Um die Machbarkeit der in Kapitel 5.1 getroffenen Annahmen zu überprüfen, wurden zwei Testläufe mit verschiedenen Vorgaben spezifiziert. Durch diesen Vorabtest sollte die Machbarkeit des Experiments überprüft und eventuelle Fehler des Settings im Vorfeld ausgeschlossen und rechtzeitig gegengesteuert werden.

Basis für diese Tests waren 10% zufällig ausgewählter Adjektive aus dem gesamten Set häufiger deutscher Adjektive. Die beiden folgenden Abschnitte beschreiben die Abläufe beim Annotieren der zufällig ausgewählten Adjektive. Anschließend werden die Ergebnisse zusammengefasst um daraus Erkenntnisse für das Experiment zur Erzeugung des Referenzlexikons abzuleiten.

5.2.1 *Testlauf für die Annotation der Relevanz*

Der erste Testlauf bezieht sich auf die Annotation relevanter Begriffe. Untersucht werden sollte die Übereinstimmung der Annotatoren – mittels der Maße Cohens Kappa [12] und Fleiss' Kappa [19] – bei der Aufgabe, aus dem Set der Adjektive jene zu markieren, welche eine Relevanz für das Opinion Mining besitzen.

Die Testteilnehmer bekamen dazu die Aufgabe, nur Begriffe zu markieren, die, für sich allein stehend, wirklich meinungstragend sind. Teilnehmer dieses Tests waren sieben Personen unterschiedlichen Geschlechts und Alters, mit unterschiedlichen Vorkenntnissen.

Bei der Analyse der Ergebnisse des Experiments zeigten sich große Differenzen zwischen den Annotatoren bezüglich der Relevanz der Adjektive. Die Interrater-Reliabilität der sieben Versuchsteilnehmer ergab nach Fleiss (Kapitel 3.1) einen Wert von lediglich $\kappa = 0,349$.

Tabelle 15 zeigt die Übereinstimmung der sieben Teilnehmer (A_n) im Paarvergleich. Als Maß für die Übereinstimmung wurde Cohens Kappa (Kapitel 3.1) verwendet. Es ist gut zu erkennen, dass die Übereinstimmung zwischen den Teilnehmern stark schwankte (zwischen 0,73 und 0,16) und nicht zufriedenstellend war.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7
A_1	-	0,73	0,31	0,38	0,47	0,42	0,22
A_2	0,73	-	0,39	0,47	0,48	0,44	0,28
A_3	0,31	0,39	-	0,54	0,16	0,28	0,34
A_4	0,38	0,47	0,54	-	0,22	0,33	0,32
A_5	0,47	0,48	0,16	0,22	-	0,41	0,27
A_6	0,42	0,44	0,28	0,33	0,41	-	0,33
A_7	0,22	0,28	0,34	0,32	0,27	0,33	-

Tabelle 15: Cohens Kappa – Testlauf 1

5.2.2 *Testlauf für die Annotation der Meinungsklasse*

Der zweite Testlauf sollte zeigen, ob eine direkte Klassifikation der Adjektive nach Meinungsklasse bessere Ergebnisse erzielen würde. Dazu sollte wiederum die Übereinstimmung der Annota-

toren gemessen werden. Die Testteilnehmer bekamen für diesen Test die Aufgabe, allen Adjektiven aus dem vorliegenden Set eine passende Meinungsklasse, wie weiter unten beschrieben, zuzuordnen. Die Teilnehmer sollten dabei darauf achten, dass jedes Adjektiv die zugeordnete Meinung für sich allein stehend wiedergibt und die Meinungsklasse gewählt wird, die das jeweilige Adjektiv in den meisten Fällen trägt. Teilnehmer des Tests waren dieselben sieben Personen, wie bereits bei der Annotation der Relevanz.

Die Adjektive sollten in folgende Meinungsklassen eingeordnet werden:

- 5 - stark positiv
- 4 - schwach positiv
- 3 - neutral
- 2 - schwach negativ
- 1 - stark negativ
- 0 - objektiv (nicht meinungstragend)

Die Analyse der Ergebnisse zeigte diesmal eine bessere Übereinstimmung der verschiedenen Testpersonen. Auch die Sichtung der annotierten Adjektive, vor allem der stark wertenden (nach Berechnung des Meinungswerts), war positiv. Tabelle 16 zeigt die Übereinstimmung der sieben Teilnehmer (A_n) im Paarvergleich. Als Maß für die Übereinstimmung wurde das ungewichtete Cohens Kappa verwendet. Das Maß bewegt sich relativ stabil um $\kappa = 0,4$ mit leichten Abweichungen nach unten und oben.

Untersucht wurde außerdem die Interrater-Reliabilität unter der Annahme, das auch benachbarte Meinungsklassen bei der Berechnung zugelassen, bzw. nicht bestraft, werden, mit Ausnahme der Klasse 0, der kein direkter Nachbar zugeordnet wird.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7
A_1	-	0,39	0,48	0,44	0,46	0,42	0,45
A_2	0,39	-	0,47	0,37	0,30	0,37	0,47
A_3	0,48	0,47	-	0,37	0,36	0,43	0,44
A_4	0,44	0,37	0,37	-	0,45	0,36	0,40
A_5	0,46	0,30	0,36	0,45	-	0,35	0,31
A_6	0,42	0,37	0,43	0,36	0,35	-	0,40
A_7	0,45	0,47	0,44	0,40	0,31	0,40	-

Tabelle 16: Cohens Kappa (ungewichtet) – Testlauf 2

Tabelle 17 zeigt die Ergebnisse des gewichteten Cohens Kappa im paarweisen Vergleich. Durch diese sinnvolle Gewichtung ergibt sich ein neues Bild. Der Wert für Cohens Kappa liegt hier fast durchgehend über $0,6$. Die Interpretation dieses Ergebnisses nach Landis und Koch ([28]) ergibt eine mittelmäßige bis beachtliche Übereinstimmung.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7
A_1	-	0,67	0,70	0,60	0,57	0,71	0,69
A_1	0,67	-	0,66	0,73	0,55	0,62	0,70
A_1	0,70	0,66	-	0,60	0,46	0,66	0,60
A_1	0,60	0,73	0,60	-	0,53	0,64	0,64
A_1	0,57	0,55	0,46	0,53	-	0,55	0,46
A_1	0,71	0,62	0,66	0,64	0,55	-	0,72
A_1	0,69	0,70	0,60	0,64	0,46	0,72	-

Tabelle 17: Cohens Kappa (gewichtet, benachbarte Klassen) – Testlauf 2

Durch den Testlauf wurde außerdem deutlich, dass es allen sieben Teilnehmern schwer fiel, zwischen „neutralen“ und „objektiven“ Adjektiven zu unterscheiden. Wird diese Erkenntnis nun in eine weitere Anpassung der Gewichtung des Maßes für die Interrater-Reliabilität umgesetzt, d.h. die Meinungsklassen „3 – neutral“ und „0 – objektiv“ werden gleichgesetzt, ergibt sich fol-

gendes Bild (Tabelle 18). Die Übereinstimmung der Testpersonen wird durch die neue Gewichtung nochmals besser. Cohens Kappa liegt nun im Durchschnitt bei einem Wert von 0,7.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7
A_1	-	0,75	0,74	0,76	0,70	0,86	0,74
A_2	0,75	-	0,68	0,79	0,55	0,65	0,81
A_3	0,74	0,68	-	0,66	0,44	0,69	0,68
A_4	0,76	0,79	0,66	-	0,71	0,73	0,78
A_5	0,70	0,55	0,44	0,71	-	0,62	0,55
A_6	0,86	0,65	0,69	0,73	0,62	-	0,77
A_7	0,74	0,81	0,68	0,78	0,55	0,77	-

Tabelle 18: Cohens Kappa (gewichtet, benachbarte Klassen, neutral = objektiv) – Testlauf 2

Für eine bessere Planung des Experiments wurde nach Beendigung des Testlaufs von jedem Teilnehmer die Dauer für die Klassifikation des Testsets abgefragt. Die mittlere Dauer für die Annotation von 169 Adjektiven betrug dabei 30 Minuten. Dies entspricht einer durchschnittlichen Zeitspanne für die Klassifikation eines Adjektivs von $\approx$ 11s.

5.2.3 *Fazit der durchgeführten Tests*

Durch die im Vorfeld des eigentlichen Experiments durchgeführten Vortests konnten Schwächen beim Versuchsaufbau identifiziert und entsprechend gegengesteuert werden.
Zum einen wurde deutlich, dass die Klassifikation der Adjektive nach Relevanz nicht präzise genug funktionierte und dieser Schritt daher im Folgenden verworfen werden musste. Aus diesem Grund musste auch die, in Kapitel 5.1.3 geplante, Aufnahme von Phrasen in das Lexikon vorerst verworfen werden, da der Annotationsaufwand ohne vorherige Entfernung nicht relevanter

Adjektive zu hoch erschien.
Zum anderen konnte die Annotationsanleitung durch die durchgeführten Tests sukzessive verbessert und wichtige Erkenntnisse hinsichtlich Dauer der Annotation und Fehlerquellen gewonnen werden.

5.3 Versuchsaufbau und -durchführung

Für das geplante Experiment wurde ein eigener *Scraper* entwickelt[7], der die benötigten Daten, d.h. die häufigsten deutschen Adjektive sowie zusätzliche Informationen, der in Abschnitt 5.1 benannten Quelle *Duden online* automatisch extrahierte.

Die komplette Liste der extrahierten Adjektive enthielt 23.757 Einträge[8]. Für eine manuelle Annotation war diese große Anzahl von Adjektiven jedoch ungeeignet. Ein Filterschritt, durch den lediglich die Adjektive aus den obersten drei Häufigkeitsklassen des Duden-Korpus, wie in Kapitel 5.1.3 beschrieben, ausgewählt wurden, ergab schließlich eine handhabbare Menge von 1.749 deutschen Adjektiven. Diese wurden ein weiteres Mal gefiltert, um fehlerhafte Einträge (50 großgeschriebene Begriffe) auszuschließen. Am Ende dieser beiden Filterschritte, entstand ein Set mit 1.699 deutschen Adjektiven.

Für die Durchführung des Experiments zur Generierung eines Referenzlexikons wurden anschließend folgende Anforderungen festgelegt:

- Aufteilung der 1.699 Adjektive in 4 Sets (A, B, C, D) mit je 425 (und einmal 424) Adjektiven
- min. 16 / max. 20 Teilnehmer in 4 Gruppen (G_1, G_2, G_3, G_4) zu je 4 / 5 Personen

7 Dieses Tool kam bereits für die Extraktion der Adjektive für den Vortest (Kapitel 5.2) zum Einsatz.
8 Aufgrund der Datenlage wurden zunächst 40.000 Adjektive geschätzt.

- 4 Slots (S_1, S_2, S_3, S_4) für die serielle Bearbeitung der 4 Sets mit ausreichend kalkulierter Dauer (siehe Vorabtest)
- kalkulierte Zeit pro Entscheidung, d.h. Klassifikation eines Adjektivs, ca. 12s - 15s (auf Basis der Vortests plus Puffer)
- Klassifikation der Adjektive in folgende Meinungsklassen: 5 – *stark positiv*, 4 – *schwach positiv*, 3 – *neutral*, 2 – *schwach negativ*, 1 – *stark negativ* und 0 – *objektiv*

Auf Basis der Vortests (siehe Kapitel 5.2) wurde ein Ablaufplan skizziert, der sowohl die Ergebnisse dieser Tests als auch die räumlichen und zeitlichen Rahmenbedingen berücksichtigte. Tabelle 19 gibt einen Überblick über den geplanten Ablauf und stellt die Slots (S_n) dar, während der die Annotatoren, unterteilt in Gruppen (G_n), die verschiedenen Sets (A, B, C und D) bearbeiten sollten.

	G_1	G_2	G_3	G_4	Uhrzeit	Dauer (min)	Zeit pro ADJ (s)	Pause (min)
S_1	A	B	C	D	09:00 - 10:30	90	12,7	15
S_2	B	A	D	C	10:45 - 12:15	90	12,7	60
S_3	C	D	A	B	13:15 - 15:00	105	14,8	30
S_4	D	C	B	A	15:30 - 17:15	105	14,8	-

Tabelle 19: Geplanter Ablauf des Experiments zur Erzeugung eines Referenzlexikons

Für das Experiment wurden 20 Studierende als wissenschaftliche Hilfskräfte angeworben und angestellt. Dabei wurde darauf geachtet, nur Personen als Kandidaten einzubeziehen, die keine Verbindung zur Forschungsgruppe „Analytische Informationssysteme“ hatten. Statistiken über die Zusammensetzung dieser Gruppe, d.h. Geschlecht, Studiengang (an der Hochschule Hof) und Semester, finden sich, ebenso wie die Übersicht des tatsächlichen Ablaufs, in Abschnitt 5.4.1.

5.4 Ergebnisse

Im folgenden Kapitel werden die Ergebnisse des Experiments zur Erzeugung eines Referenzlexikons für die Evaluation deutschsprachiger lexikalischer Ressourcen vorgestellt. Zunächst werden in Kapitel 5.4.1 Statistiken über die Teilnehmer des Experiments sowie zum Ablauf präsentiert. Anschließend wird in Kapitel 5.4.2 die Plausibilität der erhaltenen Ergebnisse diskutiert. In Kapitel 5.4.3 wird ein Teil dieser Ergebnisse schließlich in Tabellenform vorgestellt.

5.4.1 *Statistische Auswertungen*

Dieser Abschnitt enthält grundlegende Statistiken des durchgeführten Experiments sowie eine Übersicht des tatsächlichen Ablaufs.

Tabelle 20 gibt einen Überblick über die Teilnehmer des Experiments. Die Tabelle enthält eine laufende Nummer, die zugeordnete Gruppe, die Bezeichnung (auf den Klarnamen wurde aus Gründen des Datenschutzes an dieser Stelle verzichtet), den Studiengang, das Semester und das Geschlecht der Teilnehmer.

Nr.	Gruppe	Bezeichnung	Studiengang[9]	Semester	Geschlecht
1	A	Tagger 1	Inf	4	m
2	B	Tagger 2	Inf	4	m
3	C	Tagger 3	MC	4	m
4	D	Tagger 4	MC	4	m
5	A	Tagger 5	MC	6	m
6	B	Tagger 6	Inf	4	m
7	C	Tagger 7	WI	6	m
8	D	Tagger 8	WI	6	m
9	A	Tagger 9	MC	4	m
10	B	Tagger 10	MC	4	m
11	C	Tagger 11	MC	4	m
12	D	Tagger 12	MC	4	m
13	A	Tagger 13	MC	4	m
14	B	Tagger 14	MC	4	m
15	C	Tagger 15	MC	4	m
16	D	Tagger 16	MI	4	w
17	A	Tagger 17	MI	4	w
18	B	Tagger 18	MI	4	w
19	C	Tagger 19	MI	4	w
20	D	Tagger 20	MI	4	w

Tabelle 20: Teilnehmer des Experiments zur Erzeugung eines Referenzlexikons

Tabelle 21 illustriert den tatsächlichen Ablauf des Experiments (im Vergleich dazu siehe den geplanten Ablauf in Tabelle 19). Es zeigte sich, dass die Teilnehmer des Experiments die gestellte Klassifikationsaufgabe deutlich schneller als erwartet erledigten. Die Zeiten variieren dabei in und zwischen den vier Slots[10]. Die schnellsten Teilnehmer benötigten für die Klassifikation in S_1 25 min, in S_2 28 min, in S_3 25 min und in S_4 23 min. Der Gesamtdurchschnitt für die Klassifikation von 425 Adjektiven lag dabei bei nur 42 min, deutlich schneller als durch die Vortests (siehe Kapitel 5.2) erwartet.

9 Studiengänge der Hochschule Hof im Sommersemster 2017: Inf – Informatik, MC – Mobile Computing, WI – Wirtschaftsinformatik, und MI – Medieninformatik.

10 Uhrzeiten und Dauer in Tabelle 21 beziehen sich jeweils auf jene Teilnehmer des Experiments, die zuletzt fertig wurden (maximale Dauer).

	G_1	G_2	G_3	G_4	Uhrzeit	Dauer (min)	Zeit pro ADJ (s)	Pause (min)
S_1	A	B	C	D	09:40 - 10:49	69	9,7	1
S_2	B	A	D	C	10:50 - 11:47	57	8,1	103
S_3	C	D	A	B	13:30 - 14:24	54	7,6	16
S_4	D	C	B	A	14:40 - 15:22	42	5,9	-

Tabelle 21: Realer Ablauf des Experiments zur Erzeugung eines Referenzlexikons

5.4.2 *Evaluation der Ergebnisse*

Bei der Auswertung der Ergebnisse wurde zunächst die Interrater-Reliabilität der 20 Teilnehmer bestimmt. Beim paarweisen Vergleich durch Cohens Kappa wurde zunächst der Datensatz von „Tagger 5“ ausgeschlossen, da dieser mit einem Kappa von durchschnittlich $\kappa = 0,28$ weit hinter den Ergebnissen der anderen Teilnehmer zurücklag. Anschließend wurde die Übereinstimmung der verbleibenden 19 Teilnehmer gemessen. Das gewichtete durchschnittliche Kappa zeigte dabei mit $\kappa \approx 0,53$ eine mittelmäßige Übereinstimmung der Rater. Die Überprüfung der endgültigen Liste mit Adjektiven und dazugehörigen Sentiment Values erschien plausibel. Vor allem die als wertend klassifizierten Adjektive mit einem $|SV| > 0,33$ erscheinen dabei vielversprechend. Durch einen weiteren Evaluationsschritt, bei dem einem zufällig ausgewählten Teil der 1.699 Adjektive manuell Ränge zugeordnet wurden, konnten die Ergebnisse unabhängig verifiziert werden (Kapitel 6.1).

5.4.3 *Übersicht meinungstragender Adjektive*

Die Tabellen 22 - 28 enthalten alle, von den Teilnehmern des Experiments, subjektiv klassifizierten Adjektive, inklusive der berechneten *Sentiment Values* (SV).

Adjektiv	*SV*	Adjektiv	*SV*	Adjektiv	*SV*
großartig	1,00	kompetent	0,82	fit	0,74
wunderschön	0,97	schön	0,82	freundlich	0,74
legendär	0,95	selbstbewusst	0,82	gemeinnützig	0,74
perfekt	0,95	vorbildlich	0,82	göttlich	0,74
top	0,95	clever	0,79	gut	0,74
erstklassig	0,92	gerecht	0,79	leistungsfähig	0,74
exzellent	0,92	gesund	0,79	mutig	0,74
genial	0,92	happy	0,79	positiv	0,74
grandios	0,92	herrlich	0,79	prächtig	0,74
hervorragend	0,92	herzlich	0,79	prima	0,74
attraktiv	0,89	hochwertig	0,79	produktiv	0,74
ausgezeichnet	0,89	professionell	0,79	richtig	0,74
beeindruckend	0,89	pünktlich	0,79	sexy	0,74
erfolgreich	0,89	beliebt	0,76	spitze	0,74
super	0,89	bemerkenswert	0,76	talentiert	0,74
wunderbar	0,89	effizient	0,76	vernünftig	0,74
glücklich	0,87	fair	0,76	zukunftsweisend	0,74
herausragend	0,87	köstlich	0,76	charmant	0,71
hübsch	0,87	optimistisch	0,76	erstaunlich	0,71
liebevoll	0,87	schlau	0,76	feierlich	0,71
brillant	0,84	spektakulär	0,76	froh	0,71
fleißig	0,84	stark	0,76	glaubwürdig	0,71
gebildet	0,84	tapfer	0,76	hilfreich	0,71
himmlisch	0,84	toll	0,76	innovativ	0,71
ideal	0,84	treu	0,76	lebendig	0,71
klug	0,84	überdurchschnittlich	0,76	leidenschaftlich	0,71
optimal	0,84	verlässlich	0,76	qualifiziert	0,71
sensationell	0,84	visionär	0,76	siegreich	0,71
sympathisch	0,84	wertvoll	0,76	umweltfreundlich	0,71
überragend	0,84	zuverlässig	0,76	akademisch	0,68
atemberaubend	0,82	aussichtsreich	0,74	aufmerksam	0,68
ehrlich	0,82	authentisch	0,74	detailliert	0,68
eindrucksvoll	0,82	bedeutsam	0,74	einflussreich	0,68
einwandfrei	0,82	begehrt	0,74	exakt	0,68
einzigartig	0,82	begeistert	0,74	harmonisch	0,68
faszinierend	0,82	edel	0,74	intellektuell	0,68
fröhlich	0,82	ehrgeizig	0,74	korrekt	0,68
geil	0,82	elegant	0,74	kostenlos	0,68
intelligent	0,82	erfreulich	0,74	kraftvoll	0,68

Tabelle 22: Übersicht subjektiver Adjektive des erzeugten Referenzlexikons (1)

Adjektiv	*SV*	Adjektiv	*SV*	Adjektiv	*SV*
lieb	0,68	angesehen	0,63	strahlend	0,61
motiviert	0,68	aufrecht	0,63	überzeugt	0,61
nachhaltig	0,68	außergewöhnlich	0,63	vertraut	0,61
nett	0,68	besser	0,63	vielseitig	0,61
nobel	0,68	blühend	0,63	vollkommen	0,61
präzise	0,68	cool	0,63	wettbewerbsfähig	0,61
qualitativ	0,68	ehrenamtlich	0,63	wichtig	0,61
raffiniert	0,68	friedlich	0,63	ansprechend	0,58
reibungslos	0,68	frisch	0,63	beruhigend	0,58
smart	0,68	gelungen	0,63	eifrig	0,58
spannend	0,68	gemütlich	0,63	eindeutig	0,58
süß	0,68	gigantisch	0,63	entspannt	0,58
überwältigend	0,68	großzügig	0,63	euphorisch	0,58
überzeugend	0,68	höflich	0,63	fähig	0,58
wahr	0,68	idyllisch	0,63	geduldig	0,58
zufrieden	0,68	lustig	0,63	gemeinsam	0,58
angenehm	0,66	sicher	0,63	gepflegt	0,58
begabt	0,66	sinnvoll	0,63	komfortabel	0,58
engagiert	0,66	steuerfrei	0,63	kostengünstig	0,58
erfahren	0,66	strukturiert	0,63	logisch	0,58
familiär	0,66	weise	0,63	lukrativ	0,58
fortschrittlich	0,66	bedeutend	0,61	menschlich	0,58
geschickt	0,66	dankbar	0,61	mühelos	0,58
gleichberechtigt	0,66	dynamisch	0,61	ordentlich	0,58
gründlich	0,66	flexibel	0,61	profitabel	0,58
kreativ	0,66	freiheitlich	0,61	renommiert	0,58
maßgeschneidert	0,66	gesichert	0,61	robust	0,58
nützlich	0,66	interessant	0,61	sauber	0,58
preiswert	0,66	interessiert	0,61	sozial	0,58
reizvoll	0,66	klar	0,61	übereinstimmend	0,58
romantisch	0,66	kostenfrei	0,61	übersichtlich	0,58
sorgfältig	0,66	lebhaft	0,61	unkompliziert	0,58
stolz	0,66	modern	0,61	unterhaltsam	0,58
wirksam	0,66	revolutionär	0,61	verständlich	0,58
wirkungsvoll	0,66	schick	0,61	vielversprechend	0,58
witzig	0,66	selbstständig	0,61	würdig	0,58
abwechslungsreich	0,63	seriös	0,61	zuversichtlich	0,58
aktiv	0,63	sonnig	0,61	amüsant	0,55
ambitioniert	0,63	sportlich	0,61	anerkannt	0,55

Tabelle 23: Übersicht subjektiver Adjektive des erzeugten Referenzlexikons (2)

Adjektiv	*SV*	Adjektiv	*SV*	Adjektiv	*SV*
anständig	0,55	glänzend	0,53	vornehm	0,50
behutsam	0,55	glaubhaft	0,53	anschaulich	0,47
berühmt	0,55	human	0,53	begründet	0,47
bewährt	0,55	humanitär	0,53	berechtigt	0,47
echt	0,55	imposant	0,53	bereit	0,47
eigenständig	0,55	komplett	0,53	bewegend	0,47
erotisch	0,55	kostbar	0,53	fundiert	0,47
erreichbar	0,55	malerisch	0,53	gemeinschaftlich	0,47
erwachsen	0,55	namhaft	0,53	geordnet	0,47
fein	0,55	natürlich	0,53	gewandt	0,47
grenzenlos	0,55	ordnungsgemäß	0,53	heimisch	0,47
heiter	0,55	originell	0,53	intakt	0,47
initiativ	0,55	preisgünstig	0,53	konsequent	0,47
klasse	0,55	vorsorglich	0,53	lässig	0,47
konkret	0,55	zutreffend	0,53	magisch	0,47
konstruktiv	0,55	ästhetisch	0,50	nahtlos	0,47
konzentriert	0,55	ausführlich	0,50	neuartig	0,47
kräftig	0,55	außerordentlich	0,50	olympisch	0,47
legal	0,55	beispielhaft	0,50	original	0,47
mächtig	0,55	einstimmig	0,50	passend	0,47
neu	0,55	fertig	0,50	prominent	0,47
nutzbar	0,55	festlich	0,50	realistisch	0,47
offen	0,55	genau	0,50	überlegen	0,47
problemlos	0,55	günstig	0,50	universal	0,47
reichlich	0,55	hochkarätig	0,50	verblüffend	0,47
schlüssig	0,55	individuell	0,50	verfügbar	0,47
sehenswert	0,55	luxuriös	0,50	warm	0,47
souverän	0,55	nachvollziehbar	0,50	zugelassen	0,47
unermüdlich	0,55	praktisch	0,50	aufregend	0,45
vollständig	0,55	rein	0,50	ausgewogen	0,45
willkommen	0,55	repräsentativ	0,50	elementar	0,45
beachtlich	0,53	schlank	0,50	exklusiv	0,45
brav	0,53	selbstverständlich	0,50	fachlich	0,45
einverstanden	0,53	signifikant	0,50	flott	0,45
etabliert	0,53	stabil	0,50	fortgeschritten	0,45
frei	0,53	studiert	0,50	freiwillig	0,45
fresh	0,53	unabhängig	0,50	ganz	0,45
geeignet	0,53	verdient	0,50	geregelt	0,45
gehoben	0,53	vielfältig	0,50	gewählt	0,45

Tabelle 24: Übersicht subjektiver Adjektive des erzeugten Referenzlexikons (3)

Adjektiv	*SV*	Adjektiv	*SV*	Adjektiv	*SV*
gewiss	0,45	mehrheitlich	0,42	erregend	0,37
heilig	0,45	modisch	0,42	ersichtlich	0,37
hundertprozentig	0,45	pädagogisch	0,42	förmlich	0,37
künstlerisch	0,45	reich	0,42	fromm	0,37
kunstvoll	0,45	relevant	0,42	gleichmäßig	0,37
monumental	0,45	sinnlich	0,42	groß	0,37
munter	0,45	solidarisch	0,42	gültig	0,37
musikalisch	0,45	stattlich	0,42	handwerklich	0,37
nennenswert	0,45	transparent	0,42	hell	0,37
real	0,45	zahlreich	0,42	konstant	0,37
rechtmäßig	0,45	zulässig	0,42	kontrolliert	0,37
sanft	0,45	adäquat	0,39	legitim	0,37
schillernd	0,45	aktuell	0,39	persönlich	0,37
strategisch	0,45	angestrebt	0,39	populär	0,37
traditionsreich	0,45	beständig	0,39	rhythmisch	0,37
treffend	0,45	beweglich	0,39	ruhig	0,37
unbegrenzt	0,45	bewusst	0,39	sachlich	0,37
wissenschaftlich	0,45	bunt	0,39	simpel	0,37
wohlhabend	0,45	direkt	0,39	sparsam	0,37
zart	0,45	einig	0,39	spielerisch	0,37
ausgebildet	0,42	erlaubt	0,39	spontan	0,37
ausgefallen	0,42	expert	0,39	traut	0,37
bescheiden	0,42	final	0,39	umfassend	0,37
demokratisch	0,42	golden	0,39	unbestritten	0,37
erneuerbar	0,42	handfest	0,39	uneingeschränkt	0,37
erwartungsgemäß	0,42	harmlos	0,39	vorstellbar	0,37
folgerichtig	0,42	international	0,39	zukünftig	0,37
führend	0,42	locker	0,39	angemessen	0,34
gelernt	0,42	maximal	0,39	ausgeglichen	0,34
geschäftsführend	0,42	offenkundig	0,39	beispiellos	0,34
gewünscht	0,42	reif	0,39	dazugehörig	0,34
greifbar	0,42	satt	0,39	digital	0,34
heil	0,42	strukturell	0,39	diplomatisch	0,34
kompakt	0,42	überschaubar	0,39	einheitlich	0,34
kompatibel	0,42	ungebrochen	0,39	entschieden	0,34
königlich	0,42	verheiratet	0,39	geschätzt	0,34
leicht	0,42	bestimmt	0,37	hochrangig	0,34
leuchtend	0,42	brauchbar	0,37	informell	0,34
machbar	0,42	einfach	0,37	jugendlich	0,34

Tabelle 25: Übersicht subjektiver Adjektive des erzeugten Referenzlexikons (4)

Adjektiv	*SV*	Adjektiv	*SV*	Adjektiv	*SV*
markant	0,34	fremd	-0,37	rückläufig	-0,42
maßgeblich	0,34	gering	-0,37	starr	-0,42
mobil	0,34	kommerziell	-0,37	steuerpflichtig	-0,42
moralisch	0,34	lautstark	-0,37	stürmisch	-0,42
offiziell	0,34	militant	-0,37	träge	-0,42
ökonomisch	0,34	nass	-0,37	ungleich	-0,42
plausibel	0,34	skurril	-0,37	winzig	-0,42
poetisch	0,34	steuerlich	-0,37	befristet	-0,45
potenzial	0,34	unbestimmt	-0,37	geladen	-0,45
programmatisch	0,34	verlegen	-0,37	gleichgültig	-0,45
riesig	0,34	verrückt	-0,37	grotesk	-0,45
spezifisch	0,34	vorsätzlich	-0,37	heimlich	-0,45
spielend	0,34	beklagt	-0,39	komisch	-0,45
stetig	0,34	billig	-0,39	link	-0,45
systematisch	0,34	flüchtig	-0,39	notorisch	-0,45
tropisch	0,34	gebraucht	-0,39	oberflächlich	-0,45
ultimativ	0,34	gentechnisch	-0,39	populistisch	-0,45
unternehmerisch	0,34	knapp	-0,39	spät	-0,45
vereinbar	0,34	mäßig	-0,39	spekulativ	-0,45
wirtschaftlich	0,34	platt	-0,39	überhöht	-0,45
wünschenswert	0,34	prekär	-0,39	ungewohnt	-0,45
ausverkauft	-0,34	psychiatrisch	-0,39	unklar	-0,45
betroffen	-0,34	rigoros	-0,39	unwahrscheinlich	-0,45
entfernt	-0,34	schal	-0,39	vage	-0,45
hart	-0,34	schwer	-0,39	altmodisch	-0,47
ideologisch	-0,34	sturm	-0,39	anstrengend	-0,47
lau	-0,34	undenkbar	-0,39	bedenklich	-0,47
laut	-0,34	abrupt	-0,42	diffus	-0,47
massenhaft	-0,34	abweichend	-0,42	finster	-0,47
rasend	-0,34	allein	-0,42	fraglich	-0,47
roh	-0,34	alt	-0,42	grob	-0,47
streng	-0,34	besorgt	-0,42	kommunistisch	-0,47
unerwartet	-0,34	blass	-0,42	müde	-0,47
zwangsläufig	-0,34	bürokratisch	-0,42	paradox	-0,47
albern	-0,37	dramatisch	-0,42	schwerwiegend	-0,47
banal	-0,37	kritisch	-0,42	seltsam	-0,47
brisant	-0,37	melancholisch	-0,42	skeptisch	-0,47
chronisch	-0,37	militärisch	-0,42	spärlich	-0,47
forsch	-0,37	niedrig	-0,42	ungewiss	-0,47

Tabelle 26: Übersicht subjektiver Adjektive des erzeugten Referenzlexikons (5)

Adjektiv	*SV*	Adjektiv	*SV*	Adjektiv	*SV*
ungläubig	-0,47	gegnerisch	-0,55	drastisch	-0,61
unrealistisch	-0,47	heikel	-0,55	erschreckend	-0,61
zögerlich	-0,47	kahl	-0,55	gnadenlos	-0,61
begrenzt	-0,50	kostspielig	-0,55	problematisch	-0,61
dürr	-0,50	merkwürdig	-0,55	ratlos	-0,61
kompliziert	-0,50	mörderisch	-0,55	schwierig	-0,61
langwierig	-0,50	schade	-0,55	trotzig	-0,61
mühsam	-0,50	scheu	-0,55	ungeduldig	-0,61
nationalistisch	-0,50	tabu	-0,55	unnötig	-0,61
rau	-0,50	umstritten	-0,55	unzulässig	-0,61
restriktiv	-0,50	unbequem	-0,55	verdächtig	-0,61
schief	-0,50	unerlaubt	-0,55	vermisst	-0,61
schräg	-0,50	ungeklärt	-0,55	verwirrend	-0,61
turbulent	-0,50	unsicher	-0,55	willkürlich	-0,61
wahnsinnig	-0,50	verschlafen	-0,55	zerfallen	-0,61
widersprüchlich	-0,50	verschlossen	-0,55	bedroht	-0,63
zynisch	-0,50	zweifelhaft	-0,55	beschränkt	-0,63
dick	-0,53	zwingend	-0,55	besessen	-0,63
dubios	-0,53	angeschlagen	-0,58	blind	-0,63
eigenartig	-0,53	ärgerlich	-0,58	chaotisch	-0,63
falsch	-0,53	einseitig	-0,58	einsam	-0,63
fehlend	-0,53	feind	-0,58	eiskalt	-0,63
islamistisch	-0,53	karg	-0,58	erschlagen	-0,63
kalt	-0,53	mager	-0,58	hektisch	-0,63
langsam	-0,53	misstrauisch	-0,58	lahm	-0,63
marode	-0,53	naiv	-0,58	langweilig	-0,63
stumpf	-0,53	nervös	-0,58	leid	-0,63
teuer	-0,53	nieder	-0,58	schleppend	-0,63
überfällig	-0,53	riskant	-0,58	traurig	-0,63
unangemessen	-0,53	strittig	-0,58	unheimlich	-0,63
unberechenbar	-0,53	trist	-0,58	unübersichtlich	-0,63
uralt	-0,53	ungünstig	-0,58	unverständlich	-0,63
wund	-0,53	unruhig	-0,58	verlassen	-0,63
ablehnend	-0,55	unterlegen	-0,58	abhängig	-0,66
angespannt	-0,55	unwirksam	-0,58	absurd	-0,66
betrunken	-0,55	vergeblich	-0,58	belastend	-0,66
bitter	-0,55	abgelaufen	-0,61	erfolglos	-0,66
foul	-0,55	ängstlich	-0,61	fett	-0,66
fragwürdig	-0,55	behindert	-0,61	furchtbar	-0,66

Tabelle 27: Übersicht subjektiver Adjektive des erzeugten Referenzlexikons (6)

Adjektiv	*SV*	Adjektiv	*SV*	Adjektiv	*SV*
gefährdet	-0,66	strafbar	-0,74	tragisch	-0,84
gemein	-0,66	unbefriedigend	-0,74	ungerecht	-0,84
irr	-0,66	unmöglich	-0,74	wütend	-0,84
schmerzlich	-0,66	verletzt	-0,74	aggressiv	-0,87
schmutzig	-0,66	antisemitisch	-0,76	arrogant	-0,87
übel	-0,66	arbeitslos	-0,76	feindlich	-0,87
umständlich	-0,66	drohend	-0,76	giftig	-0,87
unterirdisch	-0,66	fehlerhaft	-0,76	depressiv	-0,89
defekt	-0,68	hilflos	-0,76	gefälscht	-0,89
defekt	-0,68	rechtswidrig	-0,76	gewaltsam	-0,89
eisig	-0,68	sinnlos	-0,76	gewalttätig	-0,89
erbittert	-0,68	tot	-0,76	grausam	-0,89
gefährlich	-0,68	blutig	-0,79	schlecht	-0,89
kaputt	-0,68	gefangen	-0,79	brutal	-0,92
nuklear	-0,68	hoffnungslos	-0,79	dumm	-0,92
obdachlos	-0,68	mangelhaft	-0,79	fürchterlich	-0,92
schuldig	-0,68	pessimistisch	-0,79	korrupt	-0,92
schwach	-0,68	radikal	-0,79	kriminell	-0,92
überflüssig	-0,68	radioaktiv	-0,79	lebensgefährlich	-0,92
unangenehm	-0,68	unerwünscht	-0,79	tödlich	-0,92
unfreiwillig	-0,68	ungeliebt	-0,79	unerträglich	-0,92
unrecht	-0,68	ungenügend	-0,79	bedrohlich	-0,95
unsinnig	-0,68	verboten	-0,79	böse	-0,95
unzureichend	-0,68	bewaffnet	-0,82	elend	-0,95
verfassungswidrig	-0,68	fahrlässig	-0,82	hässlich	-0,95
wüst	-0,68	fatal	-0,82	krank	-0,95
bizarr	-0,71	schlimm	-0,82	rechtsextrem	-0,95
enttäuscht	-0,71	unglücklich	-0,82	miserabel	-0,97
lächerlich	-0,71	unzufrieden	-0,82	rassistisch	-0,97
peinlich	-0,71	verheerend	-0,82	rechtsradikal	-1,00
stur	-0,71	verzweifelt	-0,82	terroristisch	-1,00
ungeheuer	-0,71	enttäuschend	-0,84		
verfallen	-0,71	illegal	-0,84		
verloren	-0,71	katastrophal	-0,84		
faul	-0,74	nationalsozialistisch	-0,84		
gebrochen	-0,74	negativ	-0,84		
lästig	-0,74	schädlich	-0,84		
mies	-0,74	schmerzhaft	-0,84		
sauer	-0,74	schrecklich	-0,84		

Tabelle 28: Übersicht subjektiver Adjektive des erzeugten Referenzlexikons (7)

5.5 Zusammenfassung

Die Erzeugung des beschriebenen Referenzlexikons mit 1.699 deutschen Adjektiven verlief problemlos. Nach Auswertung der Datensets der 20 Teilnehmer und Überprüfung der Interrater-Reliabilität, musste lediglich der Datensatz eines Teilnehmers („Tagger 5“, siehe Kapitel 5.4.2), aufgrund einer zu geringen Übereinstimmung, von der Berechnung der Meinungswerte ausgeschlossen werden.[11]

Nach Berechnung der Meinungswerte (siehe Formel in Kapitel 4.3.3.3), zeigte sich folgende Verteilung:

- 133 stark positive Adjektive (Intervall $[+1; +\frac{2}{3}]$)
- 356 schwach positive Adjektive (Intervall: $]+\frac{2}{3}; +\frac{1}{3}]$)
- 888 neutrale / objektive Adjektive (Intervall: $]+\frac{1}{3}; -\frac{1}{3}[$)
- 221 schwach negative Adjektive (Intervall: $[-\frac{1}{3}; -\frac{2}{3}[$)
- 101 stark negative Adjektive (Intervall: $[-\frac{2}{3}; -1]$)

Obwohl durch den Vortest in Kapitel 5.2.2 deutlich wurde, dass die Teilnehmer Probleme bei der Unterscheidung von „neutralen“ und „objektiven“ Adjektiven hatten, wurde diese Aufteilung vorerst beibehalten und beide Meinungsklassen getrennt erhoben. Die erhobenen Daten sollen zu einem späteren Zeitpunkt der genaueren Analyse dieser beiden Meinungsklassen dienen, um zu Überprüfen ob Adjektive existieren, die eindeutig „neutral“ bzw. „objektiv“ sind, bzw. wo die „Grenze“ zwischen beiden Klassen verläuft. Für die Ermittlung der Meinungswerte wurden beide Meinungsklassen jedoch zusammengefasst.

Durch das Experiment wurden von den häufigsten deutschen Adjektiven (1.699) laut Duden, 811 Adjektive (48, 0%) als subjektiv, d.h. meinungstragend, und 888 Adjektive (52, 0%) als objektiv

11 Die Ergebnisse änderten sich dadurch allerdings nur marginal.

bzw. neutral, d.h. nicht meinungstragend, klassifiziert. Durch die große Anzahl von Annotatoren sowie deren zufriedenstellender Übereinstimmung ($\kappa_{all} \approx 0,53$) ist dieses Ergebnis als valide und plausibel anzusehen.

Geplant als Referenz für die Evaluation automatisch generierter Lexika, kann das erzeugte Referenzlexikon zudem als Seed für die Generierung neuer lexikalischer Ressourcen (Wörterbuchbasierter Ansatz) verwendet werden. Zudem kommt eine Verwendung als eigenständiges, sehr präzises, Lexikon in Frage. Dazu sind allerdings weitere Schritte notwendig, wie die Erweiterung um wichtige Phrasen sowie weiterer Wortarten, wie etwa Nomen.

Das erzeugte Referenzlexikon für die deutsche Sprache wird nun genutzt (siehe Kapitel 6), um alle in Kapitel 2.3.2 beschriebenen deutschen Ressourcen zu evaluieren und insbesondere auch die Effektivität der, in dieser Arbeit beschriebenen, neu erzeugten Ressource zu messen. Zu beachten ist dabei die Aussagekraft der Evaluationsergebnisse. Da es sich bei der erzeugten Referenzressource um eine Liste bestehend aus einzelnen Adjektiven handelt, kann die Vollständigkeit (Recall) eines Lexikons nicht ermittelt werden.[12] Sowohl andere Wortarten, wie Verben oder Nomen, als auch Phrasen sind nicht in dieser Ressource vorhanden.[13] Allerdings kann das erzeugte Referenzlexikon als kleinster gemeinsamer Nenner aller deutschen Ressourcen angesehen werden, mit dessen Hilfe eine Aussage zum Vorhandensein und zur Genauigkeit, betreffend Polarität, und, falls vorhanden, Stärke wichtiger meinungstragender deutscher Adjektive getroffen werden kann.

12 Fairerweise muss hier erwähnt werden, dass es im Allgemeinen unmöglich erscheint, die Vollständigkeit einer lexikalischen Ressource zu ermitteln. Die relevanten Begriffe unterscheiden sich je nach Domäne und zu analysierendem Text.

13 Es ist allerdings geplant, diese Ressource sukzessive um Phrasen sowie um Nomen zu erweitern.

6 EVALUATION

Das nachfolgende Kapitel behandelt die Evaluation der erzeugten lexikalischen Ressource. Dazu wird in Abschnitt 6.1 zunächst eine qualitative (manuelle) Evaluation des zu verwendeten Referenzkorpus durchgeführt. Anschließend werden die in Kapitel 2.3.2 beschriebenen deutschsprachigen lexikalischen Ressourcen sowie die im Rahmen dieser Arbeit erzeugte Ressource in Kapitel 6.2 schließlich mit dem in Kapitel 5 erzeugten Referenzkorpus evaluiert.

6.1 Qualitative Evaluation des Referenzlexikons

In einem ersten Evaluationsschritt wurde die Plausibilität des erzeugten Referenzlexikons nochmals manuell überprüft. Zu diesem Zweck wurde ein Tool konzeptioniert und entwickelt, mit dessen Hilfe einer vorgegebenen Anzahl von Wörtern, durch paarweisen Vergleich, manuell Ränge zugeordnet werden konnten. Abbildung 14 zeigt die grafische Oberfläche des entwickelten Evaluationstools. Die grundlegende Idee bei der Bestimmung der Ränge war der direkte Vergleich zweier Wörter, bei welchem die Testteilnehmer immer den, aus ihrer Sicht, jeweils positiveren Begriff auswählen mussten. Ein Wort wurde dabei so oft mit anderen, bereits in der Liste vorhandenen, Wörtern verglichen, bis es an der richtigen Position eingefügt werden konnte. Durch die zugrunde liegende Datenstruktur entstand durch das Einfügen sukzessiv eine nach Positivität der Wörter absteigend sortierte Liste.

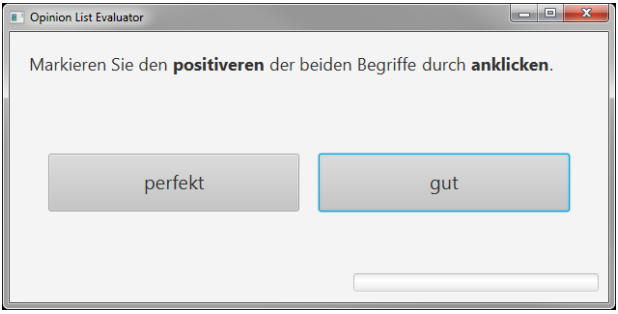


Abbildung 14: Evaluationstool

Für diesen Evaluationsschritt wurde zunächst ein Subset der im Referenzlexikon vorkommenden Adjektive ausgewählt. Um den Aufwand für die Annotatoren dabei im Rahmen zu halten, wurden, nach Abschätzung des benötigten Aufwand, 125 der 1.699 ($\approx 7,4\%$) Adjektive zufällig bestimmt und für die Verwendung mit dem Evaluationstool vorbereitet. Anschließend wurden zwei unabhängige Testteilnehmer bestimmt, die den Annotationssprozess jeweils für die ausgewählten 125 Adjektive durchführen sollten.

Zum Ende dieses Prozesses entstanden zwei Listen mit Adjektiven, jeweils absteigend geordnet nach der Positivität der enthaltenden Adjektive. Um die Übereinstimmung der beiden Testpersonen zu messen, wurde ein Rangtest nach Spearman durchgeführt, der mit $\rho = 0,87$ (p-value $< 2,2e{-}16$) eine sehr gute Übereinstimmung zeigte.

Für die weitere Evaluation wurden schließlich die Ränge der 125 zufällig ausgewählten Adjektive des Referenzlexikons durch Sortierung nach SV bestimmt und anschließend überprüft, wie groß die Ähnlichkeit diese Rangfolge denen der beiden manuell erzeugten Adjektiv-Listen entsprach. Dazu wurde wiederum Spearman's ρ verwendet. Die Bestimmung der Rangkorrelation zwischen den Ergebnissen von Annotator 1 und den Daten des

Referenzlexikons ergab einen Wert von $\rho = 0,89$ (p-value $< 2,2e{-}16$), der Vergleich der Ränge mit den Ergebnissen von Annotator 2 war mit $\rho = 0,87$ (p-value $< 2,2e{-}16$) nur unwesentlich schlechter.

Auf Basis dieser Ergebnisse können die Daten des Referenzlexikons als plausibel angesehen werden.

6.2 Evaluationen durch das erzeugte Referenzlexikon

In diesem Abschnitt werden alle vorhandenen deutschen Meinungslexika sowie die selbst erzeugte lexikalische Ressource mittels des erstellten Referenzlexikons evaluiert. Dabei wird zunächst überprüft, wie viele der häufigsten deutschen meinungstragenden Adjektive in den jeweiligen Listen enthalten sind. Anschließend wird die Güte der lexikalischen Ressourcen durch die Ermittlung des Korrelationskoeffizienten nach *Pearson* evaluiert. Dabei wird überprüft, ob die Verteilung der Meinungswerte der Adjektive der verschiedenen Lexika denen des Referenzlexikons entspricht.

Bei der Interpretation der Evaluationsergebnisse sind verschiedene Punkte zu beachten:

1. Betrachtet werden ausschließlich Einzelwörter. Listen, die neben einzelnen Wörtern auch Phrasen beinhalten, werden bei der Evaluation insofern „benachteiligt“, als dass ihr eigentliches Potenzial nicht gemessen werden kann.

2. Das Referenzlexikon enthält zu diesem Zeitpunkt ausschließlich häufige deutsche Adjektive.

3. Das Referenzlexikon ist (per Definition) nicht vollständig.

4. Die lexikalischen Ressourcen sind, durch verschiedene Wertebereiche, Art der Erzeugung sowie verschiedentlich ent-

haltene Informationen (teilweise POS-Tags und Flexionen), nur bedingt vergleichbar.

Es existieren also diverse Einschränkungen. Dennoch wird durch das erzeugte Referenzlexikon ein Instrument geboten, durch das eine erste Einschätzung der „Vollständigkeit“ sowie der „Güte“ der verschiedenen Lexika erfolgen kann. Zudem kann das Referenzlexikon zukünftig, auf Basis der in Kapitel 5.3 beschriebene Methode, um weitere Wortarten und Phrasen ergänzt werden.

6.2.1 *Übersicht Referenzlexikon*

Das erzeugte Referenzlexikon enthält 1.699 Adjektive. Davon wurden 811 (48,0%) als subjektiv, d.h. wertend, klassifiziert – 489 (60,3%) positive und 322 (39,7%) negative. Weiterhin wurden 234 (28,9%) dieser 811 Adjektive der Klasse *stark wertend* zugeordnet – 133 (56,8%) positive und 101 (43,2%) negative.

Abbildung 15 zeigt die Verteilung der wertenden Adjektive des Referenzlexikons. Dabei zeigte sich eine ähnliche Verteilung der Sentiment Values bei den beiden verschiedenen Gruppen (positiv und negativ).

In den folgenden Abschnitten werden zunächst die in Kapitel 2.3.2 beschriebenen lexikalischen Ressourcen und anschließend die in dieser Arbeit erzeugte Ressource durch das Referenzlexikon evaluiert.

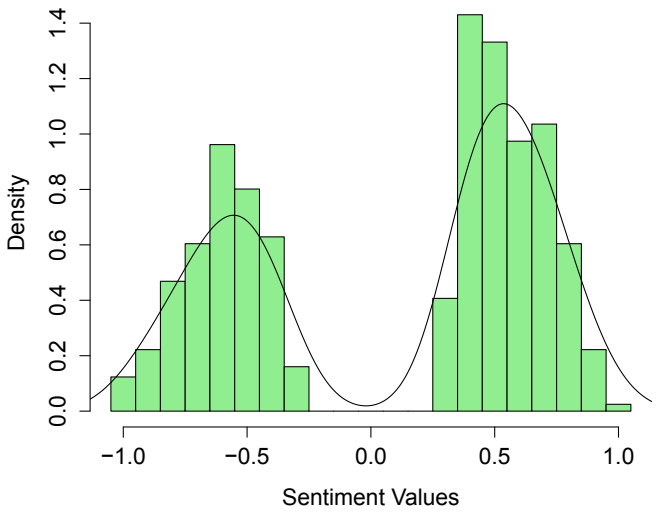


Abbildung 15: Referenzlexikon: Verteilung der meinungstragenden Wörter nach *Sentiment Values* (*SV*)

6.2.2 *SentiWS*

SentiWS wurde durch die Kombination verschiedener Verfahren und Ressourcen erstellt. Eine detaillierte Beschreibung der Erzeugung findet sich in Kapitel 2.3.2.1. Das Lexikon enthält 3.468 lemmatisierte Begriffe, unterteilt in zwei Listen mit positiven und negativen Wörtern. Eine Definition neutraler Begriffe existiert nicht. Nach der Filterung der Adjektive ergab sich die in Abbildung 16 gezeigte Verteilung. Dabei wurde deutlich, dass sich die meisten der 1.509, laut Liste, meinungstragenden Adjektive in Bin 0 befanden. Dagegen befanden sich am oberen und unteren Ende der Skala nur sehr wenige Wörter, was auf eine Verschiebung der

Werte in Richtung Mitte, d.h. auf Sentiment Values um 0, hindeutete.

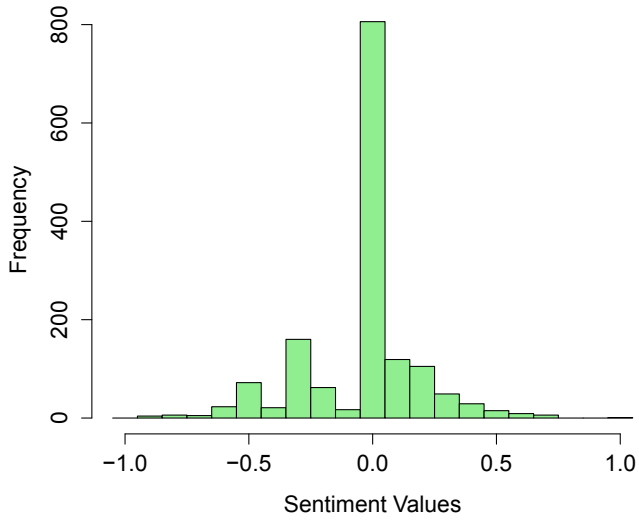


Abbildung 16: SentiWS: Verteilung der meinungstragenden Wörter nach *Sentiment Values* (*SV*)

Bei der Evaluation der lexikalischen Ressource *SentiWS* mit dem erzeugten Referenzlexikon zeigten sich folgenden Ergebnisse, die in Tabelle 29 dargestellt werden. Das Lexikon enthält 64,4% der als subjektiv klassifizierten häufigsten deutschen Adjektive. Werden ausschließlich als „stark wertend“ eingestufte Adjektive betrachtet, erreicht die Liste eine Trefferquote von 83,3%.

Meinungsklasse	Anzahl relevanter Adjektive	Anzahl enthaltener Adjektive (absolut)	Anzahl enthaltener Adjektive (relativ)
subjektiv	811	522	64,4%
schwach wertend	577	327	56,7%
stark wertend	234	195	83,3%

Tabelle 29: SentiWS: Anzahl enthaltener Referenzwörter

Die Bestimmung des Korrelationskoeffizienten r nach Pearson, zur Feststellung des Zusammenhangs der Sentiment Values für die gemeinsamen Adjektive von *SentiWS* und des Referenzlexikons, ergab folgende Ergebnisse für die verschiedenen Meinungsklassen (siehe Tabelle 30). Diese Werte sprechen sowohl bei subjektiven als auch bei stark wertenden Adjektiven für einen signifikanten Zusammenhang der Sentiment Values vorhandener Adjektive beider Listen.

Meinungsklasse	Pearson's r	p-value
subjektiv	$0,752$	$< 10e{-}15$
stark wertend	$0,828$	$< 10e{-}15$

Tabelle 30: Korrelationskoeffizienten SentiWS und Referenzlexikon

Eine abschließende Untersuchung der Differenz der Sentiment Values ($|SV|$) beider Listen – für alle subjektiven Adjektive – sollte zeigen, ob diese gleiche bzw. ähnliche Meinungswerte besaßen oder ob, wie durch die Verteilung in Abbildung 16 vermutet, die Werte verschoben waren. Obwohl in Tabelle 30 gezeigt werden konnte, dass ein starker Zusammenhang zwischen den Meinungswerten der wertenden Adjektive aus SentiWS und denen des Referenzlexikons bestand, zeigt Abbildung 17 eine klare Verschiebung der Werte um $\bar{x} = 0,5$. Dies könnte sich negativ bei der Anwendung des Lexikons auswirken, da die tatsächliche Stärke einer Meinungsäußerung nicht zuverlässig erkannt werden kann.

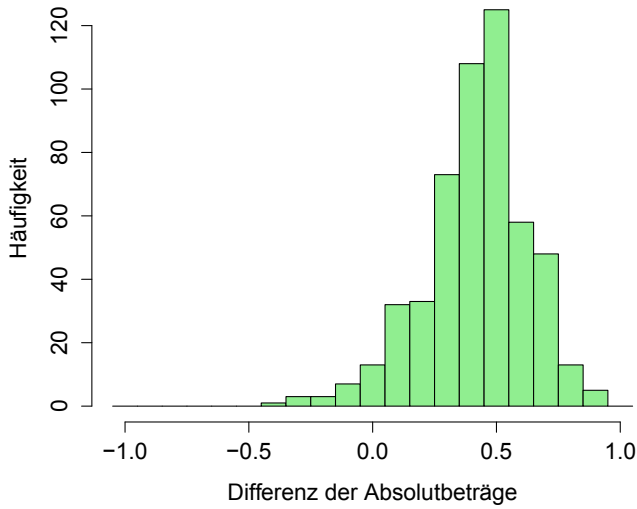


Abbildung 17: SentiWS: Schwankung der absoluten *Sentiment Values* (*SV*) gegenüber Referenzlexikon

6.2.3 *GermanPolarityClues*

GermanPolarityClues wurde 2010 von Waltinger [60] entwickelt und enthält 10.141 meinungstragende Wörter. Eine detaillierte Beschreibung der Erzeugung findet sich in Kapitel 2.3.2.2. Nach einer Datenvorverarbeitung, bei der alle Nomen und Verben entfernt wurden, blieben für die Evaluation 2.600 Adjektive erhalten. Es ergab sich eine Verteilung der Sentiment Values, die stark der Verteilung von *SentiWS* ähnelt (siehe Abbildung 18). Da in der Erzeugung der letzten Version dieser Liste alle Wörter und Werte von *SentiWS* integriert wurden, war dieses Verhalten allerdings nicht überraschend.

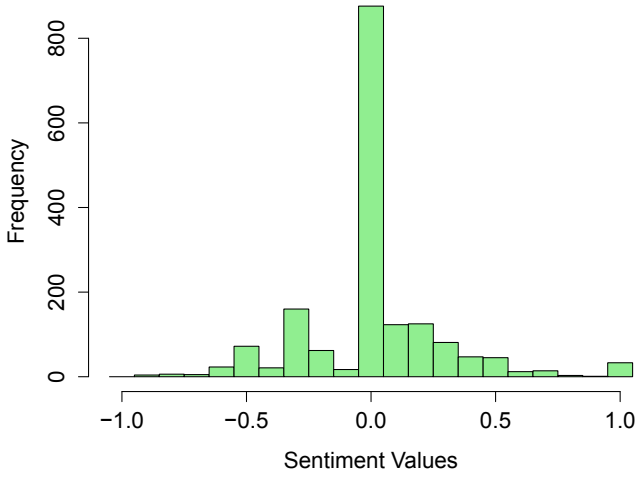


Abbildung 18: GermanPolarityClues: Verteilung der meinungstragenden Wörter nach Sentiment Values

Bei der Evaluation dieser Ressource mit dem erzeugten Referenzlexikon zeigten sich ebenfalls Parallelen zu *SentiWS*. In Tabelle 31 werden die Ergebnisse der Evaluation dargestellt. Die Ressource enthält 69,1% der als subjektiv klassifizierten häufigsten deutschen Adjektive. Von den „stark wertenden“ Adjektiven des Referenzlexikons enthält die Liste 85,5%. Damit schnitt *GermanPolarityClues* in allen Klassen leicht besser ab als *SentiWS*, die Teilmenge dieser Liste ist.

Meinungsklasse	Anzahl relevanter Adjektive	Anzahl enthaltener Adjektive (absolut)	Anzahl enthaltener Adjektive (relativ)
subjektiv	811	560	69,1%
schwach wertend	577	360	62,4%
stark wertend	234	200	85,5%

Tabelle 31: GermanPolarityClues: Anzahl enthaltener Referenzwörter

Die Bestimmung des Korrelationskoeffizienten r nach Pearson, ergab für diese Liste folgende Ergebnisse (siehe Tabelle 32) für die verschiedenen Meinungsklassen. Die jeweiligen Werte für r waren etwas schlechter, als es bei *SentiWS* der Fall war, sprachen aber dennoch für einen signifikanten Zusammenhang der Sentiment Values von Referenzlexikon und *GermanPolarityClues*.

Meinungsklasse	Pearson's r	p-value
subjektiv	0,720	$< 10e{-}15$
stark wertend	0,814	$< 10e{-}15$

Tabelle 32: Korrelationskoeffizienten GermanPolarityClues und Referenzlexikon

Auch die Untersuchung der Differenz der Sentiment Values ($|SV|$) beider Listen zeigte eine starke Ähnlichkeit zu der Verteilung von *SentiWS* (Abbildung 17). Die Werte der lexikalischen Ressource *GermanPolarityClues* waren ebenfalls um $\bar{x} = 0,40$ verschoben, siehe Abbildung 19.

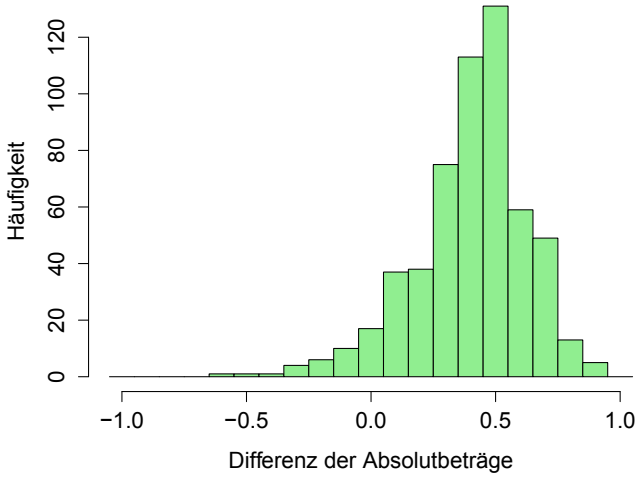


Abbildung 19: GermanPolarityClues: Schwankung der absoluten Sentiment Values gegenüber dem Referenzlexikon

6.2.4 *GermanLex*

GermanLex (siehe Kapitel 2.3.2.3) war die einzige lexikalische Ressource im Test, bei der die Werte nicht auf einer kontinuierlichen Skala verteilt waren. Stattdessen wurde die Stärke der Meinung durch fünf diskrete Werte dargestellt, $1,0$ und $0,7$ jeweils für positiv und negativ sowie 0 für neutrale Begriffe. Abbildung 20 zeigt die Verteilung der gefilterten 4.215 Adjektive (vorhandene Nomen und Verben wurden entfernt). Aus Gründen der Übersichtlichkeit wurden die Werte für negativ klassifizierte Begriffe mit einem negativen Vorzeichen versehen.

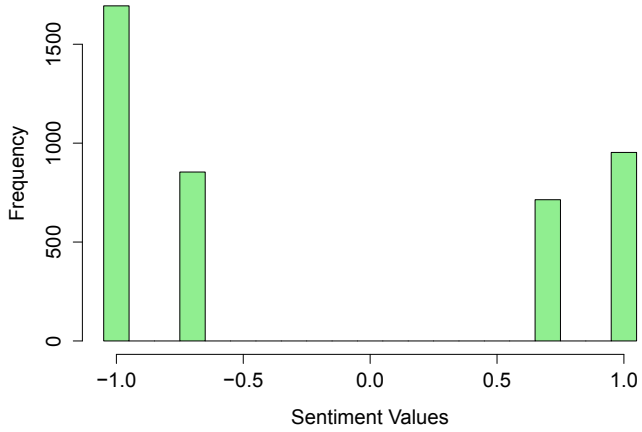


Abbildung 20: GermanLex: Verteilung der meinungstragenden Wörter nach Sentiment Values

Aufgrund der fehlenden kontinuierlichen Werteskala wurde im Folgenden lediglich eine Evaluation der „Vollständigkeit", bezogen auf vorhandene Adjektive des Referenzkorpus, durchgeführt. Dabei wurden wiederum verschiedene Meinungsklassen bzw. Teilmengen dieser betrachtet. Tabelle 33 zeigt die Ergebnisse der Evaluation. *GermanLex* enthält 59,9% der als subjektiv klassifizierten Adjektive des Referenzkorpus. Werden stark wertende Adjektive gesondert betrachtet, erreicht die lexikalische Ressource dabei eine „Vollständigkeit" von 82,1%.

Meinungsklasse	Anzahl relevanter Adjektive	Anzahl enthaltener Adjektive (absolut)	Anzahl enthaltener Adjektive (relativ)
subjektiv	811	486	59,9%
schwach wertend	577	294	50,9%
stark wertend	234	192	82,1%

Tabelle 33: GermanLex: Anzahl enthaltener Referenzwörter

6.2.5 *Sentiment Phrase List (SePL)*

Die *Sentiment Phrase List* (siehe Kapitel 2.3.2.4) wurde in der Forschungsgruppe „Analytische Informationssysteme“ entwickelt und ist die erste Ressource im Test, die neben Einzelwörtern zusätzlich auch Phrasen enthält. Basis dieser Liste war ein Korpus mit Amazon Kundenrezensionen. Bei der Verteilung der Sentiment Values wurden nur die, für die Evaluation relevanten einzelnen Adjektive betrachtet. Phrasen und andere Wortformen wurden durch einen Filterschritt ausgeschlossen. Abbildung 21 zeigt die Verteilung der 1.183 wertenden Adjektive. Auffällig war dabei vor allem das starke Übergewicht positiver Adjektive, mit einer hohen Anzahl an Begriffen zwischen $0,8$ und $1,0$.

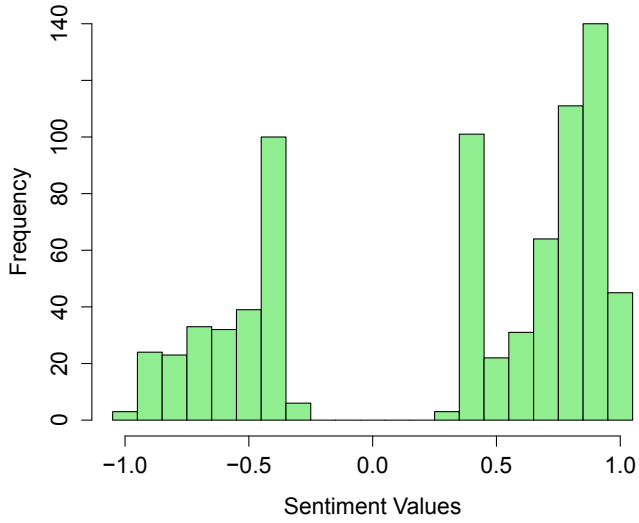


Abbildung 21: SePL: Verteilung der meinungstragenden Wörter nach Sentiment Values

Tabelle 34 zeigt die Ergebnisse der Evaluation durch den Referenzkorpus. Die *Sentiment Phrase List* enthält 53,8% der relevanten subjektiven Adjektive. Von den als „stark wertend" klassifizierten Adjektiven wurden 67,1% getroffen.

Meinungsklasse	**Anzahl relevanter Adjektive**	**Anzahl enthaltener Adjektive (absolut)**	**Anzahl enthaltener Adjektive (relativ)**
subjektiv	811	436	53,8%
schwach wertend	577	279	48,4%
stark wertend	234	157	67,1%

Tabelle 34: SePL: Anzahl enthaltener Referenzwörter

Die Bestimmung des Korrelationskoeffizienten ergab folgende Ergebnisse für die gemeinsam vorkommenden Adjektive in den

verschiedenen Meinungsklassen (siehe Tabelle 35). Auch in diesem Fall zeigten die Werte, für beide betrachteten Klassen, einen starken Zusammenhang zwischen den Sentiment Values der Adjektive der *Sentiment Phrase List* und denen des Referenzlexikons.

Meinungsklasse	**Pearson's r**	**p-value**
subjektiv	0,826	$< 10e{-}15$
stark wertend	0,910	$< 10e{-}15$

Tabelle 35: Korrelationskoeffizienten SePL und Referenzlexikon

Die abschließende Betrachtung der Differenz der Sentiment Values ($|SV|$) der *Sentiment Phrase List* und des Referenzlexikons zeigte keine Auffälligkeiten (siehe Abbildung 22). Die Differenz schwankte zwar, allerdings zeigte ein Großteil der Begriffe kaum Abweichungen von 0 ($\bar{x} = 0,01$). Das heißt, dass neben dem starken Zusammenhang zwischen den Meinungswerten der beiden Lexika – gezeigt in Tabelle 35 – auch die absoluten Werte der gefundenen subjektiven Adjektive der *Sentiment Phrase List* denen der Referenz entsprachen.

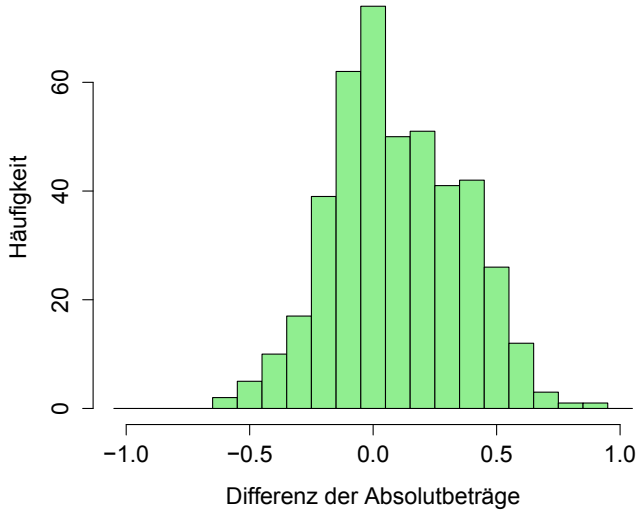


Abbildung 22: SePL: Schwankung der absoluten Sentiment Values gegenüber dem Referenzlexikon

6.2.6 *SentiMerge*

SentiMerge stellt, als das Produkt der Zusammenführung verschiedener anderer lexikalischer Ressourcen, einen Sonderfall der betrachteten Lexika dar. Eine detaillierte Beschreibung der Erzeugung findet sich in Kapitel 2.3.2.5. Bei der Betrachtung der Verteilung der Sentiment Values, wurde der von den Autoren der Liste vorgeschlagene „neutrale Bereich“ $[-0,23; 0,23]$ ausgeschlossen. Außerdem wurden abermals nur Adjektive betrachtet, andere Wortformen wurden gefiltert. Abbildung 23 zeigt die Verteilung der 5.701 subjektiven Adjektive von *SentiMerge*, die der des Referenzlexikons ähnelt.

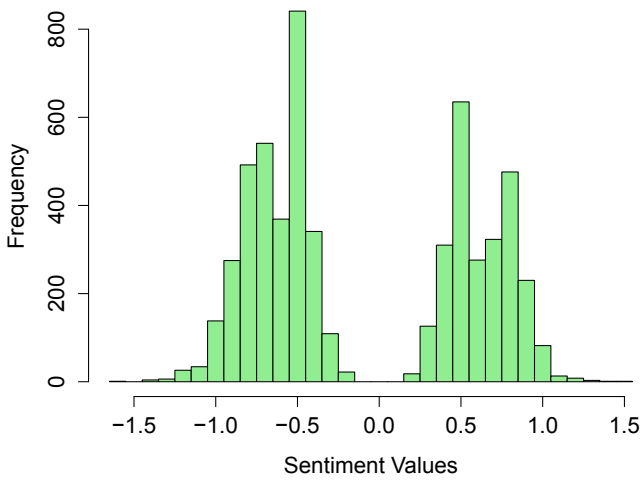


Abbildung 23: SentiMerge: Verteilung der meinungstragenden Wörter nach Sentiment Values

Bei der Evaluation durch das Referenzlexikon erreichte *SentiMerge* die besten Ergebnisse aller überprüfter Lexika. Tabelle 36 zeigt die Ergebnisse der Evaluation. Die lexikalische Ressource enthält 94,5% aller subjektiven Adjektive des Referenzlexikons. Dieser Wert wurde bei der Betrachtung der stark wertenden Adjektive mit einer Trefferquote von 98,3% sogar noch übertroffen.

Meinungsklasse	**Anzahl relevanter Adjektive**	**Anzahl enthaltener Adjektive (absolut)**	**Anzahl enthaltener Adjektive (relativ)**
subjektiv	811	766	94,5%
schwach wertend	577	536	92,9%
stark wertend	234	230	98,3%

Tabelle 36: SentiMerge: Anzahl enthaltener Referenzwörter

Auch bei der Überprüfung der Korrelation der Sentiment Values nach Pearson erzielte die lexikalische Ressource die besten Ergebnisse (siehe Tabelle 37). Mit einem Korrelationskoeffizienten $r = 0,898$ für alle subjektiven Adjektive und $r = 0,955$ für stark wertende, konnte ein signifikanter Zusammenhang zwischen den Sentiment Values von *SentiMerge* und denen des Referenzlexikons nachgewiesen werden.

Meinungsklasse	**Pearson's r**	**p-value**
subjektiv	0,898	$< 10e{-}15$
stark wertend	0,955	$< 10e{-}15$

Tabelle 37: Korrelationskoeffizienten SentiMerge und Referenzlexikon

Auch die Untersuchung der Differenz der Sentiment Values ($|SV|$) in Abbildung 24 zeigte keine Auffälligkeiten ($\bar{x} = 0$). Die absoluten Werte der gefundenen subjektiven Adjektive entsprachen also im Wesentlichen denen des Referenzlexikons.

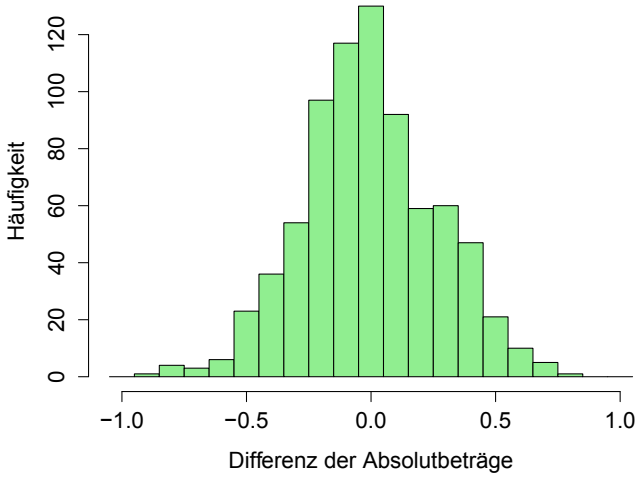


Abbildung 24: SentiMerge: Schwankung der absoluten Sentiment Values gegenüber Referenzlexikon

6.2.7 *Eigene Ressource*

Die im Rahmen dieser Arbeit erzeugt lexikalische Ressource wurde durch das in Kapitel 4 beschriebene Verfahren erzeugt. Dieses Lexikon ist das einzige im Test, welches keine Lemmata, sondern verschiedene Flexionen eines Wortes, enthält, z.B. „gute“, „guter“ und „gutes“. Durch den konsequenten Ausschluss sämtlicher NLP-Methoden, sind zudem keine POS-Tags enthalten, was die Evaluation weiter erschwerte. Abbildung 25 zeigt die Verteilung der 9.515 meinungstragenden Begriffe[1]. Hier zeigte sich, wie bereits bei der *Sentiment Phrase List* (Abbildung 21), das starke Übergewicht der positiven Begriffe. Dies hängt vor allem mit

1 Darin enthalten sind jegliche Wortformen und Flexionen.

dem verwendeten Amazon-Korpus zusammen, der mehr positive als negative Kundenrezensionen listet.

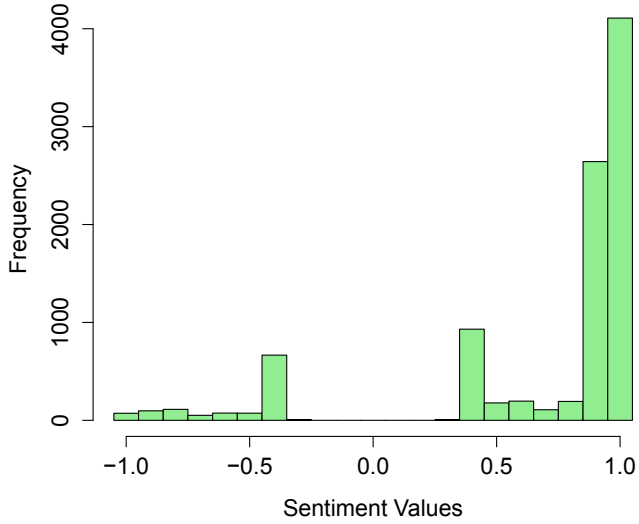


Abbildung 25: Eigene Ressource: Verteilung der meinungstragenden Wörter nach Sentiment Values

Bei der Evaluation der neu erzeugten lexikalischen Ressource durch das Referenzlexikon zeigten sich folgenden Ergebnisse (siehe Tabelle 38). Das Lexikon enthält 44,9% der häufigsten deutschen subjektiven Adjektive. Werden wiederum ausschließlich stark wertende Adjektive betrachtet, erreicht die neue Ressource eine Trefferquote von 56,4%.

Meinungsklasse	**Anzahl relevanter Adjektive**	**Anzahl enthaltener Adjektive (absolut)**	**Anzahl enthaltener Adjektive (relativ)**
subjektiv	811	360	44,9%
schwach wertend	577	228	39,5%
stark wertend	234	132	56,4%

Tabelle 38: Eigene Ressource: Anzahl enthaltener Referenzwörter

Tabelle 39 zeigt die Werte des Korrelationskoeffizienten r nach Pearson für die Adjektive der verschiedenen Meinungsklassen. Auch die Werte der neuen Liste zeigten einen signifikanten Zusammenhang der Sentiment Values der in beiden Listen vorhandenen Adjektive.

Meinungsklasse	**Pearson's r**	**p-value**
subjektiv	0,701	$< 10e{-}15$
stark wertend	0,816	$< 10e{-}15$

Tabelle 39: Korrelationskoeffizienten eigene Ressource und Referenzlexikon

Abschließend wurde die Differenz aller Sentiment Values ($|SV|$) gebildet, um die Abweichung der absoluten Meinungswerte vom Referenzlexikon festzustellen (siehe Abbildung 26). Die Werte der neuen Ressource schwankten zwar etwas stärker als jene von *SePL* (siehe Abbildung 22), allerdings waren die Abweichungen noch vertretbar ($\bar{x} = 0,08$).

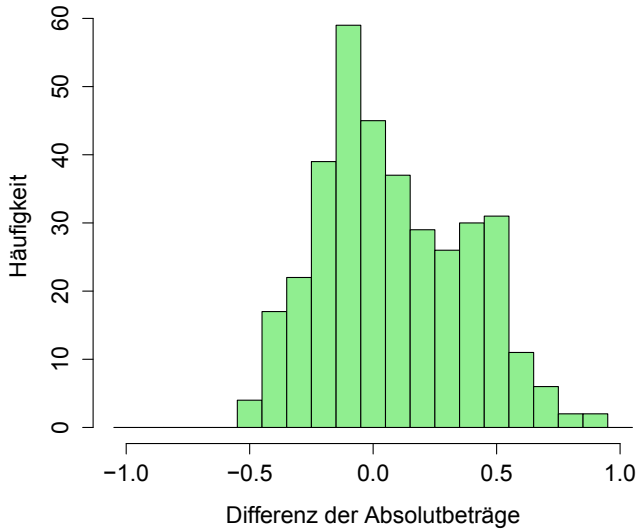


Abbildung 26: Eigene Ressource: Schwankung der absoluten Sentiment Values gegenüber Referenzlexikon

6.3 Zusammenfassung

Alle evaluierten Lexika zeigten zufriedenstellende Ergebnisse bei der Überprüfung der „Vollständigkeit", bezogen auf die Adjektive des Referenzlexikons, und gute bis sehr gute Ergebnisse bei der Evaluation der „Güte" dieser Adjektive, anhand der Ermittlung des Korrelationskoeffizienten nach Pearson. Aufgrund der in Abschnitt 6.2 benannten Einschränkungen sollten die Ergebnisse allerdings mit Vorsicht interpretiert werden.

Dennoch konnten durch die Evaluation Stärken und Schwächen der jeweiligen lexikalischen Ressourcen aufgezeigt werden. Außerdem stellten sich einige Besonderheiten heraus, wie die nahe „Verwandtschaft" der beiden Listen *SentiWS* und *GermanPo-*

larityClues, die bei jedem Test nahezu identische Werte und Verteilungen aufwiesen. Auch die enge Verbindung von *Sentiment Phrase List* und der im Rahmen dieser Arbeit erstellten Ressource wurde ersichtlich. Diese ergibt sich aus der Verwendung des selben Korpus sowie der identischen Formel zur Berechnung von Sentiment Values.

Etwas überraschend waren die sehr guten Ergebnisse der lexikalischen Ressource *SentiMerge*, die sowohl bei „Vollständigkeit", als auch bei „Güte" die besten Werte erzielte. Andererseits war diese Ressource eine der umfangreichsten im Test und brachte als Produkt der Kombination von vier deutschen Lexika die Voraussetzungen für diese Ergebnisse mit.

Obwohl *Sentiment Phrase List* und vor allem die in dieser Arbeit beschriebene Ressource bei der Evaluation weniger gut abschnitten als andere Lexika, sagt dies vorerst nichts über die Güte des Verfahrens zur Erzeugung und die Performance der Liste in *Opinion Mining* Anwendungen aus. Einerseits muss davon ausgegangen werden, dass der Amazon-Korpus, auf dem beide Lexika beruhen, nicht zwangsläufig die Vielfalt der deutschen Sprache abbildet und damit für das Opinion Mining wichtige Adjektive schlicht fehlen. Andererseits konnte die große Stärke beider Ressourcen, die Listung von Mehrwortphrasen, wie z.B. „nicht gut" oder „einfach nur schlecht", nicht gemessen bzw. bewertet werden, da weder das Referenzlexikon, noch die anderen deutschen Lexika solche Phrasen enthalten.

Aufgrund der oben genannten Einschränkungen überraschen die Ergebnisse des hier vorgestellten Algorithmus, der, trotz einiger Schwächen, aus einem gegebenen Textkorpus, ohne weitere Kenntnisse, vorgegebenen Seed oder NLP-Methoden und nur auf Basis statistischer Methoden, ein Lexikon mit meinungstragenden Wörtern und Phrasen erzeugte, das auch bei der Evaluation einzelner Adjektive befriedigende Ergebnisse zeigte.

Die Tabellen 40 und 41 stellen nochmals eine Zusammenfassung der Evaluationsergebnisse aller lexikalischen Ressourcen dar.

Ressource	subjektiv	schwach wertend	stark wertend
SentiWS	64,4%	56,7%	83,3%
GPC	69,1%	62,4%	85,5%
GermanLex	59,9%	50,9%	82,1%
SePL	53,8%	48,4%	67,1%
SentiMerge	94,5%	92,9%	98,3%
EIGENE	44,9%	39,5%	56,4%

Tabelle 40: Zusammenfassung Evaluation Referenzwörter

Ressource	Pearson's r (subjektiv)	Pearson's r (stark wertend)
SentiWS	0,752	0,828
GPC	0,720	0,814
GermanLex	-	–
SePL	0,826	0,910
SentiMerge	0,898	0,955
EIGENE	0,701	0,816

Tabelle 41: Zusammenfassung Evaluation Korrelationskoeffizienten

7

DISKUSSION MÖGLICHER ANWENDUNGSSZENARIEN

Im folgenden Kapitel werden mögliche Anwendungsszenerien des entwickelten Algorithmus sowie der erzeugten lexikalischen Ressource diskutiert. Sowohl der ursprüngliche Ausgangspunkt der Promotion, die *Analyse literarischer Texte* (Kapitel 7.2), als auch das klassische Einsatzgebiet des Opinion Mining in der Wirtschaft, die Unterstützung des (Online)-Marketings, exemplarisch in Form von *Beschwerde- und Reputationsmanagement* (Kapitel 7.1), werden dabei thematisiert.

Während die beiden erstgenannten Anwendungsszenerien die Ergebnisse des in der Arbeit vorgeschlagenen Algorithmus in Form neu erzeugter lexikalischer Ressourcen betreffen, bezieht sich das Anwendungsszenerio *Untersuchung von Sprachvarietäten* in Kapitel 7.3 auf die Anwendbarkeit des entwickelten Algorithmus in einem anderen Forschungsgebiet.

7.1 Beschwerde- und Reputationsmanagement

Ein Hauptanwendungsgebiet der Opinion Mining Algorithmen und Ressourcen, abseits der Forschung, ist der Einsatz in Systemen zur Erfassung der Kundenmeinungen über verschiedene Unternehmensbereiche. Sowohl die Wirksamkeit neuer Marketingmaßnahmen als auch die Zufriedenheit der Kunden mit der Serviceabteilung und selbst die Meinung der Kunden zu neuen Produkten, die Einfluss auf Neuentwicklungen haben kann, können mit solchen Systemen fortlaufend überprüft werden.

Viele namhafte Softwarehersteller bieten entsprechende Module für bestehende Analysesysteme an. Darunter *IBM*, mit *IBM WATSON*[1], *SAP*, mit *SAP HANA*[2] oder *Microsoft*, mit *Microsoft Azure*[3].

Alle Systeme werben damit, Meinungen oder Stimmungen in Texten verschiedener Sprachen erkennen und bewerten zu können. Im Detail treten dabei jedoch häufig Probleme auf, die auf unzureichende Algorithmen und Ressourcen zurückzuführen sind. Wie in Kapitel 2.2 ausgeführt wurde, existieren in jeder Sprache bestimmte Besonderheiten, die für eine exakte Analyse berücksichtigt werden müssen. Zudem ist eine Meinungsanalyse auf Dokumenten- oder Satzebene, wie sie oft von vorhandenen Businesstools durchgeführt wird, nicht exakt genug, da bei der Bestimmung der Tonalität auf dieser Ebene nicht klar ist, auf welchen Aspekt des Unternehmens sich die Meinungen der Kunden konkret beziehen. Die beiden größten Probleme und Fehler treten jedoch beim Einsatz der eingesetzten lexikalischen Ressourcen auf. Zum einen fehlen den, oft aus einem kleinen Set meinungstragender Wörter bestehenden, Listen wichtige Begriffe bei der Analyse von Texten aus einer bestimmten Domäne, bzw. wurden domänenabhängigen Begriffen falsche Polaritäten oder Meinungswerte zugeordnet[4]. Der Grund dafür sind manuell erstelle lexikalische Ressourcen oder Übersetzungen solcher. Wie bereits in Kapitel 2.3.1.2 deutlich gemacht wurde, funktionieren solche Übersetzung nicht oder nur sehr unzuverlässig. Wobei die Abdeckung vieler verschiedener Sprachen, wie sie von den Herstellern angeboten werden, momentan nicht anders ermöglicht werden kann. Die entsprechenden Ressourcen existieren schlicht nicht. Zum anderen basieren vorhandene Systeme auf der Analy-

1 https://www.ibm.com/watson/
2 https://www.sap.com/germany/products/hana.html
3 https://azure.microsoft.com/de-de/
4 Das Adjektiv „gruselig“ ist ein solcher Begriff. Es wird in der Domäne „Bücher“ als positiver, in der Domäne „Elektronik“ jedoch als negativer Begriff verwendet.

se von Einzelwörtern. Das bedeutet, dass Negationswörter sowie verstärkende und abschwächende Begriffe gesondert behandelt werden müssen, um korrekte Ergebnisse zu gewährleisten. Diese meinungsveränderten Begriffe müssen also erkannt und algorithmisch korrekt verarbeitet werden. Außerdem können solche Systeme aufgrund der genannten Beschränkung auf Einzelwörter (länderspezifische) Redewendungen, die ebenfalls wichtige Meinungsträger sein können, weder erkennen noch bewerten[5].

7.1.1 *Exemplarisches Analysesystem*

Im Folgenden wird ein System speziell für das Opinion Mining vorgeschlagen, das die oben genannten Probleme überwinden soll und speziell auf die Anwendung neuer lexikalischer Ressourcen, wie in dieser Arbeit vorgeschlagen, ausgerichtet ist. Die Grundlagen eines solchen kompletten Systems – speziell für die deutsche Sprache – wurden in der Forschungsgruppe „Analytische Informationssysteme“ bereits geschaffen. Einzelne Module müssen dabei jedoch noch überarbeitet bzw. finalisiert werden. Außerdem wurde dieses System für die semi-automatische Analyse von Texten erdacht, da frühere Projekte und Versuche zeigten, dass eine vollständig automatische Analyse von unbekannten Texten zu unbefriedigenden Ergebnissen führt. Vor allem die Aufbereitung der Rohdaten sowie die Generierung und Bereitstellung der lexikalischen Ressourcen, d.h. Lexika mit meinungstragenden Wörtern und Phrasen sowie Aspektlisten, variieren stark, abhängig von Datenquelle und Domäne.

Abbildung 27 illustriert den exemplarischen Aufbau des Systems für die sukzessive Bearbeitung und Analyse eines Textes mit dem Ziel der Extraktion der darin enthaltenen Meinungen, von der Datenakquise bis zur Ergebnisvisualisierung. Die dazu benö-

5 Ein gutes Beispiel dafür ist die Redewendung „klein aber oho“.

tigten Ressourcen werden ebenfalls abgebildet, wobei sowohl die Generierung von Lexika meinungstragender Wörter und Phrasen als auch die Erzeugung der benötigten Aspekte als Subtasks ausgelagert wurden.

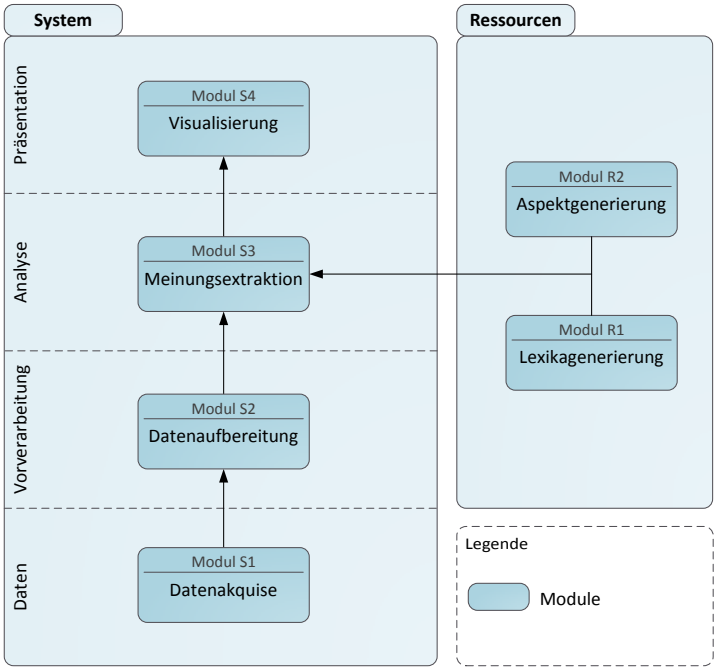


Abbildung 27: Opinion Mining System für Beschwerde- und Reputationsmanagement

Im Folgenden soll der Aufbau dieses Systems anhand der einzelnen Module verdeutlicht werden.

MODUL R1 – LEXIKAGENERIERUNG Für das vorgeschlagene Aspect-based Opinion Mining System werden Ressourcen mit meinungstragenden Wörtern und Phrasen benötigt. Erst anhand dieser Lexika können Meinungsäußerungen in Texten erkannt

werden. Innerhalb des Moduls R1 findet die Generierung dieser Ressourcen für die jeweilige Sprache und Domäne statt. Dabei kommt der in dieser Arbeit beschriebene Algorithmus zum Einsatz, der für die automatische Generierung der benannten Ressourcen ausschließlich einen passenden Textkorpus mit Bewertungen benötigt, wie er bei vielen Bewertungsportalen in Form von Kundenrezensionen zu finden ist. Zusammen mit den Informationen aus Modul R2, welches nachfolgend beschrieben wird, tragen beide Ressourcen, in Verbindung mit einem innovativen Algorithmus (Modul S3), zur Meinungsextraktion bei.

MODUL R2 – ASPEKTGENERIERUNG Neben Meinungslexika sind Listen mit relevanten Aspekten essentiell für das Aspekt-basierte Opinion Mining. Neben der manuellen Definition wichtiger Aspekte, können in diesem Modul verschiedene Algorithmen für deren Erzeugung bzw. Extraktion eingesetzt werden – beispielsweise mittels erweiterter Frequenzanalyse oder Topic Detection Methoden, wie der *Latent Dirichlet Allocation* (LDA) [4], welche hier jedoch nicht näher beschrieben werden sollen.

MODUL S1 – DATENAKQUISE Die große Bandbreite an verfügbaren Datenquellen, die Meinungsäußerungen enthalten, ist eine der größten Herausforderungen bei der Entwicklung dieses Moduls. Es wird dabei zwischen internen Daten, z.B. E-Mails, und externen Daten, z.B. Texte von Nutzern des Social Web, unterschieden. Während interne Daten durch vorhandene Schnittstellen oft einfach für eine Analyse verwendet werden können, sind externe Datenquellen, aufgrund der Heterogenität, eine größere Herausforderung. Es müssen entsprechende Schnittstellen entwickelt werden, um Zugriff auf die entsprechenden Daten zu erhalten.

MODUL S2 – DATENAUFBEREITUNG Innerhalb dieses Moduls findet die, für Opinion Mining Analysen, wichtige Datenvorverarbeitung statt. Dabei werden unnötige Textabschnitte entfernt und die Daten anschließend in ein standardisiertes Format überführt. Zu den nicht relevanten Bestandteilen eines Textes zählen dabei Anreden, z.B. „Sehr geehrte Damen und Herren", Grußformeln, z.B. „Mit freundlichen Grüßen", Signaturen oder Werbung. Außerdem werden doppelt auftretende Textstellen, wie Zitate, entsprechend markiert, um eine mehrfache Wertung zu vermeiden. Um den Text abschließend in ein standardisiertes Format zu überführen, wird dieser automatisch in Sätze und Wörter zerlegt und die Wortart der Wörter bestimmt.

MODUL S3 – MEINUNGSEXTRAKTION Wie bereits beschrieben, sind bestehende Verfahren nicht in der Lage Phrasen-basierte Lexika zu verarbeiten. Für dieses Kernmodul des Opinion Mining Prozesses wurde deshalb ein entsprechender Algorithmus zur Extraktion und Zuordnung der meinungstragenden Phrasen zu Aspekten entwickelt. Dazu werden die lexikalischen Ressourcen der Module R1 und R2 verwendet. Dieser Algorithmus berücksichtigt dabei auch Sonderfälle, wie auseinander stehende Phrasen. Ein Beispiel dafür ist die Phrase „nicht gut", die entweder zusammen – *„Der Service ist nicht gut."* – oder getrennt – *„Gut finde ich den Service nicht."* – verwendet werden kann. Auch das Auftreten von mehreren Phrasen pro Satz kann von diesem Algorithmus behandelt werden. Nach der Extraktion und Zuordnung von meinungstragenden Phrasen und Aspekten, werden diese in Form von Opinion Quintupeln (siehe Kapitel 2.2.2.3) bereitgestellt. Dazu müssen abschließend noch Autor, Zeitpunkt und Entität zugeordnet werden.

MODUL S4 – VISUALISIERUNG Dieses Output-Modul dient der Präsentation der ermittelten Daten. Eine aussagekräftige Präsentation der Ergebnisse steht dabei im Vordergrund. Die in Modul S3 erzeugten Opinion Quintupel können in diesem Modul auf verschiedenen Ebenen aggregiert und angezeigt werden. Da je nach Einsatzzweck verschiedene Detaillierungsgrade notwendig sein können, wird eine Visualisierung von Meinungsäußerungen auf Aspektbasis vorgeschlagen. Diese ermöglicht eine schnelle Übersicht über einzelne Produkte und Dienstleistungen sowie deren Eigenschaften. Durch die Betrachtung aller Meinungsäußerungen zu einem Aspekt, z.B. sortiert nach dem Grad der Negativität, kann gezielt analysiert werden, in welchem Bereich vermehrt (negative) Äußerungen auftreten.

Anschließend soll der Ablauf, beim Einsatz eines solchen Systems, anhand zweier Anwendungsfälle verdeutlicht werden. Die genannten Anwendungsfälle unterscheiden sich dabei lediglich durch den Erkenntnisgewinn der eingesetzten Opinion Mining Methoden bzw. durch den Detaillierungsgrad.

7.1.2 *Anwendungsfall 1: Beschwerdemanagement*

Fallbeispiel: *Der Kunde eines Versicherungsunternehmens hat Probleme bei der Schadensregulierung und beschwert sich per E-Mail. Dabei droht er mit der Kündigung seines Vertrags.*

Dieser erster Anwendungsfall adressiert die Optimierung des Beschwerdemanagements von Unternehmen und beinhaltet die Analyse von unternehmensinternen Datenquellen. Solche Datenquellen können E-Mails oder Texte aus unternehmenseigenen Serviceplattformen sein. Das Ziel dieses Anwendungsfalls ist es, einem entsprechenden Unternehmen die technischen Möglichkei-

ten zur schnellen und gezielten Reaktion auf Kundenbeschwerden zu bieten.

Im oben genannten Fallbeispiel ist eine möglichst kurze Reaktionszeit essentiell, um den Kunden nicht zu verlieren. Dafür kann der Einsatz von Opinion Mining Methoden von Vorteil sein. Durch die Analyse des gesamten E-Mail-Verkehrs können kritische Szenarien – wie eine Beschwerde mit Kündigungsdrohung – erkannt und automatisch Benachrichtigungen an die zuständigen Mitarbeiter der entsprechenden Abteilung gesendet werden. Für diesen Anwendungsfall ist es nicht notwendig, jede Meinungsäußerung im Text zu identifizieren. In den meisten Fällen ist es ausreichend, mindestens eine stark negative Meinung zu erkennen, um dem entsprechenden Dokument damit eine hohe Priorität zuzuweisen und eine Benachrichtigung auszulösen.

7.1.3 *Anwendungsfall 2: Reputationsmanagement*

Fallbeispiel: *Ein Unternehmen startet eine neue Imagekampagne, da in letzter Zeit vorwiegend negativ, sowohl über die Produkte, als auch über das Unternehmen im Web 2.0 gesprochen wurde. Die Marketingabteilung möchte wissen, wie sich diese Kampagne zeitlich auf das Meinungsbild des Unternehmens im Web 2.0 auswirkt.*

Der zweite Anwendungsfall thematisiert die Verwendung von Opinion Mining Methoden im Bereich des Reputationsmanagements. Dazu können frei verfügbare Texte des Web 2.0 verwendet werden, um Meinungsbilder für Unternehmen zu erstellen. Auf diese Weise wird ein effektives Monitoring der Meinungen zu Unternehmen bzw. zu deren Produkten oder Dienstleistungen ermöglicht. Neben einmaligen Analysen, die den Stand der Meinungen zu einem bestimmten Zeitpunkt widerspiegeln, sind auch Zeitreihenanalysen möglich. Dadurch wird es möglich, den Verlauf der Meinungen zu einem Unternehmen über einen länge-

ren Zeitraum zu untersuchen und somit zum Beispiel die Wirksamkeit von Imagekampagnen zu prüfen, wie im Fallbeispiel beschrieben. Auslöser für negative Meinungsäußerungen können dadurch schnell identifiziert und Gegenmaßnahmen zeitnah initiiert werden.

Für diesen Anwendungsfall werden sowohl hinreichend gute Lexika mit meinungstragenden Wörtern und Phrasen, als auch möglichst vollständige Aspektlisten für das Aspekt-basierte Opinion Mining benötigt. Die Meinungslexika sollten dabei vor allem auch die wichtigsten meinungstragenden Wörter und Phrasen aus der entsprechenden Domäne enthalten. Auch der Zugriff auf geeignete Datenquellen spielt eine entscheidende Rolle.

7.2 Analyse literarischer Texte

Das folgende Anwendungsszenario adressiert die ursprüngliche Idee der Promotion, die Analyse literarischer Texte mit Methoden aus dem Bereich des Opinion Mining. Dieser hypothetische Anwendungsfall kann damit dem Forschungsbereich der *Digital Humanities* zugeordnet werden. In einer ersten eigenen Arbeit zu diesem Thema wurde die Verwendbarkeit vorhandener lexikalischer Ressourcen [42] in diesem Bereich überprüft.

Die grundlegende Idee hinter dem Einsatz von Opinion Mining Methoden im Bereich der Literaturwissenschaften ist die Entwicklung einer neuen Informationsebene. Bisher beschränken sich die Informationen über literarische Texte auf deren Metadaten sowie quantitative Größen, wie beispielsweise Anzahl der Seiten, Erscheinungsdatum, Titel des Werkes, Name des Autors, Gattung bzw. Genre etc. Des Weiteren stehen oft weitere verwandte Informationen in Form von Kritiken und, im Zeitalter des Web 2.0, Kundenrezensionen von verschiedenen Bewertungsportalen zur

Verfügung. Anhand dieser Informationen können schließlich diverse Suchanfragen formuliert werden, zum Beispiel:

- *Wie lauten die Titel aller Märchen der Brüder Grimm?*
- *Welche Werke des Autors Stephen King erschienen im Jahr 2013?*

Durch den Einsatz von Opinion Mining Methoden, d.h. durch die Analyse der Texte, einer anschließender Aufbereitung und Speicherung der Ergebnisse, entsteht eine gänzlich neue Dimension für die Suche – eine Suche basierend auf Stimmungen und Emotionen. Erste Arbeiten zu diesem Thema finden sich in den Veröffentlichungen von Mohammad and Turney [35] und Mohammad [34]. Die Autoren erzeugten eine lexikalische Ressource mit englischsprachigen emotionstragenden Begriffen und analysierten damit exemplarisch die Märchen der Brüder Grimm. Den Begriffen wurden dabei jeweils einer der acht Basisemotionen nach Plutchik [39] – *Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger, Anticipation* – zugeteilt.

Durch die neue Informationsebene entstehen auch neue Möglichkeiten bei der intratextuellen sowie der intertextuellen Suche. Im Folgenden sollen zwei Beispiele für Suchanfragen die Mächtigkeit bzw. die Möglichkeiten dieser neuen Suche verdeutlichen:

- *Welches ist das düsterste Märchen der Brüder Grimm?* (intertextuelle Suche)
- *In welchem Textabschnitt gibt es die stärksten Stimmungsschwankungen?* (intratextuelle Suche)

Durch den Einsatz von Opinion Mining Methoden und Ressourcen können Texteinheiten, beispielsweise Sätzen, Absätzen oder Kapiteln, Meinungen und Stimmungen zugeordnet werden. Abbildung 28 illustriert vereinfacht diese Zuordnung mittels phrasenbasiertem Opinion Mining auf Kapitelebene. Der so entstehende „Fingerabdruck“ bzw. die „Landkarte“ eines literarischen

Textes kann anschließend in einer entsprechenden Datenbank gespeichert und verwendet werden.

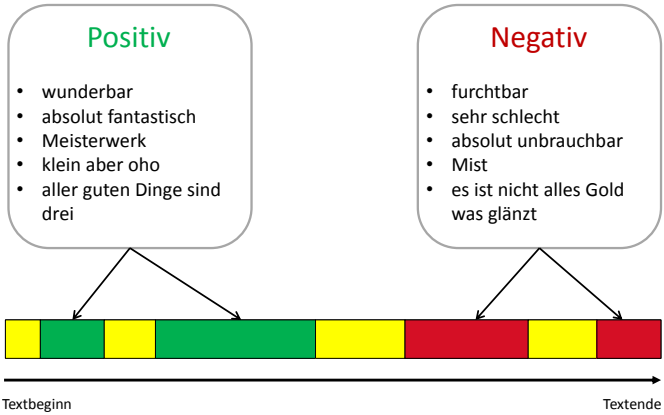


Abbildung 28: Opinion Mining für literarische Texte

Neben der oben beschriebenen Erweiterung der Suche, ermöglichen die Opinion Mining Analysen bzw. die automatische Zuordnung von Textabschnitten zu Meinungen und Stimmungen die Betrachtung weiterer Fragestellungen. Zum einen könnten mögliche Unterschiede hinsichtlich Grundstimmung von Werken aus unterschiedlichen Literaturepochen untersucht werden – Fragestellung: *Sind Werke der Nachkriegsliteratur grundsätzlich negativer als Werke der Moderne?*
Zum anderen könnte die Güte von Übersetzungen überprüft werden – Fragestellung: *Weißt ein übersetztes Werk denselben bzw. einen ähnlichen Stimmungsverlauf wie das Original auf?*
Des Weiteren sind diverse Untersuchungen zu Autoren vorstellbar – Fragestellung: *Wie verändert sich die Stimmung in den Werken des Autors über einen Zeitraum von mehreren Jahren?*

Für die Umsetzung dieses Anwendungsszenarios wird, ähnlich zur Anwendung in Abschnitt 7.1, zum einen ein Lexikon

meinungstragender Wörter und Phrasen und zum anderen ein entsprechender Algorithmus benötigt, der diese Phrasen identifizieren und Textabschnitten aggregierte Meinungen zuordnen kann. Ein für dieses Anwendungsszenario benötigtes System hätte einen ähnlichen Aufbau wie das zuvor vorgestellte System in Abbildung 27. Modul R2 „Aspektgenerierung“ wird dabei nicht benötigt. Außerdem wird die Datenvorverarbeitung in Modul S2 durch grammatikalisch und orthografisch einwandfreie Texte erheblich vereinfacht. Die größte Änderung betrifft den Übergang von der Analyseschicht in die Präsentationsschicht. Nachdem Meinungen und Stimmung extrahiert bzw. zugeordnet wurden, müssen die Informationen in einer geeigneten Datenbank, für die spätere Verwendung, abgespeichert werden. Eine Visualisierung des Textes („Fingerabdruck“, siehe Abbildung 28) ist hingegen optional. Die verwendete Datenbank sollte über die entsprechenden Schnittstellen verfügen, um die oben genannten Suchanfragen und Fragestellungen bearbeiten zu können.

Grundlage des beschriebenen Szenarios bilden lexikalische Ressourcen, die zur Analyse literarische Texte geeignet sind. Mit dem in dieser Arbeit beschriebenen Verfahren ist es möglich, solche Ressourcen automatisch zu erzeugen – geeignete Ausgangskorpora vorausgesetzt. Die Besonderheit von lexikalischen Ressourcen für literarische Texte ist die variantenreiche Sprache, die sich zudem je nach Epoche und Autor teils stark unterscheidet. Erste Versuche mit den Märchen der Brüder Grimm zeigten dies deutlich [42]. Vorhandene lexikalische Ressourcen mit aktuellen Wörtern und Phrasen konnten nur einen kleinen Teil der im Text vorhandenen Meinungen und Stimmungen erkennen. Für eine aussagekräftige Analyse muss also zunächst eine geeignete Quelle identifiziert werden, durch die es dem in der Arbeit beschriebenen Algorithmus ermöglicht wird, eine lexikalische Ressource für literarische Texte zu erzeugen. Dabei wird zwar keine abso-

lute Vollständigkeit benötigt, was ohnehin unmöglich ist, allerdings müssen relevante meinungstragende Begriffe vorhanden sowie deren Polarität korrekt sein.

7.3 Untersuchung von Sprachvarietäten

Während die beiden vorher genannten Anwendungsszenarien eher dem Bereich des „klassischen" Opinion Mining zuzuordnen sind, da in diesen entsprechende lexikalische Ressourcen und Algorithmen angewendet werden, stammt die in diesem Kapitel beschriebene Anwendung aus dem Bereich der Sprachstatistik. Dabei soll vor allem die Anwendbarkeit des neu entwickelten Algorithmus in einem anderem Forschungsfeld diskutiert werden. Die Idee der Untersuchung von Sprachvarietäten entstand durch entsprechende Kontakte im Laufe der Promotion und zeigt die Möglichkeiten der interdisziplinären Anwendung der entwickelten Methode.

Die grundlegende Idee bei dieser Analyse ist die Verwendung des Algorithmus zur automatischen Erzeugung von Lexika meinungstragender Wörter und Phrasen in einem anderen Kontext. Dabei wird das im Algorithmus verwendete Bewertungssystem zur Bestimmung der Polarität und Intensität von meinungstragenden Wörtern und Phrasen dahingehend angepasst, sodass die „Sterne" nunmehr den zu analysierenden Sprachvarianten entsprechen. Als Beispiel dafür dient ein Textkorpus mit zwei verschiedenen Sprachvarianten der englischen Sprache, z.B. diverse Texte aus Großbritannien sowie den USA. Der entwickelte Algorithmus wird so angepasst, dass den Texten aus Großbritannien fiktive fünf Sterne und den Texten aus den USA jeweils ein fiktiver Stern zugeordnet werden. Diese Zuordnung ermöglicht es, signifikante Phrasen für bestimmte Sprachvarianten zu erkennen. Im erwähnten Beispiel bekämen signifikante Phrasen, die haupt-

sächlich in Großbritannien verwendet werden, nach der Analyse einen fiktiven Meinungswert um $+1$. Signifikate Phrasen aus den USA hätten dagegen einen fiktiven Wert um -1. Phrasen, die in beiden Ländern häufig verwendet werden, lägen bei einem Wert um 0. Die Ergebnisse dieser Analysen könnten anschließend mit anderen Ansätzen verglichen werden.

Um die Machbarkeit des oben beschriebenen Ansatzes zu überprüfen, wurden erste Versuche mit dem *International Corpus of English (ICE)* [57] durchgeführt. Dieser Korpus enthält verschiedene Varietäten der englischen Sprache, u.a. Texte aus den Ländern Großbritannien, Kanada, USA und den Philippinen. Für erste Versuche wurden dazu Pressetexte aus Großbritannien (GB) mit denen der Philippinen (PHI), mittels des beschrieben und angepassten Algorithmus, verglichen. Dazu wurden die ursprünglichen Texte zunächst bereinigt und für die Analyse aufbereitet, d.h. alle Sätze wurden mit fiktiven Sternebewertungen markiert.

In den nachfolgenden Tabellen werden die Ergebnisse dieser Analyse auszugsweise dargestellt. Tabelle 42 enthält Beispiele für signifikante Wörter und Phrasen des Philippinischen. Tabelle 43 enthält im Gegensatz dazu Beispiele für signifikante Wörter und Phrasen des Britisch Englischen. Tabelle 44 zeigt Begriffe, die in beiden in etwa mit der gleichen Häufigkeit verwendet werden.

Wort / Phrase	**Länge**	**Frequenz**	**Land**
estrada	1	104	PHI
philippine	1	59	PHI
manila	1	57	PHI
marcos	1	52	PHI
ramos	1	44	PHI
president said	2	64	PHI
estrada president	2	60	PHI
president ramos	2	19	PHI
estrada president said	3	23	PHI

Tabelle 42: Beispiele für signifikante Wörter und Phrasen (PHI)

Wort / Phrase	Länge	Frequenz	Land
london	1	39	GB
labour	1	38	GB
scottish	1	30	GB
thatcher	1	21	GB
liverpool	1	20	GB
mrs thatcher	2	17	GB
rather than	2	12	GB
government labour	2	11	GB

Tabelle 43: Beispiele für signifikante Wörter und Phrasen (GB)

Wort / Phrase	Länge	Frequenz	Land
to	1	1901	PHI + GB
of	1	1888	PHI + GB
in	1	1405	PHI + GB
from	1	405	PHI + GB
or	1	161	PHI + GB
good	1	36	PHI + GB

Tabelle 44: Beispiele für signifikante Wörter und Phrasen (PHI + GB)

Obwohl die signifikantesten bzw. häufigsten Wörter und Phrasen erwartungsgemäß länderspezifisch geprägt waren, konnte anhand der ersten Ergebnisse gezeigt werden, dass der Ansatz prinzipiell funktioniert. Allerdings müsste der Korpus für aussagekräftigere Ergebnisse noch erweitert werden – die Ergebnisse der oben gezeigten Analyse basierten auf ca. 4200 Sätzen. Auffällig bei den Analyseergebnissen war die Phrase „rather than“, die signifikant häufig in Texten aus Großbritannien vorkam. Allerdings wurde ebenfalls deutlich, dass für den Einsatz des Algorithmus im Umfeld der Untersuchung von Sprachvarietäten entsprechendes Fachwissen benötigt wird, um geeignete Fragestellungen formulieren und die Ergebnisse korrekt bewerten zu können.

8

DISKUSSION DER ERGEBNISSE

In diesem abschließenden Kapitel werden zunächst die Ergebnisse der vorliegenden Arbeit zusammengefasst und diskutiert. Im Anschluss folgen Vorschläge für Anknüpfungspunkte zukünftiger Arbeiten.

8.1 Zusammenfassung und Schlussfolgerungen

Die vorliegende Dissertation beschäftigte sich im Allgemeinen mit dem Aufbau, der Struktur und den Einsatzmöglichkeiten vorhandener lexikalischer Ressourcen für das Opinion Mining und im Besonderen mit der Generierung einer neuen lexikalischen Ressource basierend auf Textkorpora und statistischen Verfahren. Dabei wurde die Notwendigkeit für die Erzeugung neuer Lexika für spezielle Anwendungsfälle sowie für verschiedene Sprachen verdeutlicht. Ebenfalls wurden die Vorteile bei der Verwendung von Wortphrasen, im Gegensatz zu Einzelwörtern, innerhalb solcher Ressourcen aufgezeigt.

Im Rahmen der Promotion wurde ein generisches Verfahren entwickelt, mit dessen Hilfe lexikalische Ressourcen für das Opinion Mining, d.h. Lexika mit meinungstragenden Wörtern und Phrasen, automatisch aus geeigneten Textkorpora – das Vorhandensein eines Textes (Titel) sowie einer Bewertung vorausgesetzt, siehe Kapitel 4.1 – erzeugt werden können. Der entwickelte Algorithmus benötigt dabei keine besondere Kenntnis des Korpus und arbeitet sprachunabhängig, d.h. ohne den Einsatz von NLP-Methoden, wie POS-Tagging oder Lemmatisierung.

Mittels des entwickelten Algorithmus wurde schließlich ein neues Lexikon mit meinungstragenden Wörtern und Phrasen für die deutsche Sprache erzeugt. Basis dafür waren Kundenrezensionen von Amazon Deutschland. Trotz einiger bestehender Probleme, wie dem Fehlen der Reihenfolge bei Phrasen oder dem Auftreten unpassender Phrasen, konnte die Funktionsweise des neuen Ansatzes demonstriert und akzeptable Ergebnisse erzielt werden.
Bezogen auf die Einschränkungen vorhandener Lexika, wie in Kapitel 1.2 dargestellt, trägt der entwickelte Algorithmus dazu bei, die identifizierte Forschungslücke in diesem Bereich zu schließen. Vor allem die Betrachtung von Phrasen statt Einzelwörtern und die durch den Algorithmus ermöglichte Identifikation und Bewertung von meinungstragenden Redewendungen sprechen dabei für das neue sprachunabhängige Verfahren.

Des Weiteren wurde zum Zweck einer umfassenden Evaluation im Rahmen dieser Arbeit ein Referenzlexikon, bestehend aus den häufigsten deutschen Adjektiven, manuell erzeugt. Die Adjektive wurden dabei von der Website des *Duden Online* bezogen, inklusive der angegeben Häufigkeit. Das Experiment zur Erzeugung des Referenzlexikons mit 20 Teilnehmern verlief dabei ohne Probleme, sodass am Ende eine Liste mit 1.699 bewerteten Adjektiven entstand.
Bei der anschließenden Evaluation wurden die Stärken und Schwächen der vorhandenen deutschsprachigen lexikalischen Ressourcen sowie der im Rahmen dieser Arbeit erzeugten Ressource deutlich gemacht. Die eigene Ressource schnitt bei dieser Evaluation zwar nur durchschnittlich ab, allerdings konnten durch das Referenzlexikon nur Einzelwörter untersucht werden. Die größte Stärke der eigenen Ressource, die Verwendung von Phrasen, konnte dabei nicht gemessen werden.

Zum Abschluss dieser Arbeit wurden diverse Anwendungsszenarien für den entwickelten Algorithmus sowie für die dadurch entstandene lexikalische Ressource vorgestellt. Der Schwerpunkt lag dabei auf dem klassischen Einsatzgebiet des Opinion Mining in der Wirtschaft, der Unterstützung von Marketingabteilungen in Unternehmen, exemplarisch in Form des Beschwerde- und Reputationsmanagements. Es wurde demonstriert, wie die neu entwickelte Ressource gewinnbringend in geeigneten Opinion Mining Systemen eingesetzt werden kann, um die Meinungsanalyse für Unternehmen zu verbessern und detailliertere Ergebnisse zu erhalten. Zudem wurde der Einsatz bei der Analyse literarische Texte diskutiert und mit der Untersuchung von Sprachvariatäten ein interdisziplinäres Einsatzszenario für den entwickelten Algorithmus beschrieben.

8.2 Ausblick

Sowohl für den entwickelten Algorithmus zur sprachunabhängigen Generierung lexikalischer Ressourcen, als auch für das erzeugte Referenzlexikon existieren verschiedene Anknüpfungspunkte für zukünftige Arbeiten.

Zunächst sollten mit dem entwickelten Algorithmus weitere Experimente mit anderen Textkorpora durchgeführt und die Ergebnisse mit denen des Amazon-Korpus verglichen werden. Interessant wäre dabei vor allem die Identifikation von domänenspezifischen Begriffen. Des Weiteren sollte der Algorithmus auch mit Texten anderer Sprachen getestet werden, um die Funktionsweise zu überprüfen, Probleme zu identifizieren und die Grenzen dieses Verfahrens auszuloten – Fragestellung: *Können für alle Sprachen der indogermanischen Sprachfamilie lexikalische Ressourcen erzeugt werden?*

Außerdem sollten die in Kapitel 4.7 beschriebenen Probleme und Einschränkungen des vorgestellten Verfahrens näher untersucht und falls möglich behoben werden. Vor allem die fehlende Reihenfolge bei den Wörtern einer Phrase sowie das Auftreten unpassender Phrasen und Entitäten innerhalb der lexikalischen Ressource beeinträchtigen die Verwendung des Lexikons und sollten deshalb behoben werden.

Für den Einsatz in (eigenen) Aspekt- und Phrasen-basierten Opinion Mining Systemen entstand die Idee der Kombination der beiden Verfahren zur Erzeugung von *Sentiment Phrase List* (SePL) und der im Rahmen dieser Arbeit erstellten Ressource. Ziel dieses hybriden Verfahrens wäre eine umfangreiche lexikalische Ressource, die die Vorteile beider Verfahren kombiniert und sowohl lemmatisierte meinungstragende Phrasen in richtiger Reihenfolge als auch Redewendungen enthält.

Das erzeugte Referenzlexikon erwies sich als wertvolle Ressource bei der Evaluation lexikalischer Ressourcen. Um zukünftig die Bewertung von Phrasen-basierten Lexika zu ermöglichen, sollte dieser Korpus entsprechend der vorgestellten Methode weiterentwickelt werden. Denkbar dafür wäre die Erweiterung vorhandener meinungstragender Adjektive mit passenden Verstärkern, Abschwächern und Negationen. Die damit gebildeten Phrasen müssten anschließend wiederum von mehreren Annotatoren, im Rahmen eines entsprechenden Experiments, bewertet werden. Für die Auswahl der korrekten Meinungsshifter könnte ein „Vorschlagssystem“, beispielsweise basierend auf der *Google*-Suche[1], entwickelt werden, welches alle möglichen Kombinationen von *Partikel* + *Adjektiv* überprüft und die Häufigste, und damit die Wahrscheinlichste, zurückgibt.

1 https://www.google.de/

LITERATURVERZEICHNIS

[1] Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. In *Proceedings of the 20th International Conference on Very Large Data Bases*, 1994.

[2] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining Association Rules between Sets of Items in Large Databases. In *ACM SIGMOD Record*, volume 22, pages 207–216. ACM, 1993.

[3] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *LREC*, volume 10, pages 2200–2204, 2010.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of machine Learning research*, 3: 993–1022, 2003.

[5] Julian Brooke, Milan Tofiloski, and Maite Taboada. Cross-Linguistic Sentiment Analysis: From English to Spanish. In *RANLP*, pages 50–54, 2009.

[6] Erik Cambria, Robert Speer, Catherine Havasi, and Amir Hussain. SenticNet: A Publicly Available Semantic Resource for Opinion Mining. In *AAAI fall symposium: commonsense knowledge*, volume 10, 2010.

[7] Erik Cambria, Catherine Havasi, and Amir Hussain. SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis. In *FLAIRS conference*, pages 202–207, 2012.

[8] Erik Cambria, Björn Schuller, Yunqing Xia, and Catherine Havasi. New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems*, 28(2):15–21, 2013.

[9] Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal. SenticNet 3: A Common and Common-Sense Knowledge Base for Cognition-Driven Sentiment Analysis. In *Twenty-eighth AAAI conference on artificial intelligence*, 2014.

[10] Kenneth Ward Church and Patrick Hanks. Word Association Norms, Mutual Information, and Lexicography. *Computational linguistics*, 16(1):22–29, 1990.

[11] Simon Clematide and Manfred Klenner. Evaluation and Extension of a Polarity Lexicon for German. In *Proceedings of the First Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 7–13, 2010.

[12] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[13] Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[14] Kushal Dave, Steve Lawrence, and David M Pennock. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM, 2003.

[15] Ted Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational linguistics*, 19(1):61–74, 1993.

[16] F.L. Dyer and T.C. Martin. *Edison: His Life and Inventions*. Number Bd. 2 in Edison: His Life and Inventions. Harper & Brothers, 1910.

[17] Guy Emerson and Thierry Declerck. SentiMerge: Combining Sentiment Lexicons in a Bayesian Framework. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing*, 2014.

[18] Andrea Esuli and Fabrizio Sebastiani. SentiWordNet: A High-Coverage Lexical Resource for Opinion Mining. *Institute of Information Science and Technologies (ISTI) of the Italian National Research Council (CNR)*, 2006.

[19] Joseph Fleiss, Bruce Levin, and Cho Paik Myunghee. The Measurement of Interrater Agreement. *Statistical Methods for Rates and Proportions*, 2(212-236):22–23, 1981.

[20] Joseph L Fleiss and Domenic V Cicchetti. Inference About Weighted Kappa in the Non-Null Case. *Applied Psychological Measurement*, 2(1):113–117, 1978.

[21] Bibliographisches Institut GmbH. Duden – Die Verteilung der Wortarten im Rechtschreibduden, 2017. `http://www.duden.de/sprachwissen/sprachratgeber/die-verteilung-der-wortarten-im-rechtschreibduden`, zuletzt abgerufen am 25.10.2017.

[22] Bibliographisches Institut GmbH. Duden – Häufigkeit, 2017. `http://www.duden.de/hilfe/haeufigkeit`, zuletzt abgerufen am 25.10.2017.

[23] Vasileios Hatzivassiloglou and Kathleen R McKeown. Predicting the Semantic Orientation of Adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 174–181. Association for Computational Linguistics, 1997.

[24] Gerhard Heyer, Martin Läuter, Uwe Quasthoff, Thomas Wittig, and Christian Wolff. Learning Relations Using Collocations. In *Workshop on ontology learning*, volume 38, 2001.

[25] Minqing Hu and Bing Liu. Mining and Summarizing Customer Reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[26] Nan Hu, Paul A Pavlou, and Jie Jennifer Zhang. Why do Online Product Reviews have a J-shaped Distribution? Overcoming Biases in Online Word-of-Mouth Communication. *Communications of the ACM*, 2007.

[27] Nan Hu, Jie Zhang, and Paul A. Pavlou. Overcoming the J-shaped Distribution of Product Reviews. *Communications of the ACM*, 52:144–147, 2009.

[28] J Richard Landis and Gary G Koch. The Measurement of Observer Agreement for Categorical Data. *biometrics*, pages 159–174, 1977.

[29] Bing Liu. Sentiment Analysis and Subjectivity. *Handbook of natural language processing*, 2:627–666, 2010.

[30] Bing Liu. Sentiment Analysis and Opinion Mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[31] Bing Liu. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, 2015.

[32] George A Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[33] Amazon.de (User: M.K.). Customer Review, 2017. `https://www.amazon.de/gp/customer-reviews/R2VE0167UCB7H/`, zuletzt abgerufen am 14.12.2017.

[34] Saif Mohammad. From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114. Association for Computational Linguistics, 2011.

[35] Saif M Mohammad and Peter D Turney. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics, 2010.

[36] Tetsuya Nasukawa and Jeonghee Yi. Sentiment Analysis: Capturing Favorability using Natural Language Processing. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 70–77. ACM, 2003.

[37] Offene Daten Berlin. Liste der häufigen Vornamen 2013, 2013. `https://daten.berlin.de/datensaetze/liste-der-h%C3%A4ufigen-vornamen-2013`, zuletzt abgerufen am 17.12.2017.

[38] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: Sentiment Classification using Machine Learning Techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[39] Robert Plutchik. A General Psychoevolutionary Theory of Emotion. *Theories of emotion*, 1(3-31):4, 1980.

[40] Uwe Quasthoff. *Wörterbuch der Kollokationen im Deutschen*. Walter de Gruyter, 2011.

[41] Uwe Quasthoff and Christian Wolff. The Poisson Collocation Measure and its Applications. In *Second International Workshop on Computational Approaches to Collocations*. IEEE, 2002.

[42] Dirk Reinel. Evaluation der Qualität lexikalischer Ressourcen zur Stimmungserkennung in literarischen Texten. In *LWA*, pages 168–172, 2013.

[43] Dirk Reinel and Jörg Scheidt. Automatische Auswertung von Kundenmeinungen – Opinion Mining am Beispiel eines Projekts für die Versicherungswirtschaft. In *Dialogmarketing Perspektiven 2014/2015*, pages 129–149. Springer, 2015.

[44] Dirk Reinel, Jörg Scheidt, Andreas Henrich, and Niko Brucker. Sentiment Phrase Generation Using Statistical Methods. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, 2018. (in Kürze erscheinend).

[45] Robert Remus, Uwe Quasthoff, and Gerhard Heyer. SentiWS – A Publicly Available German-language Resource for Sentiment Analysis. In *LREC*, 2010.

[46] Sven Rill. *Themenerkennung in Twitter unter Berücksichtigung von Meinungsäußerungen: PoliTwi: eine Analyse politischer Themen*. PhD thesis, Goethe University Frankfurt, 2016.

[47] Sven Rill, Sven Adolph, Johannes Drescher, Dirk Reinel, Jörg Scheidt, Oliver Schütz, Florian Wogenstein, Roberto V Zicari, and Nikolaos Korfiatis. A Phrase-Based Opinion List for the German Language. In *KONVENS*, pages 305–313, 2012.

[48] Sven Rill, Jörg Scheidt, Johannes Drescher, Oliver Schütz, Dirk Reinel, and Florian Wogenstein. A Generic Approach to Generate Opinion Lists of Phrases for Opinion Mining Applications. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 7. ACM, 2012.

[49] Sven Rill, Dirk Reinel, Jörg Scheidt, and Roberto V Zicari. Politwi: Early Detection of Emerging Political Topics on Twitter and the Impact on Concept-Level Sentiment Analysis. *Knowledge-Based Systems*, 69:24–33, 2014.

[50] A Schiller, S Teufel, C Stöckert, and C Thielen. Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, Stuttgart, Germany: Institut für maschinelle Sprachverarbeitung, 1999.

[51] Nakatani Shuyo. Language Detection Library for Java, 2010. `http://code.google.com/p/language-detection/`, zuletzt abgerufen am 17.12.2017.

[52] Philip J Stone, Dexter C Dunphy, and Marshall S Smith. The General Inquirer: A Computer Approach to Content Analysis. *Computing and Computers*, 1966.

[53] Carlo Strapparava, Alessandro Valitutti, et al. WordNet Affect: an Affective Extension of WordNet. In *LREC*, volume 4, pages 1083–1086, 2004.

[54] Hiroya Takamura, Takashi Inui, and Manabu Okumura. Extracting Semantic Orientations of Words Using Spin Model. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 133–140. Association for Computational Linguistics, 2005.

[55] Peter D Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

[56] Peter D Turney and Michael L Littman. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems (TOIS)*, 21 (4):315–346, 2003.

[57] University College London UCL. The International Corpus of English, 2016. `http://www.ucl.ac.uk/english-usage/projects/ice.htm`, zuletzt abgerufen am 25.10.2017.

[58] Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. The viability of web-derived polarity lexicons. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 777–785. Association for Computational Linguistics, 2010.

[59] Ulli Waltinger. Polarity Reinforcement: Sentiment Polarity Identification By Means Of Social Semantics. In *AFRICON'09*, pages 1–6. IEEE, 2009.

[60] Ulli Waltinger. GermanPolarityClues: A Lexical Resource for German Sentiment Analysis. In *LREC*, 2010.

[61] Ulli Waltinger. Sentiment Analysis Reloaded – A Comparative Study on Sentiment Polarity Identification Combining Machine Learning and Subjectivity Features. In *WEBIST (1)*, pages 203–210, 2010.

[62] Janyce Wiebe. Learning Subjective Adjectives from Corpora. *AAAI/IAAI*, 2000.

[63] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.

[64] Florian Wogenstein, Johannes Drescher, Dirk Reinel, Sven Rill, and Jörg Scheidt. Evaluation of an Algorithm for Aspect-Based Opinion Mining Using a Lexicon-Based Approach. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 5. ACM, 2013.

University
of Bamberg
Press

Die ständig zunehmende Anzahl nutzergenerierter textueller Daten im World Wide Web, wie beispielsweise Kundenrezensionen, und die Begehrlichkeiten diese Daten hinsichtlich der darin ausgedrückten Meinungen zu Unternehmen, Produkten, Dienstleistungen etc. maschinell auszuwerten, erfordern gut funktionierende, angepasste Methoden des Opinion Mining. Die Grundlage für viele dieser Methoden bilden lexikalische Ressourcen in Form von Lexika meinungstragender Wörter und Phrasen. Diese Lexika existieren bisher allerdings nur für ausgewählte Sprachen, haben diverse inhaltliche Lücken, und sind automatisch (für verschiedene Sprachen) nur mit großem Aufwand zu erzeugen.

In dieser Arbeit wird deshalb ein neues Verfahren vorgestellt, das dazu beitragen soll, die benannten Probleme – durch den Einsatz statistischer Methoden – zu überwinden. Zudem wurde, mittels dieses Verfahrens, der Prototyp eines neuen Lexikons mit meinungstragenden Wörtern und Phrasen für die deutsche Sprache generiert und anschließend evaluiert. Dafür wurde im Rahmen eines Experiments mit 20 Teilnehmern ein Basis-Referenzlexikon für die deutsche Sprache manuell erzeugt.

Klassische Einsatzgebiete der Opinion Mining Algorithmen und Ressourcen, und damit des vorgestellten Verfahrens, sind Systeme zur Erfassung von Kundenmeinungen zu verschiedenen Unternehmensbereichen zur Unterstützung des Beschwerde- und Reputationsmanagements. Allerdings sind die Möglichkeiten des neu entwickelten Verfahrens nicht auf diese klassischen Anwendungsfelder begrenzt. Auch der interdisziplinäre Einsatz, z.B. zur Untersuchung von Sprachvarietäten im Forschungsfeld der Sprachstatistik, ist denkbar.

ISBN: 978-3-86309-594-9
9 783863 095949
www.uni-bamberg.de/ubp