TECHNISCHE UNIVERSITÄT DRESDEN

Diplomarbeit

Extraktion und Klassifikation von bewerteten Produktfeatures auf Webseiten

bearbeitet von

Michéle Sprejz

Geboren am 06.07.1986 in Bad Muskau

Fakultät Informatik
Institut für Systemarchitektur
Lehrstuhl Rechnernetze

Professor:
Prof. Dr. rer. nat. habil. Dr. h. c. Alexander Schill

Betreuer:
Dipl.-Medien-Inf. David Urbansky
Dipl.-Wirt.-Inf. Christopher Schulz
Dipl.-Wirt.-Inf. Richard Raue

Aufgabenstellung

(diese Seite wird im Druck durch das Original ersetzt)

Erklärung

Hiermit erkläre ich, Michéle Sprejz, die vorliegende Diplomarbeit zum Thema

„Extraktion und Klassifikation von bewerteten Produktfeatures auf Webseiten“

selbstständig und ausschließlich unter Verwendung der im Quellenverzeichnis aufgeführten Literatur- und sonstigen Informationsquellen verfasst zu haben.

Dresden, am 31. Dezember 2011

Danksagung

Während der Bearbeitungszeit der vorliegenden Arbeit erhielt ich von vielen Menschen Unterstützung, dafür möchte ich mich hiermit bedanken. David Urbansky gilt besonderer Dank für die außerordentliche Betreuung während der Bearbeitung des Großen Beleges und des Diploms. Hervorheben möchte ich, dass er es sehr gut versteht Menschen zu motivieren und mit konstruktiver Kritik immer wieder voranzutreiben. Die Erstellung beider Arbeiten wäre ohne ihn für mich nie derart lehrreich und erfüllend gewesen.

Diese Arbeit ist in Kooperation mit der T-Systems Multimedia Solutions GmbH entstanden, was durch meine beiden dortigen Betreuer Christopher Schulz und Richard Raue möglich war und wofür ich ihnen sehr dankbar bin. Für mich war es eine tolle Erfahrung die Strukturen und Abläufe bei der Entwicklung eines großen Projektes, wie der SemaSuite kennenzulernen. Bedanken möchte ich mich für die zu jeder Zeit vorhandene Unterstützung von Christopher Schulz bei jeder Art von Fragen oder Problemen. Ebenso für die inhaltlichen Anregungen und Anmerkungen die er während dem letzten halben Jahr gemacht hat.

Meinen Eltern möchte ich außerordentlich großen Dank aussprechen. Sie haben mir das Studium ermöglicht und mich in jeglicher Art und Weise immer wieder unterstützt. Ihnen und meinem Bruder danke ich dafür, dass sie immer an mich geglaubt haben. Ein ganz besonderer Dank gilt auch meinem Freund, Christian Langenhan, der mir während der gesamten Bearbeitungszeit eine wundervolle Unterstützung war. Ich bin dankbar für die vielen Gespräche in denen er mich immer wieder auf neue Ideen brachte und mich ständig motivierte sowie für seine nicht überstrapazierbare Geduld mit mir. Sehr großen Dank möchte ich an dieser Stelle auch an die vielen großartigen Menschen, die mir in den letzten Jahren begegnet sind richten, besonders an meine Kommilitonen Johannes Apel, Janine Hellriegel und Timm Fredebold. Ohne sie wäre das Studium nicht zu solch einer schönen Zeit geworden und vielleicht nicht einmal schaff bar gewesen. Danke für die ständige Motivation und das gemeinsame Bewältigen der Herausforderungen währenddessen. Einen großen Dank auch an meine fleißigen Korrekturleser, die diese Arbeit besser gemacht haben.

Dresden, im Dezember 2011 *Michéle Sprejz*

Kurzfassung

Nirgendwo sonst als im Internet können positive und negative Erfahrungen schneller und weitläufiger ausgetauscht werden. Für heutige Unternehmen wird somit immer wichtiger vorhandene Meinungen von Kunden zu kennen, zu analysieren und zu überwachen, um schnellstmöglich auf einen negativen Ruf reagieren zu können. Da nahezu jeder Mensch in der Lage ist, seine Meinungen online auszudrücken – und dies auch tut – entsteht eine fast unendliche Masse an Daten, welche analysiert werden muss. Hier ist das manuelle Durchsuchen nicht mehr ausreichend und automatisierte Methoden werden benötigt. In der vorliegenden Arbeit wird ein solches Verfahren auf Eignung für die Klassifikation von Sätzen hinsichtlich ihrer Polarität überprüft. Ein Klassifikator aus dem Bereich des maschinellen Lernens wird mit unterschiedlichen Datensätzen und Einstellungen getestet und anschließend prototypisch in ein bestehendes Informationsextraktions-System integriert.

Abkürzungsverzeichnis

POS	Part of Speech (zu dt.: Wortart)
NER	Named Entity Recognition (zu dt.: Eigennamenerkennung)
SVM	Support Vector Machine (zu dt.: Stützvektormaschine)
SO	Semantische Orientierung
TDF	Term document frequency (zu dt.: Vorkommenshäufigkeit)
IDF	Inverse document frecquency (zu dt.: inverse Dokumenthäufigkeit)
FAQ	Frequently asked questions (zu dt.: häufig gestellte Fragen)
TU	Technische Universität
SentiWs	Sentiment Wortschatz
P	Positiv
N	Negativ
O	Objektiv

Inhaltsverzeichnis

1 Einleitung

Das Internet beeinflusst unser Leben in der heutigen Zeit maßgeblich. Es gibt kaum noch einen Menschen, der wichtige Entscheidungen trifft, ohne sich vorher online über die vorhandenen Möglichkeiten und jeweiligen Meinungen dazu zu informieren. Sei es der Einstieg in eine neue Firma, vor dem ausführlich über den Umgang mit den Mitarbeitern recherchiert wird oder der Umzug in eine neue Stadt, vor dem Foren durchgeforstet werden, um Meinungen der Ansässigen über die Gegend zu erhalten. Auch die Anschaffung eines neuen technischen Gerätes wird kaum noch durchgeführt, ohne vorheriges Lesen der Erfahrungsberichte anderer Nutzer. Das Internet wird zudem nicht mehr nur für die Informationsbeschaffung, sondern mittlerweile auch zur Informationsverbreitung verwendet. Nahezu jeder Mensch ist in der heutigen Zeit in der Lage, sich über diverse Kanäle im Internet auszudrücken. Somit entsteht eine große Masse an Daten, welche manuell kaum noch zu überblicken ist. In vielen dieser Daten stecken jedoch nützliche Informationen für die unterschiedlichsten Parteien, sei es für den Privatnutzer oder für kommerzielle und nichtkommerzielle Unternehmungen. Für die Zukunft ist es sinnvoll, wenn nicht sogar unumgänglich, das manuelle durch maschinelles Recherchieren zu ersetzen.

1.1 Motivation

Da mittels des Internets der Zugang zu scheinbar unendlich vielen Meinungen über ebenso unendlich vielen Themen gelegt ist, werden für diese Masse an Daten automatische Analysetechniken benötigt. Nicht nur die große Menge der Daten ist ein Grund für die Notwendigkeit maschineller Techniken. Auch ist es für einen Menschen nicht immer möglich, objektiv Meinungen zu klassifizieren. Häufig sind sich verschiedene Personen nicht einig darüber, ob eine Aussage eher positiver oder negativer Natur ist. Eventuell können Maschinen hier eine konsistentere Klassifikation bieten. Einer Maschine können dafür Regeln beigebracht werden oder sie können selbst anhand von Beispielen Muster erlernen. Letzteres bezieht sich auf den Ansatz des maschinellen Lernens. Wenn ein Computer anhand von Beispielsätzen zukünftig neue Sätze automatisch in die Kategorien positiv und negativ einordnen könnte, wie hilfreich wäre dies beispielsweise für einen Produktersteller? Er könnte sich geäußerte Meinungen zu dem neu auf dem Markt erschienenen Produkt automatisch zusammenfassen lassen. Wie viele Nutzer könnten vor der Unzufriedenheit mit fehlerhaften Produkten bewahrt werden, wenn schnell erkennbar ist, dass diese Geräte eventuell Produktionsfehler haben? Und wie viele Hersteller könnten rechtzeitig auf diesen Mangel ihres Produktes reagieren? Ausgehend von diesen Fragestellungen wird in der vorliegenden Arbeit die automatische Analyse von Meinungen betrachtet.

1.2 Ziel der Arbeit

Der Fokus dieser Arbeit liegt auf dem automatischen Klassifizieren von aus dem Internet gewonnenen Meinungen bezüglich ihrer Orientierung. Das heißt, es ist das Ziel automatisch eine große Menge an Meinungen positiven, negativen oder gegebenenfalls objektiven Kategorien zuzuordnen. Diese semantische Analyse von Daten wird als Opinion Mining (zu dt.: „Stimmungsanalyse“) bezeichnet. Die Verfahren aus diesem Bereich sind vorrangig für unstrukturierte Daten konzipiert. Dies ist auch häufig die Form der aus dem Internet extrahierten Meinungen. Die Sprache, in der diese Klassifizierung umgesetzt wird, ist Deutsch. Der Großteil der Forschungen betrachtet englisch-sprachige Dokumente, die deutsche Sprache ist hinsichtlich des Opinion Mining relativ unerforscht.

Der spezielle Anwendungsfall für das Opinion Mining in dieser Arbeit ist, dass ein Hersteller die Meinungen über seine Produkte auf dem Markt automatisch analysiert haben möchte. Für das Unternehmen ist es essentiell zu erfahren auf

welche Themen sich diese Äußerungen beziehen und ob sie sich auf das gesamte Produkt oder nur auf einzelne Eigenschaften beziehen. Außerdem liegt die Priorität des Herstellers höchstwahrscheinlich auf den negativen Meinungen. Sind diese bekannt, können gezielter Verbesserungen an den Produkten durchgeführt und somit Kunden gleichzeitig zufriedener gestellt werden.

Die Grundlage für das Bilden von Problemklassen sind Rezensionen von Nutzern zu zufällig gewählten Produkten. Diese sind im Web u. a. auf diversen Social Media Kanälen zu finden.

Für das Erkennen und die Extraktion von Meinungen bezüglich verschiedener Entitäten[1] und deren Aspekten sind unterschiedliche Vorgehensweisen möglich, diese können wiederum in mehrere Teilschritte zerlegt werden. Die einzelnen Schritte sind das Klassifizieren von Aussagen bezüglich ihrer Polarität sowie das Erkennen, von welchen Produkten bzw. Eigenschaften diese Meinungen handeln. Nähere Erläuterungen zu den einzelnen Teilschritten aus denen das Opinion Mining besteht sind in Abschnitt 2.2.2 zu finden. Das Erkennen der Polarität ist eine Problemstellung des Opinion Mining, wohingegen die Extraktion der Entitäten, auf welche sich die Meinungen beziehen, ihre Ursprünge in der Informationsextraktion hat. Letzteres kann also mit Hilfe von Methoden aus diesem Bereich umgesetzt werden. Für den Praxisteil dieser Arbeit ergeben sich somit drei Möglichkeiten für die Reihenfolge und die Inhalte dieser zwei Teilschritte:

I. Zu Beginn wird das gewählte Produkt durch eine Ontologie-ähnliche Struktur dargestellt. Ziel dabei ist es, die vorhandenen Aspekte abzubilden und anschließend Sätze auf deren Vorkommen zu untersuchen. Nach der Sammlung aller Sätze, welche die bekannten Aspekte enthalten, werden diese in die Kategorien positiv, objektiv und negativ eingeordnet. Dies geschieht entweder mit Hilfe eines Wörterbuches oder eines maschinellen Klassifikators.

II. Vorhandenen Kundenbewertungen werden hinsichtlich des gemeinsamen Vorkommens von Nomen und Adjektiven untersucht, aus den häufig auftretenden wird ein semantisches Netz gebildet. Es wird davon ausgegangen, dass Nomen häufig Aspekte beschreiben und dass Adjektive eine Polarität ausdrücken. Welche Polarität beschrieben wird, kann anschließend mit Hilfe eines speziellen Wörterbuches identifiziert werden.

[1] In dieser Arbeit sind die Entitäten die Produkte aus den Kategorien Handy und Notebook.

III. Es werden im ersten Schritt Sätze in gewählte Kategorien eingeordnet. Hierfür wird ein Goldstandard, mit dessen Hilfe ein maschineller Klassifikator trainiert wird, erstellt. Nachdem alle vorhandenen Dokumente klassifiziert sind, erfolgt die Extraktion der beinhalteten Themen der Sätze.

Die erste Variante des Vorgehens erscheint als nicht allgemein anwendbar, da hier vorher spezifisch zu einem Produkt Informationen als Ontologie dargestellt werden müssen. Das heißt, dass es noch vor dem Klassifizieren der Meinungen wichtig ist zu wissen, aus welchen Aspekten das Objekt besteht. Die zweite Variante wird ebenso verworfen, da hier für die Suche nach Schlüsselwörtern, bezüglich der positiven und negativen Aussagen, ein Wörterbuch angewendet werden muss. Dieses Verfahren wurde in der Vergangenheit bereits mehrfach erforscht und als nicht ausreichend bewertet.

In dieser Arbeit wird die Aufgabe mit dem dritten Ansatz gelöst, welcher allgemeingültiger ist. Es sollen erst Sätze in positive, negative und gegebenenfalls objektive Kategorien eingeordnet werden, was unabhängig von den Themen der Sätze geschieht. Interessant erscheint die Frage, ob mit Hilfe von maschinellen Lernverfahren, ohne jegliche Vorverarbeitung und Regelbildung, Sätze hinsichtlich ihrer Polaritäten klassifiziert werden können. Die Herangehensweise mit Hilfe eines Wörterbuches oder auch das Bilden von Regeln zur Extraktion von Meinungen wird besonders in der englischen Sprache oft in wissenschaftlichen Arbeiten umgesetzt und erbrachte gute, jedoch nicht vollkommen befriedigende Ergebnisse.[2] Eventuell kann ein Klassifikator Muster und Strukturen in Sätzen erkennen, welche nicht durch grammatikalische Regeln oder einzelne Stichworte erfasst werden können. Ein Goldstandard soll erstellt werden und als Grundlage für den Klassifikator dienen. Um anschließend aus den klassifizierten Sätzen Problemklassen bilden zu können, wird versucht mit Hilfe von Informationsextraktions-Methoden die Produkte und deren Aspekte zu erkennen. Zusätzlich kann mittels dieser Methoden eine genauere, nicht nur satzweise, sondern auf der Ebene von Wortgruppen erfolgende Zuordnung von Meinungen zu den einzelnen Aspekten erfolgen. In Abbildung 1 ist der Ablauf dargestellt.

[2] In Abschnitt 2.2.2 werden diese Herangehensweisen ausführlich erläutert.

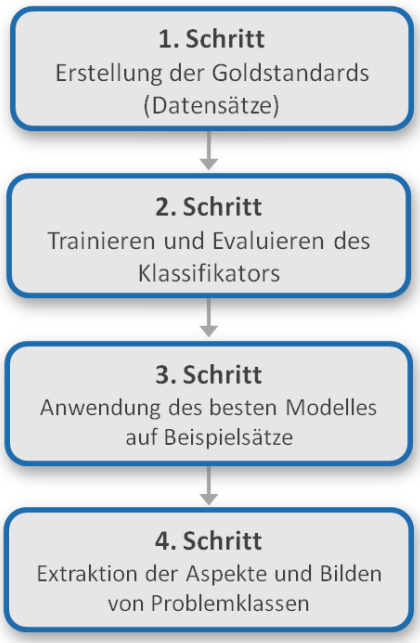

Abbildung 1: Ablauf der Herangehensweise dieser Arbeit an die Problemstellung des Opinion Mining

1.3 Forschungsfragen

Kann durch das satzweise Einordnen von Bewertungen in die Kategorien positiv, negativ und neutral das Bilden von Problemklassen bezüglich eines Produktes unterstützt werden?

Für Unternehmen ist es nicht nur wichtig zu wissen, wie gut ein Produkt o. ä. von den Nutzern angenommen wird, sondern auch welche Kritikpunkte oder Fehler diesbezüglich geäußert werden. Es ist essentiell zu wissen, welche Probleme bestehen, um Verbesserungen der Qualität umzusetzen und gezielte auf Kundenwünsche eingehen zu können. Es ergibt sich die Frage, ob mit Hilfe von Opinion Mining-Verfahren automatische Zusammenfassungen über Produkte in deutscher Sprache hinsichtlich ihrer Mängel erstellt werden können.

Können mit Hilfe von automatischer Textklassifikation Sätze ohne spezielle Vorverarbeitung in positive, neutrale und negative Kategorien eingeordnet werden?

Der größte Teil der bisherigen Forschung hinsichtlich der Kategorisierung von Dokumenten nach ihrer semantischen Bedeutung wurde mit Hilfe von Wörterbüchern umgesetzt. Diese Wörterbücher enthalten Schlüsselwörter, welche direkt oder auch indirekt eine bestimmte Emotion oder Bedeutung wiedergeben. Nach

diesen wird in den zu klassifizierenden Dokumenten gesucht. Bei dieser Methode werden aber häufig weitere Indikatoren für eine beinhaltende Meinung übersehen. Andere Herangehensweisen der Untersuchungen auf semantische Bedeutungen werden mit Hilfe von sogenannten Mustern durchgeführt. Hier werden Regeln, die beschreiben in welcher Form Meinungen auftreten, verwendet. Ein Beispiel dafür ist, dass negative Adjektive vor einem Nomen eine negative Meinung ausdrücken. Generelle Probleme bergen die Erkennung von Ironie oder Sarkasmus. Ebenso wenig fällt der Konjunktiv[3] selten in den Hauptfokus der Forschungen. Es müssten alle wichtigen Formen in einem Wörterbuch aufgenommen werden (Beispiele hierfür sind: „hätte“, „wäre“, usw.). Dies ist jedoch mit einem enorm hohen manuellen Aufwand verbunden. Der Konjunktiv stellte sich jedoch in vergangenen Arbeiten als wichtiger Indikator für negative Meinungen heraus.

Es stellt sich daher die Frage, ob es eine Variante der Herangehensweise gibt, welche ohne ein manuell erstelltes Wörterbuch und ohne tiefgründige Vorverarbeitung der Eingabedokumente, Erfolg bringend ist. Eventuell können überwachte, maschinelle Lernmethoden dafür eingesetzt werden. Es wird in dieser Arbeit überprüft, ob ein Klassifikator in einer Trainingsmenge aus positiven, negativen und objektiven Sätzen Regeln und Muster in diesen erkennt und damit neue Sätze ebenso in diese Kategorien einordnen kann.

Welche Methode - die regelbasierte oder die automatisierte - arbeitet hinsichtlich des Einordnens von Dokumenten in Polaritätskategorien mit größerer Genauigkeit und Zuverlässigkeit?

Hinsichtlich Genauigkeit und Trefferquote ist es möglich, eine wörterbuchbasierte Variante mit einem Klassifikator aus dem Bereich des maschinellen Lernens zu vergleichen. Es ist außerdem erstrebenswert, ein eventuell vorhandenes Potential zur Kopplung beider Methoden zu entdecken. Bei dem wörterbuchbasierten Ansatz werden meist Regeln manuell erstellt. Eine Regel ist beispielsweise Adjektive für die Klassifizierung von Texten zu verwenden. Zusätzlich wird ein Wörterbuch genutzt. In diesem sind meinungsbehaftete Wörter mit einer Gewichtung hinsichtlich ihrer Polarität aufgelistet. Es soll herausgefunden werden, ob Meinungen besser anhand solcher Schlüsselwörter oder anhand von maschinell erlernten Mustern eingeordnet werden können.

[3] Häufig werden negative Meinungen, beispielsweise hinsichtlich eines Produktes, im Konjunktiv geäußert: „Es wäre schön gewesen, wenn das Handy auch noch einen Mp3-Player gehabt hätte.“. Vgl. (Sprejz 2011)

Ist ein erstelltes Modell eines Textklassifikators für eine Domäne, auch erfolgreich anwendbar auf eine weitere Domäne?

Wenn es möglich ist, Sätze mit Hilfe eines maschinellen Klassifikators nach ihrer Polarität zu ordnen, stellt sich anschließend die Frage, wie domänenabhängig dieses Verfahren ist. Schlüsselwörterbücher sind meist domänenunabhängig. Es wird also evaluiert, ob ein erstelltes Modell aus einer Trainingsmenge einer Domäne auch auf eine andere erfolgreich angewendet werden kann.

Können mit Hilfe von Methoden des Text-Mining und der Informationsextraktion aus den bereits kategorisierten Dokumenten Problemklassen eines Produktes gebildet werden?

Um einen Nutzen aus der Kategorisierung von Dokumenten nach Meinungen zu gewinnen, ist es nicht nur notwendig diese korrekt in positive, negative und objektive Kategorien einzuordnen, sondern gleichzeitig herauszufinden, was die Auslöser für die jeweiligen Meinungsintensitäten sind. Bei Produktbewertungen ist es beispielsweise von Vorteil zu wissen, welche Aspekte für besonders gut oder schlecht befunden wurden. Es soll überprüft werden, ob es möglich ist, Methoden aus den Bereichen Text-Mining und Computerlinguistik anzuwenden, um oft negativ bzw. positiv erwähnte Features zu identifizieren.

1.4 Struktur des Dokuments

In Kapitel 2 wird das Opinion Mining vorgestellt und in Fachbereiche eingeordnet. Die Problemstellungen werden erläutert und vorhandene Lösungsansätze in der Literatur werden gegeben.

Das Kapitel 3 beschreibt, wie die Datensätze, welche die Grundlage für die Umsetzung der hier vorliegenden Aufgabenstellung sind, erstellt wurden und welche Besonderheiten bei der Extraktion von Rezensionen aus dem Internet aufgetreten sind.

Kapitel 4 beschreibt die Vorgehensweise, welche genutzt wurde, um zu überprüfen ob maschinelle Lernverfahren für die Klassifikation von Sätzen hinsichtlich ihrer Polarität geeignet sind. Außerdem erfolgt hier ein Vergleich mit einem weiteren Ansatz zur Erkennung von Polaritäten.

Die prototypische Umsetzung der Aufgabenstellung wird in Kapitel 5 beschrieben. Und in Kapitel 6 erfolgt die Zusammenfassung sowie der Ausblick auf zukünftige Forschungen.

2 Grundlagen

In diesem Kapitel wird ein Überblick über den Bereich des Opinion Mining gegeben. Im ersten Abschnitt werden zunächst grundlegende Definitionen erläutert. Im Abschnitt 2.2 erfolgt das Einordnen des Opinion Mining in einzelne Fachbereiche. Anschließend werden Problemstellungen erläutert bzw. bisherige Lösungsansätze dargestellt. Im Abschnitt 2.3 werden angewendete Verfahren aus anderen Bereichen sowie speziell für das Klassifizieren von Dokumenten hinsichtlich ihrer Polarität neu erstellte Methoden beschrieben. Abschließend werden im Abschnitt 2.4 aktuelle Forschungsarbeiten zu diesem Thema vorgestellt.

2.1 Grundlegende Definitionen

Im Allgemeinen werden die Bezeichnungen Daten, Wissen und Informationen mit abweichenden Bedeutungen verwendet. Insbesondere der Begriff Information wird in der Literatur auf verschiedene Weisen definiert. (Ferber 2003, 26)

Daten sind eine Folge von Einheiten, welche eine syntaktische Struktur besitzen. Wenn zusätzlich die Bedeutungen dieser vorhanden sind, entsteht eine semantische Bedeutung und damit Wissen. Dadurch entsteht eine Information für eine konkrete Situation, die mit diesem vorhandenen Wissen bewältigt werden kann. Wissen bedient also einen Informationsbedarf und wird dadurch zu einer Information. Es gibt andere Ansätze, bei denen die Information als Voraussetzung für

Wissen gesehen wird. Mit ihr wird „gearbeitet“ und sie kann für eine Problemlösung eingesetzt werden. (Ferber 2003, 27), (Heyer, Quasthoff und Wittig 2006, 8)

Wissen wird mittels Text dargestellt und in unterschiedlichsten Formen, wie Fachbücher, Lexika etc. angeboten. Um aus Texten Wissen zu generieren, müssen Relationen zwischen den Zeichenketten, aus denen Texte bestehen, gefunden und deren Semantik erkannt werden. Dies ist für Menschen ein ganz natürlicher Prozess, für eine Maschine jedoch eine große Herausforderung. (Heyer, Quasthoff und Wittig 2006, 8)

Damit Maschinen ähnlich „Denken“ bzw. Wissen erfassen können wie Menschen, beschäftigen sich verschiedene Forschungsbereichen mit der Lösung dieser Aufgabe. Bereiche wie die Computerlinguistik, das Text Mining oder auch das Opinion Mining haben im Laufe der Jahre eine große Bedeutung erlangt, da im Internet immer mehr Informationen zur Verfügung stehen und von Maschinen bearbeitet werden können bzw. auf Grund der Masse bearbeitet werden müssen.

Die Grundlage für das Entwickeln von Methoden für die maschinelle Verarbeitung von Dokumenten stellen linguistische Korpora dar. Im weitetesten Sinne handelt es sich dabei um maschinell lesbare, digitalisierte Sprachdaten (Sasaki und Witt 2004, 195). Allgemein werden sie u. a. für das Testen von computerlinguistischen Methoden oder auch für das Trainieren und Testen von maschinellen Klassifikatoren eingesetzt (Carstensen, et al. 2010, 482).

In dieser Arbeit wird mit Korpora aus Texten gearbeitet. Für viele Anwendungen ist es hilfreich, diese Texte in linguistische Einheiten zu unterteilen. Das können Wörter, Buchstaben, aber auch Sätze sein. Neben dem Text sind in solchen Korpora häufig auch Annotationen, also Zusatzinformationen zu finden, welche beispielsweise mittels POS-Tags[4] eingefügt werden. Dies ermöglicht zum Beispiel das Einteilen der Wörter nach ihren Wortarten. (Carstensen, et al. 2010, 483f.) Häufig werden solche Korpora auch als Goldstandard bezeichnet. Eingesetzt werden sie u. a. für das Messen der Qualität von Algorithmen der Computerlinguistik. (Carstensen, et al. 2010, 489)

[4] POS-Tags sind Markierungen, welche die Wortart eines Wortes ausdrücken.

2.2 Opinion Mining

2.2.1 Einordnung in Fachbereiche

Opinion Mining befasst sich mit der Analyse von unstrukturierten, textuellen Daten hinsichtlich auftretender Meinungen, Einstellungen, Gefühle, etc. gegenüber einem Thema (Liu 2011, 6). Häufig wird es auch als „Sentiment Detection“ (Tang, Tan und Cheng 2009, 10760) oder zu dt. „Stimmungsanalyse“ bezeichnet.[5]

Opinion Mining vereint viele unterschiedliche Fachbereiche, weil eine Vielzahl von Verfahren aus dem Bereich der Computerlinguistik, dem Information Retrieval, der Informationsextraktion und auch aus dem Bereich des Text Mining angewendet werden (Tang, Tan und Cheng 2009, 10760). Da viele Fachbereiche in das Feld des Opinion Mining einfließen, soll im Folgenden eine kurze Beschreibung dieser gegeben werden.

Die Zuordnung zu einem Anwendungsgebiet ist in der Literatur nicht eindeutig zu finden. Häufig wird es in das Gebiet des Web Content Mining eingeordnet. Dies ist eine Unterform des Web Mining, welches wiederum unter anderem ein Anwendungsgebiet des Data Mining ist (Liu 2011, 6f.). Auch das Text Mining wird oft als Fachgebiet des Opinion Mining bezeichnet. In der Literatur wird das Web Mining auch als Anwendungsfeld des Text Mining angesehen (Mehler 2004, 348) sowie als eine eigenständige Spezialisierung des Data Mining, neben dem Text Mining (Liu 2011, 6) beschrieben. Das Web Mining wird in drei Bereiche unterteilt - in das Untersuchen der Struktur (bzw. der Verlinkung) des Internets, der Analyse des Benutzerverhaltens sowie die Analyse der Inhalte des Webs. In Letzteres wird das Opinion Mining eingeordnet. Viele Methoden aus dem Bereich des Data Mining finden hier Anwendung. Diese müssen jedoch angepasst bzw. neu erfunden werden, da unstrukturierte Daten analysiert werden. (Liu 2011, 7)

Da das Opinion Mining sich ebenfalls mit der Suche nach Informationen in Dokumenten beschäftigt, wird es auch als eine Unterform des Data Mining bezeichnet. Letzteres umfasst die Suche nach vorhandenen Regelmäßigkeiten in bestehenden, strukturierten Datenmengen. (Ferber 2003, 14).

Es gibt mittlerweile speziellere Bereiche mit angepassten Verfahren für unstrukturierte Daten, wie das Web Mining oder auch das Text Mining. Verfahren, welche hier angewendet werden, kommen aus dem Bereich des maschinellen Lernens.

[5] Für eine detaillierte Übersicht über die Begriffsentstehung siehe Pang und Lee 2008, 8ff.

Dies sind die Klassifikation und das Clustering sowie die Analyse von Assoziationen. (Liu 2011, 6)

Text Mining wird entweder als Nebenbereich des Opinion Mining oder auch als übergeordneter Bereich beschrieben. Es ist eine Kombination von Methoden aus dem Bereichen Data Mining und Information Retrieval (Ferber 2003, 19).

Da die Masse an unstrukturierten Informationen im Internet immer größer wird und es nicht ausreichend ist „nur" Informationen zu finden und diese nach Relevanz zu ordnen, wurden geeignetere Methoden gesucht und im Text Mining gefunden (Mehler 2004, 329) Dabei sollen Zusammenhänge gewonnen, extrahiert und somit meist große Mengen an Texten strukturiert werden. Häufig werden Verfahren, wie das maschinelle Lernen, angewendet, welche auf der Basis von Statistiken und Mustern arbeiten. (Heyer, Quasthoff und Wittig 2006, 1ff.)

Da sich das Opinion Mining mit der semantischen Analyse von Texten und somit mit der Verarbeitung natürlicher Sprache beschäftigt kommt es auch vor, dass es dem Fachbereich der Computerlinguistik, welcher sich der algorithmischen Verarbeitung natürlicher Sprache widmet, zugeschrieben wird (Ensuli und Sebastiani 2006, 1). Es werden beispielsweise Struktur und Syntax von Texten analysiert, aber auch Wortformen und deren Beugungen verarbeitet. Verfahren wie die maschinelle Übersetzung von Texten können somit umgesetzt werden. (Carstensen, et al. 2010, 1)

Da das Einordnen von Dokumenten in positive, negative und neutrale Kategorien ein Klassifizierungsproblem ist, kann das Opinion Mining auch der Textklassifikation zugeordnet werden. (Manning, Raghavan und Schütze 2009, 254). Die Textklassifikation ist wiederum dem Bereich der Informationsextraktion zugeordnet ist. Auch hier werden textuelle Daten in einzelne Kategorien eingeordnet (Xia, Zong und Li 2010, 1138).

Unter Textklassifizierung wird die Einordnung textueller Daten bzw. Dokumente in Kategorien verstanden. Die meisten Forschungen beschäftigten sich bisher mit der traditionellen Klassifikation nach Themen der Dokumente. Mittlerweile sind weitere Klassifikationsarten, wie beispielsweise die Klassifikation nach einem Genre oder einem Schreibstil entstanden. (Ensuli und Sebastiani 2005, 1)

Für die Einordnung in Stimmungen werden oft zwei Kategorien genutzt, und zwar positiv und negativ. Es wurde jedoch gezeigt, dass es von Vorteil sein kann, eine

weitere „neutrale“ Kategorie, wodurch bessere Ergebnisse beim Kategorisieren in positive und negative Dokumente erzielt werden. (Koppel und Schler 2005, 15)

In der Literatur wird die neutrale Kategorie jedoch unterschiedlich verwendet. Häufig steht sie für das Fehlen einer Meinung, also einen Fakt. Jedoch gibt es auch Ansätze, die diese Kategorie dafür benutzen, dass eine Meinung genau zwischen positiv und negativ einzuordnen ist, was wiederum auch eine Meinung ausdrückt. (Pang und Lee 2008, 27)

Es gibt drei Ansätze für die Textklassifikation. Kategorien und deren Eigenschaften können durch sogenanntes Expertenwissen manuell erstellt oder auch mittels Regeln und Mustern definiert werden. Diese beiden Ansätze sind jedoch sehr umständlich und zeitaufwändig. Ein dritter Ansatz erfolgt über statistische Methoden, das heißt, die Umsetzung wird mit maschinellen Lernverfahren[6]vollzogen. Dieser Ansatz soll auch in der vorliegenden Arbeit angewendet werden. (Manning, Raghavan und Schütze 2009, 255)

Um Dokumente hinsichtlich ihrer Polarität klassifizieren zu können, müssen sie zunächst aus dem Internet extrahiert werden. Diese Aufgabenstellung gehört in den Bereich des Information Retrieval. Das Gebiet beschäftigt sich mit dem Auffinden gewünschter, bereits bestehender Informationen in vorhandenen Datensammlungen. Es unterstützt die Informationsgewinnung von Nutzern. Eine weitere Hauptaufgabe ist die maschinengerechte Darstellung der menschlichen Anfrage sowie die Umwandlung der Ergebnisse in eine für Menschen lesbare Form. (Ferber 2003, 21)

Eine weitere, wichtige Aufgabe des Information Retrieval ist das schnelle Finden relevanter Informationen in großen Textquellen, auch Volltextsuche genannt (Carstensen, et al. 2010, 584f.). Das Information Retrieval wird gelegentlich auch als Anwendungsgebiet der Computerlinguistik bezeichnet (Yu und Hatzivassiloglou 2003, 1).

Da neben dem Erkennen von positiven und negativen Meinungen ebenso Inhalte dieser Meinungen extrahiert werden, ist das Opinion Mining dem Bereich der Informationsextraktion sehr ähnlich. Dieser Fachbereich beschäftigt sich mit der gezielten Suche nach Informationen in unstrukturierten Texten (Carstensen, et al. 2010). Es sollen vorkommende Objekte in Dokumenten und ihre Beziehungen zu

[6] Auf diese Verfahren wird in Abschnitt Verfahren aus dem Bereich Maschinelles Lernen näher eingegangen bzw. im zweiten Teil der Arbeit in Kapitel 5 angewendet.

einander erkannt werden. Die Textanalyse wird hier also tiefgründiger bzw. auf semantischer Ebene vollzogen. (Carstensen, et al. 2010, 594f.)

Auch in diesem Bereich finden häufig Methoden des maschinellen Lernens Anwendung. Ein Beispiel ist das automatische Erlernen von Mustern um aus vorhandenen Dokumenten Informationen zu extrahieren (Carstensen, et al. 2010, 596). Es wird, ebenso wie das Information Retrieval, als Anwendungsgebiet der Computerlinguistik betrachtet (Yu und Hatzivassiloglou 2003, 1).

2.2.2 Übersicht der Problemstellungen des Opinion Mining

Die Stimmungsanalyse kann in eine Vielzahl von Teilproblemen zerlegt werden. In den folgenden Abschnitten wird ein Überblick über die auftretenden Schwierigkeiten bei der Analyse von Meinungen gegeben werden. Es wird gezeigt, dass die Klassifikation nach Polaritäten nicht ebenso „simpel“, wie die Klassifikation nach Themen ist. Bei letzterer ist die Suche nach Schlüsselwörtern in einem Dokument ausreichend. Auch beim Opinion Mining ist es naheliegend, Meinungen anhand von Adjektiven, welche hier die Schlüsselwörter, nach denen es zu suchen gilt, sind, zu erkennen. Dass dies jedoch nicht ausreichend ist, wird im nachfolgend dargestellt. (Pang und Lee 2008, 16ff.)

Hauptprobleme sind die Subjektivitätsklassifikation, welches das Klassifizieren von textuellen Informationen in subjektive und objektive Dokumente beinhaltet; die anschließende Klassifikation hinsichtlich der Polarität, welche auf unterschiedlichen Ebenen eines Textes vollzogen werden kann[7]; sowie die Extraktion und Darstellung der gefundenen Meinungen. (Tang, Tan und Cheng 2009, 10760)

Eine weitere Herausforderung ist das Erkennen des Objektes, auf welches sich eine Meinung bezieht. Das Objekt wird in der Literatur als Entität bezeichnet. Dies kann eine Person, ein Produkt u. v. m. sein. Es besitzt Komponenten und Attribute, welche vereinfacht unter dem Begriff Aspekte zusammengefasst werden.[8] Da Aspekte und Entitäten nicht immer explizit von einem Autor erwähnt werden, sondern Meinungen darüber auch implizit durch die Verwendung von Vergleichen, Synonymen, Abkürzungen oder Koreferenzen[9] beschrieben werden können, wird die Erkennung und Extraktion dieser erschwert. Für unterschiedliche Anwendungsfälle ist es zudem möglich, dass es nicht nur von Bedeutung ist,

[7] Ebenen sind: Dokument, Satz, Wortgruppe und Wort
[8] In der Literatur wird auch häufig von Features gesprochen. (Vgl. Liu, 2010)
[9] Umschreibungen durch Pronomen

worüber eine Meinung geäußert wurde, sondern zusätzlich von welcher son[10]diese stammt. (Liu 2011, 460f.)

Meinungen können in unterschiedlichen Ausprägungen auftreten. Sie werden regulär über ein Objekt bzw. eine Entität geäußert oder auch indirekt gegeben, beispielsweise durch einen Vergleich. (Liu 2011, 463)

Die Analyse von Vergleichen zwischen einzelnen Objekten ist ein Teilgebiet des Opinion Mining. Die Problemstellung ist hier aufgrund der Art und Weise wie Subjektivität ausgedrückt wird eine andere. Vergleiche enthalten häufig nur indirekt subjektive Aussagen.[11] Zusätzlich gibt es nicht nur eine, sondern zwei Entitäten. Eine direkte Schlussfolgerung auf die Meinung gegenüber jeder Entität ist meist nicht möglich (Jindal und Liu 2006, 1)[12]

Eine Meinung kann eine Einstellung, ein Gefühl oder eine Beurteilung gegenüber einer Entität oder einem Aspekt einer Entität, mit der Orientierung positiv, negativ oder neutral sein (Liu 2010, 4). Da in einem Satz oder einem Dokument viele verschiedene Meinungen vorkommen können, ist es unter Umständen hilfreich, wenn nicht nur analysiert wird, ob diese vorhanden sind, sondern auch welche Ausprägung bzw. Stärke diese haben (Wilson, Wiebe und Hwa 2004, 1).

Einfache Wörter und Phrasen werden größtenteils für den Ausdruck einer Stimmung verwendet. Aber auch Fakten ohne ein bestimmtes meinungsbehaftetes Wort können wünschenswerte oder auch nicht wünschenswerte Zustände ausdrücken.[13] Das erwähnen der Quantität eines Aspekts kann ebenso negative oder positive Gefühle vermitteln.[14] Auch wenn ein Produkt bzw. eine Entität Ressourcen verbraucht oder produziert, können darüber Meinungen ausgedrückt werden.[15] Es lässt sich also festhalten, dass Meinungen nicht nur an einzelnen Wörtern festgemacht werden können. (Liu 2011, 483ff.)

Eine ideale Herangehensweise für die Meinungsanalyse wären folgende Teilschritte wie sie in Abbildung 2 abgebildet sind.

[10] Hierauf wird im Folgenden nicht eingegangen, da bei der Analyse von Produktbewertungen der Autor größtenteils bekannt ist. Weiteres hierzu in Bethard, et al. 2004; Choi, et al. 2005.
[11] Beispiel: „Auto x ist schneller als Auto y."
[12] In dieser Arbeit wird nicht auf die Analyse von Vergleichen eingegangen. Weiterführende Literatur: Jindal und Liu 2006, Ganapathibhotla und Liu 2008, Jindal und Liu 2006, Li, et al. 2010
[13] Beispiel: „Es ist kalt."
[14] Beispiel: „In deinem Zimmer liegt aber sehr viel Müll."
[15] Beispiel: „Der Akku wird schnell leer."

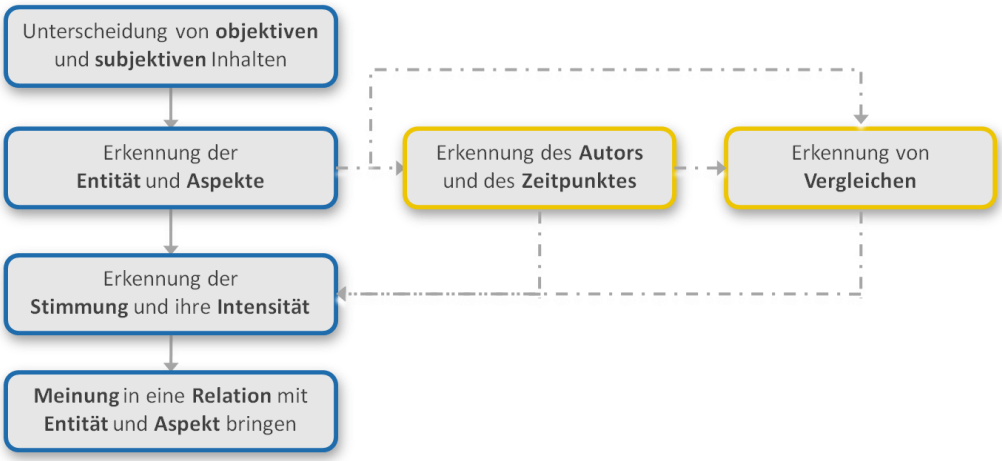

Abbildung 2: Einzelne Schritte des Opinion Mining

Der erste Schritt sollte das Unterscheiden zwischen subjektiven und objektiven Dokumenten darstellen. Anschließend könnte die Extraktion aller Entitäten mit ihren dazugehörigen Aspekten erfolgen. Im Folgenden Schritt sollte der Autor und die dazugehörige Zeit erfasst werden. Daraufhin müsste die Meinung hinsichtlich eines Aspektes einer Polarität zugeordnet werden und abschließend sollten diese in Relation zu den vorhandenen Entitäten und Aspekten gesetzt werden. Die Erkennung von Vergleichen kann ebenso als sinnvoller Teilschritt angesehen werden. Dieser wird allerdings nicht in allen Forschungsarbeiten mit einbezogen. In der Praxis sind nicht immer alle einzelnen Schritte für jeden Anwendungsfall notwendig. Hinzu kommt, dass die befriedigende Umsetzung einer jeden Teilaufgabe eine große Herausforderung ist. (Liu 2011, 465)

Eines der größten Probleme in diesem Gebiet ist es, zwischen Fakten und Meinungen zu unterscheiden. Dies ist nicht nur wichtig für den Bereich des Opinion Mining, sondern auch beispielsweise für Informationsextraktions-Systeme. Wenn hier gegebenenfalls Fakten als Wissen geliefert werden sollen, bringen subjektive Dokumente, wie einen Leserbrief, nicht die gewünschten Ergebnisse. (Yu und Hatzivassiloglou 2003, 1)

Nicht nur subjektive Sätze drücken Meinungen aus. Sie müssen auch nicht immer eine beinhalten (Liu 2011, 466). Objektive Sätze und auch objektive Wörter wiederum können ebenso indirekt Meinungen implizieren, auch wenn sie nicht eindeutig als subjektiv erkennbar sind (Zhang und Liu 2011, 2). Außerdem wurde festgestellt, dass das Unterscheiden zwischen Subjektivität und Objektivität schwieriger ist, als die anschließende Einordnung in die Polaritätskategorien (Mihalcea, Banea und Wiebe 2007, 977).

Das Finden von Subjektivität in Dokumenten ist lange ein wichtiger Vorverarbeitungsschritt der Polaritätsklassifikation gewesen und wird auch heute noch in einzelnen Forschungen verwendet. Dabei werden irrelevante Informationen eliminiert, damit diese die Kategorisierung nach Meinungen nicht verfälschen können. PANG und LEE zeigen in ihrer Arbeit, dass dadurch Reviews auf das Wesentliche beschränkt werden und hinsichtlich ihrer Polarität genauso aussagekräftig sind, wie das komplette Dokument. (Pang und Lee 2004, 7)

PANG und LEE stellen außerdem fest, dass es nicht ausreichend ist, eine Liste von Wörtern zu generieren, welche Indikatoren für subjektive Meinungen sind. Wie bereits erwähnt, ist das der zentrale Bestandteil der traditionellen Themenklassifizierung (Pang und Lee 2002, 5f.). Meinungen werden häufig sehr subtil und indirekt ausgedrückt, was nicht immer an einzelnen Schlüsselwörtern erkennbar ist (Pang und Lee 2008, 19). Hinzu kommt, dass viele Wörter in einer bestimmten Domäne eine bestimmte Polarität ausdrücken, in einer anderen aber genau das Gegenteil Wörterbuchbasierte Ansätze sind in diesem Fall nicht hilfreich (Zhang und Liu 2011, 1f.). Wörter können nicht nur in unterschiedlichen Domänen unterschiedliche Polaritäten ausdrücken, sondern sie können sogar abhängig von unterschiedlichen Aspekten eines Objektes sein (Ding, Liu und Yu 2008, 2).

Dennoch sind Wörter, welche eine Meinung vermitteln, eine besonders wichtige Grundlage für die Stimmungsanalyse. Doch das Finden dieser stellt sich als schwierig heraus. Nicht nur Adjektive können hierfür genutzt werden, auch Verben, Nomen und Phrasen kommen in Frage (Zhang und Liu 2011, 1). Gesucht sind Wörter mit einer sogenannten semantischen Orientierung, oder auch Polarität (Hatzivassiloglou und McKeown 1997, 1). Zusätzlich besitzt die semantische Orientierung der Wörter, welche entweder positiv oder negativ ist, einen Stärkegrad. Dieser sollte auch mit in die Analysen einbezogen werden. (Turney und Littman 2003, 4)

Eine manuell erstellte Liste an Schlüsselwörtern ist nicht ausreichend, da Menschen unterschiedliche Wörter als Indikatoren für ihre Meinungen empfinden. Eine maschinell erstellte Liste mit Hilfe des Zählens, wie häufig bestimmte Wörter in meinungsbehafteten Dokumenten vorkommen, lieferte bessere Ergebnisse. (Pang und Lee 2002, 3f.)

Für das Erstellen solcher Listen aus Wörtern, welche eine Meinung vermitteln, gibt es neben dem manuellen Ansatz noch zwei weitere. Der Wörterbuchbasierte Ansatz erfolgt iterativ mit Hilfe einer kleinen, manuell erstellten Startmenge aus

bekannten Wörtern. Diese erweitert sich automatisch durch Synonyme und Antonyme aus online verfügbaren Wörterbüchern[16] (Liu 2011, 478). Jedoch hat dieses Verfahren den Nachteil, dass nicht auf Domänenabhängigkeit geachtet und der Kontext nicht mit einbezogen wird (Zhang und Liu 2011, 1).

Ein anderer Ansatz wird als Korpus-basiert bezeichnet. Dieser arbeitet ebenso auf einer kleinen Startmenge an Wörtern, von denen die Polarität bekannt ist (Liu 2011, 478). Zusätzlich werden linguistische Bedingungen und die sogenannte Stimmungskonsistenz genutzt (Qiu, Liu, et al. 2009, 1). Unter anderem gehören Konjunktionen solchen sprachlichen Bedingungen an und können dafür verwendet werden, um beispielsweise weitere Adjektive und deren Polaritäten zu bestimmen, wenn von einer Startmenge ausgegangen wird (Hatzivassiloglou und McKeown 1997, 8).

Die Stimmungskonsistenz drückt aus, dass Meinungen einer Polarität meist hintereinander geäußert werden. Ein Wechsel wird häufig durch sogenannte Valence Shifter (zu dt.: „Meinungsänderer“), wie „aber“ oder „dennoch" eingeleitet. (Kanayama und Nasukawa 2006, 4) Dieser Ansatz kann auch auf Sätze übertragen werden.

Der korpusbasierte Ansatz kann auch ohne Stimmungskonsistenz mit Hilfe sogenannter POS-Tags und der Pointwise Mutual Information angewendet werden. Hierbei wird mit Hilfe einer Startmenge und einem großen Korpus berechnet, welche Wörter am engsten in Verbindung mit den bekannten Wörtern einer jeweiligen Polarität stehen. Daraus wird geschlussfolgert, dass sie ebenfalls dieser Polarität angehören müssen (Turney 2002). Letzterer Ansatz ist allein jedoch weniger effektiv als der wörterbuchbasierte Ansatz, da der Korpus sehr groß sein muss, um alle Möglichkeiten abzudecken (Liu, Web Data Mining 2011, 480).

Lexikalische Relationen sind eine weitere Methode zur Vergrößerung der Startmenge an Wörtern. Hier werden Distanzen zwischen zwei Wörtern mit Hilfe eines Wörterbuches berechnet, indem Beziehungen u. a. zwischen Synonymen analysiert werden. Es wird davon ausgegangen, dass zum Beispiel das Synonym eines positiven Wortes auch positiv sein muss. (Kamps, et al. 2004, 2ff.)

Ebenso können Definitionen und Bedeutungsbeschreibungen in online vorhandenen Wörterbüchern genutzt werden, um weitere Wörter für eine Polarität zu finden. Auch hier gilt die Annahme, dass beispielsweise positive Wörter ebenso

[16] Ein Beispiel wäre hierfür WordNet (Fellbaum 1998)

durch andere positive Wörter beschrieben werden. (Ensuli und Sebastiani 2005, 4ff.)

2.2.3 Ebenen der Durchführung des Opinion Mining

Die Klassifizierung von textuellen Informationen hinsichtlich ihrer Aussagenpolarität kann auf unterschiedlichen Ebenen erfolgen. Eine davon ist die sogenannte Dokument-Level-Stimmungs-Klassifikation. Hier wird ein ganzes Dokument einer Stimmungsrichtung zugeordnet. Eine Nutzerbewertung eines Produktes kann ein solches Dokument sein. Die Besonderheit bei Bewertungen ist, dass die Hypothese benutzt werden kann, dass das vollständige Dokument nur von einer Entität handelt und auch nur einen Autor besitzt. Somit ist es möglich ein Dokument dieser Art als Ganzes hinsichtlich einer Polarität zu klassifizieren. Jedoch ist dieser Umstand in der Realität eine Ausnahme. In Dokumenten aus Foren oder Blogs werden meist mehrere Entitäten von mehreren Autoren beschrieben. Die Analyse dieser Meinungen sollte jedoch ebenso betrachtet werden. (Liu 2011, 470f.)

Ein Problem, welches sich auf der Dokument-Ebene ergibt, ist das hier nicht wie bei der Themenklassifikation einzelne Schlüsselwörter für die Stimmungsanalyse ausreichend sind. Die Gesamtaussage eines Dokumentes ist häufig nicht die Summe der einzelnen Polaritäten der Teilabschnitte. (Pang und Lee 2008)

Dementsprechend werden viele Forschungen auf der Ebene von Sätzen umgesetzt. Diese detailliertere Betrachtung ist für viele Anwendungsfälle vorteilhafter. Auch hier wird wieder von der Annahme ausgegangen, dass in einem Satz nur eine einzelne Meinung von einem Autor enthalten ist. Dies ist in der Realität ebenso nicht immer, aber häufiger als auf der Ebene eines Dokumentes, gegeben (Liu 2011, 474f.). Sätze können auch eine Mischung aus Fakten und Meinungen darstellen. Mehrere Objekte können angesprochen werden und unterschiedliche Orientierungen können vorkommen (Wilson, Wiebe und Hwa 2004).

Daraus lässt sich schlussfolgern, dass es wichtig ist, Sätze detaillierter zu untersuchen und eventuell vorkommende Wortgruppen zu analysieren. Auch die Berechnung der Stärken der Polaritäten kann einen Mehrwert erbringen. Außerdem ist zu beachten, dass, wie bereits erwähnt, objektive Sätze ebenso indirekt eine Meinung enthalten können wie subjektive keine ausdrücken müssen. (Liu 2011, 477)

In vielen Anwendungsfällen sind die einzelnen Meinungen hinsichtlich der Aspekte einer Entität von großer Bedeutung. Daraus lässt sich schlussfolgern, dass

es hilfreich ist, alle vorhandenen Aspekte zu extrahieren, um dann die Meinungen zu diesen in Relation setzen zu können. Selbst in einem Satz können mehrere Polaritäten und Entitäten vorkommen. Dadurch entstehen größere Anforderungen an die Verarbeitung der natürlichen Sprache, wie beispielsweise das Erkennen von Wortgruppen. Bei diesem Vorgehen können jedoch auch in einem Satz alle Polaritäten erkannt werden. (Liu 2011, 480ff.)

Zudem kann es vorkommen, dass in einem Dokument mehrere Entitäten auftauchen und bewertet werden. Wie schon bei den Aspekten entsteht das gleiche Problem, dass extrahiert werden muss, was das Thema der Bewertung ist. (Pang und Lee 2008, 43)

Außerdem kann durch die Analyse auf der Wortgruppen-Ebene eventuell dem Problem begegnet werden, dass meinungsbehaftete Wörter bei unterschiedlichen Produkten auch unterschiedliche Bedeutungen ausstrahlen. Die satzweise Bestimmung der Polarität kann besonders bei Bewertungen sehr ungenau sein, da oft zwei Eigenschaften von einem Produkt und ihre dazugehörige Meinungen genannt werden. Diese müssen jedoch nicht die gleiche semantische Orientierung besitzen. Aus diesem Grund betrachten DING, LIU und YU in ihrer Arbeit alle meinungsbehafteten Wörter in der unmittelbaren Nähe eines Aspektes in einem Satz und versuchten so Konflikte zwischen Polaritäten aufzulösen. (Ding, Liu und Yu 2008, 2)

2.3 Angewendete Verfahren des Opinion Mining

Durch die Vielzahl der unterschiedlichen Problemstellungen des Opinion Mining, finden in diesem Gebiet auch unterschiedlichste Verfahren Anwendung. Von der Verarbeitung natürlicher Sprache über maschinelle Lernverfahren und Textklassifizierung bis hin zu Techniken aus dem Bereich der Informationsgewinnung. Auf die in Abbildung 3 exemplarisch visualisierten Fachgebiete und deren verwendete Verfahren beim Opinion Mining wird nun näher eingegangen.

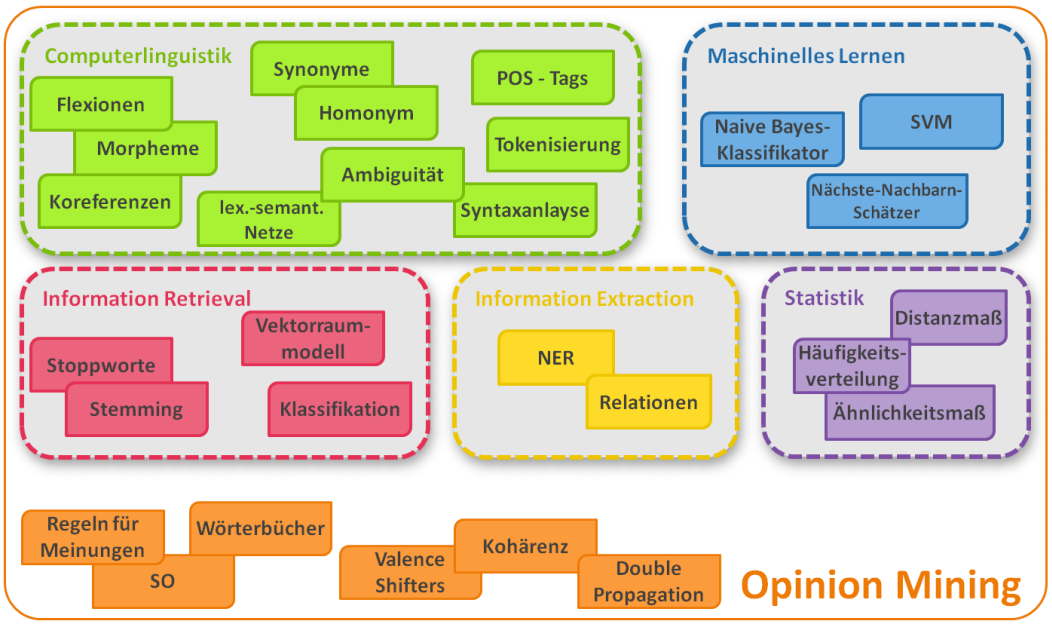

Abbildung 3: Übersicht über angewendete Methoden im Opinion Mining

2.3.1 Verfahren aus dem Bereich Maschinelles Lernen

Das maschinelle Lernen beschäftigt sich mit dem Erkennen von Regeln und Mustern aus vorhandenen Objekten sowie dem Erstellen von annähernd genauen Prognosen für zukünftige Objekte. Anwendung finden diese Verfahren im Bereich des Data Mining, aber auch in dem Gebiet der künstlichen Intelligenz. Damit letzteres beispielsweise so bezeichnet werden kann, muss ein System fähig sein, aus gegebenen Umständen zu lernen und auf neue Situationen erlerntes Wissen anwenden zu können. (Alpaydin 2004, 1f.)

Maschinelles Lernen kann u. a. für das Erlernen von Assoziationsregeln oder auch für die Klassifikation von Objekten hilfreich sein (Alpaydin 2004, 4f.). Eine weitere Bezeichnung für das Verfahren ist das Gewinnen von Wissen. Das Erkennen von Regeln und das Nutzen dieser für weitere Anwendungen wird auch als Aneignen von Wissen bzw. als Lernen eingeordnet. (Ferber 2003, 102)

Ein spezieller Anwendungsfall des maschinellen Lernens ist die Textklassifikation. Sie wird u. a. häufig im Information Retrieval benötigt, findet aber auch beim Opinion Mining Anwendung. Hier wird davon ausgegangen, dass ein Dokument in einem höher dimensionalen Raum aus Dokumenten abgelegt ist und dass eine Anzahl von Klassen, in welche Dokumente eingeordnet werden, vorhanden ist. Die Klassen werden manuell mit Hilfe von Beispieldokumenten definiert. Gesucht wird nun eine Abbildung, welche die Dokumente den richtigen Klassen zuordnet.

Anschließend wird diese Abbildung an neuen Dokumenten, welche noch keiner Klasse zugeordnet sind, getestet. (Manning, Raghavan und Schütze 2009, 256f.)

Für die Umsetzung der Verfahren gibt es drei Varianten: Das überwachte Lernen, das heißt das System lernt aus einer Menge von Beispieldaten und kennt dementsprechend Eingabewerte und die dazugehörigen Ausgabewerte. Das unüberwachte Lernen, welches nur Daten als Eingabe bekommt und in diesen Gesetzmäßigkeiten und Muster finden muss. Eine weitere Methode dafür ist beispielsweise das Clustering. Als dritte Variante existiert das bestärkende Lernen, bei dem ein System die korrekte Folge von Aktionen erlernt, also Taktiken für gewisse Zustände. Anwendungsbeispiele hierfür sind Brettspiele, wo nur die beste Taktik zum Sieg führt. (Alpaydin 2004, 11ff.)[17]

Das überwachte Lernen (engl.: supervised learning), welches auch in dieser Arbeit angewendet wird, aus Beispielen wird ebenfalls als induktives Lernen bezeichnet, aber auch als Klassifikation (Ferber 2003, 113), (Liu 2011, 63). Damit das induktive Lernen möglich wird, werden Beispiele benötigt, mit denen die Regeln erlernt werden können. Diese Beispiele werden Trainingsmenge genannt, mit welcher der Lern-Algorithmus hinsichtlich der beinhaltenden Vorschriften trainiert wird. Für die Überprüfung wird eine weitere Menge, die Testmenge, benötigt, welche ebenso aus bereits korrekt kategorisierten Beispielen besteht. Diese darf nicht für die Trainingsphase verwendet werden, damit die vorhandenen Dokumente neu für den Algorithmus sind. Beide Mengen werden zum größten Teil manuell erstellt, was einen hohen Arbeitsaufwand bedeutet. (Liu 2011, 64f.)

Für das Erlernen der Regeln können alle vorhandenen Beispiele in einem zweidimensionalen Raum abgebildet werden. Ein Klassifikator versucht daraus eine Hypothese, also eine Abbildung, zu erstellen, mit welcher anschließend neue Beispiele annähernd genau eingeordnet werden können. Hierfür wird vorher eine Hypothesenklasse gebildet, welche festlegt, welche Menge in eine bestimmte Kategorie gehört. (Alpaydin 2004, 19ff.)

Eine weitere Form des überwachten Lernens beinhaltet das Kennen aller Attribute der Trainingsbeispiele. Diese Trainingsmenge zu erstellen ist meist sehr aufwendig. Deshalb gibt es die zweite, abschwächte Form, das sogenannte bewertete Lernen. Hier sind nur die Kategorisierungen der Beispiele bekannt. (Ferber 2003, 114)

[17] In der folgenden Arbeit werden zum größten Teil nur überwachte Methoden angewendet, aus diesem Grund wird auf die beiden anderen Varianten hier nicht näher eingegangen.

Bei Trainingsbeispielen mit Attributen sollten diejenigen, die denselben Wert annehmen, auch in der gleichen Kategorie vorzufinden sein. Dieser Umstand wird als konsistent bezeichnet. Von Vorteil ist es außerdem, wenn die Wertebereiche der Attribute endlich sind und für jeden Wert ein Beispiel auftaucht. Somit wird ausgeschlossen, dass neue Attributwerte bei neuen Objekten auftauchen, welche anhand des trainierten Algorithmus nicht kategorisiert werden können. (Ferber 2003, 120f.) Bei der Klassifikation sind die Ausgabewerte einfache boolesche Typen, das heißt ein Objekt gehört in eine bestimmte Klasse oder nicht (Alpaydin 2004, 32).

Wichtig für das Erstellen einer Trainingsmenge ist die ausreichende Größe – es müssen möglichst viele Beispiele vorhanden sein und diese müssen auch in einer repräsentativen Häufigkeit auftreten. Zusätzlich müssen sie als Stichprobe der Realität diese auch ausreichend darstellen. Klassifizierungsalgorithmen können mit jeglichen Arten von Mengen trainiert werden, jedoch ist das häufige Auftreten von Beispielen ein wichtiger Faktor beim Finden von Mustern und Regeln. (Ferber 2003, 127), (Liu 2011, 66)

Die Überanpassung ist ein Problem der Klassifizierung. Hier erlernt der Algorithmus zu spezielle Regeln aus der Trainingsmenge und kann neue Beispiele nicht korrekt klassifizieren, wenn sie nicht in der gleichen Art und Weise wie die Trainingsbeispiele vorkommen. Aus diesem Grund ist es wichtig eine repräsentative Trainingsmenge zu verwenden, sie muss den Bereich, welcher bearbeitet werden soll ausreichend, aber auch allgemein darstellen. Meist kann es wichtiger sein, häufig vorkommende Beispiele richtig zu kategorisieren, als seltene Beispiele. (Ferber 2003, 129)

Es gibt unterschiedliche Arten von Verteilungen, welche in welchen die Trainingsmengen auftauchen können. Die Multinomialverteilung kommt bei dem Naive Bayes Klassifikator zum Einsatz und bedeutet hinsichtlich der Textklassifikation, dass auch die Frequenzen von Termen beachtet werden, da sich somit die Wahrscheinlichkeit eines Terms für eine bestimmte Klasse ergibt (Manning, Raghavan und Schütze 2009, 264).

Die Bernoulli-Verteilung wird bei einem Zweiklassenproblem angewendet. Hier können nur zwei mögliche Ergebnisse vorkommen, ein Ereignis tritt ein oder auch nicht (Alpaydin 2004, 67). Auf die Textklassifikation bezogen bedeutet dies, dass ein Dokument zu einem bestimmten Thema gehört oder auch nicht.

Diese Verteilung ist ein multivariates Modell, was eine Alternative zu der Multinominalverteilung ist. Bei der Textklassifikation mit dem Bernoulli-Modell wird außerdem nur das Auftreten von Termen beachtet, nicht die Frequenz des Vorkommens eines Terms. (Manning, Raghavan und Schütze 2009, 263f.)

Der Naive Bayes Ansatz mit der Bernoulli-Verteilung zum maschinellen Klassifizieren von Textdokumenten eignet sich nur, wenn für eine Kategorie ausschlaggebend ist ob ein bestimmter Term vorkommt oder nicht (Manning, Raghavan und Schütze 2009, 265f.).

Die Varianz stellt dar, wie nah alle vorhandenen Werte einer Stichprobe aneinander liegen, das heißt dem Erwartungswert entsprechen. Je größer die Werte auseinander liegen, desto größer ist die Varianz. Die Verzerrung dagegen, misst wie gut neu geschätzte Werte sind, das heißt, wie stark sie dem Erwartungswert entsprechen. Ein Schätzer, beispielsweise ein Textklassifikator, wird daran gemessen, wie gut die echten Werte geschätzt werden. Also wie gut der Erwartungswert, welcher von einer Funktion berechnet wird, ist. Er sollte bestmöglich der Realität entsprechen. Bei der Klassifikation von Objekten entsteht das sogenannte Varianz/Verzerrung-Dilemma. Das heißt, wenn ein Modell zu komplex ist, also an viele Beispieldaten angepasst ist, ist eine höhere Varianz und eine Überanpassung vorhanden. Da aber ein komplexeres Modell gut an den vorhandenen Datensatz angepasst ist, gibt es eine geringere Verzerrung. Wenn dagegen die Varianz kleiner wird, also das Modell auf speziellere Daten angepasst ist, ist die Verzerrung wiederrum größer, da es häufiger vorkommt, das ein Objekt nicht eingeordnet werden kann, da es nicht im Trainingsdatensatz vorhanden war. Das Modell ist also optimal, wenn zwischen diesen beiden Werten der bestmögliche Kompromiss gefunden wird. (Alpaydin 2004).

Ähnlichkeitsmaße geben an wie ähnlich sich zwei Vektoren sind. Beispielsweise beim Abbilden von Dokumenten im Vektorraum, wird dies zum Vergleichen genutzt, ob eine Suchanfrage zu vorhandenen Dokumenten passt. (Ferber 2003, 72) Ein Beispiel hierfür ist das Kosinus-Maß, welches den Kosinus zwischen den Winkeln zweier Vektoren berechnet, je kleiner, desto ähnlicher. Bei Distanzmaßen dagegen, spricht der kleinste Wert für das ähnlichste Paar. Hierfür wird häufig die Euklidische Distanz verwendet.

Mit Hilfe einer Testmenge werden die Klassifikationsalgorithmen hinsichtlich ihrer Ergebnisse bewertet. Die Testmenge sollte außerdem die zukünftig zu klassifizierenden Dokumente mit ausreichenden Beispielen gut darstellen. Nun kann auf

Basis der Ergebnisse, die der Algorithmus auf dieser Testmenge erzielt hat, die Genauigkeit berechnet werden, mit der der Algorithmus arbeitet. Die Genauigkeit wird oft als Vergleichswert unterschiedlicher Klassifikatoren verwendet. (Liu 2011, 65)

Verfahren, welche neu entwickelt werden, müssen anschließend hinsichtlich ihrer Qualität beurteilt werden, sei es beim Information Retrieval, beim maschinellen Lernen oder auch beim Opinion Mining selbst. Häufig genutzte Maße hierfür sind Genauigkeit[18] und Trefferquote[19], bzw. zusammengefasst das F-Maß. Im Information Retrieval, woher diese Werte stammen, wird die Precision dafür genutzt zu bestimmen, wie viele der gefundenen Dokumente, beispielsweise bei einer Suchanfrage auch wirklich relevant sind und der Recall steht dafür wie viele relevante Dokumente gefunden wurden. Gerade durch die Masse an Daten im Internet ist letzterer Wert sehr schwer zu bestimmen, da nicht alle relevanten Dokumente bekannt sind. (Thielmann und Paijmans 2004, 368f.)

Bei Klassifikationsverfahren, steht die Precision dafür, wie oft eine Entscheidung korrekt getroffen wurde und der Recall, dafür wie oft überhaupt eine Entscheidung ein Ergebnis lieferte. (Heyer, Quasthoff und Wittig 2006, 256)

Auch in der Computerlinguistik finden diese Werte Anwendung. Häufig wird ein Mittelwert aus beiden zusätzlich angegeben, das F-Maß, das gewichtete harmonische Mittel. (Carstensen, et al. 2010, 490)

Wenn für die Ergebnisse nur genau eine Klasse wichtig ist, dann ist beim maschinellen Lernen die Genauigkeit, wie eingeordnet wurde, nicht aussagekräftig genug. Und auch hier finden dann Precision und Recall, ebenso das F-Maß, Anwendung, mittels einer sogenannten Konfusionsmatrix. Hier wird festgehalten, wie korrekt und inkorrekt die Klassifikation der einen Klasse war. (Liu 2011, 81f.)

Um zu prüfen, ob die angewendeten Verfahren korrekt gearbeitet haben, werden die Ergebnisse validiert. Bei einem kleineren Datensatz, ist es oft so, dass die Kreuzvalidierung angewendet wird. Hier wird der Datensatz in n sich nicht überschneidende, gleichgroße Teile geteilt. Diese werden dann in n-1 Trainingsmengen und einer Testmenge aufgeteilt, wobei dieser Vorgang n mal wiederholt und somit jeder Teil einmal Testmenge ist. (Liu 2011, 80)

[18] Zu engl. Precision. Dieser Begriff wird in der Arbeit als Synonym für das dt. Wort Genauigkeit verwendet.
[19] Zu engl. Recall. Dieser Begriff wird ebenso weiterhin als Synonym für die Trefferquote verwendet.

Ein Beispiel für ein Verfahren aus dem Bereich des maschinellen Lernens ist die Naive Bayes Text Klassifikation. Hier wird mit Hilfe der Verteilung von Objekten einer Trainingsmenge auf verschiedene Kategorien versucht, die wahrscheinlichste Klasse für Objekte zu bestimmen (Liu 2011, 100f.). Bei der Naive Bayes Klassifikation wird davon ausgegangen, dass alle Attribute unabhängig voneinander sind (Klapdor und Felden 2005, 36).

Die Wahrscheinlichkeit eines Dokumentes in eine bestimmte Klasse zu gehören, wird in Abhängigkeit von der Auftrittswahrscheinlichkeit eines bestimmten Terms in einem Dokument einer bestimmten Klasse und von der Gesamtwahrscheinlichkeit für Dokumente, einer bestimmte Klasse anzugehören, berechnet (Manning, Raghavan und Schütze 2009, 258).

PANG und LEE zeigen in einer Arbeit, dass Unigramme als Features für einen Klassifikator bei einem „Naive Bayesian Klassifikator“ gute Ergebnisse erzielten (Pang und Lee 2002). Häufig sind jedoch auch Modelle von Klassifikatoren domänenabhängig, u. a. auch aufgrund der unterschiedlichen Wortbedeutungen in verschiedenen Domänen (Liu 2011, 469).

YU und HATZIVASSILOGLOU wenden in ihrer Arbeit einen Naive Bayes Klassifikator für das Unterscheiden zwischen subjektiven und objektiven Dokumenten an. Ihre Trainingsmenge besteht aus „Wall Street“-Artikeln und als Feature werden einzelne Wörter verwendet. Der Klassifikator erreicht hier ein besonders gutes F-Maß von 97%. (Yu und Hatzivassiloglou 2003, 3ff.)

Das „Nächste-Nachbarn-Schätzer“-Verfahren ist Art von Klassifikation werden die Trainingsobjekte auch in den mehrdimensionalen Raum übertragen. Neue Objekte werden anschließend mit Hilfe der nächsten Nachbarn eingeordnet. Mittels Abstandsfunktionen, wie der Euklidischen Distanz oder dem Kosinus-Maß bei der Textklassifikation, werden Distanzen zwischen den Objekten ermittelt (Liu 2011, 124f.). Somit werden lokale Einordnungen gefunden, in dem die ähnlichsten Vektoren ermittelt werden. Die Kategorie, welche unter den Nachbarn am häufigsten vorkommt, ist somit auch die des neu eingeordneten Objektes. (Manning, Raghavan und Schütze 2009, 297)

TURNEY wendet POS-Tags in seiner Arbeit an, um außerdem mit Hilfe des Kontextes die semantische Orientierung eines Wortes zu bestimmen. Hierfür wurden immer zwei aufeinanderfolgende Wörter extrahiert, wo das zweite für den Kontext steht. Dies wiederum geschah mit Hilfe von Mustern, welche besonders aussagekräftig sind. Solche Muster sind beispielsweise ein Adjektiv, das zusammen

mit einem Substantiv auftritt. Im zweiten Schritt wird die semantische Orientierung mit Hilfe der Pointwise Mutual Information bestimmt. Es wird berechnet, wie ähnlich das extrahierte Wort, zu bereits vorausgewählten Mengen an positiven oder negativen Wörtern ist. Das heißt, wie stark die Assoziation zu den positiven bzw. negativen Wörtern ist. Die Pointwise Mutual Information kann mit Hilfe von Termfrequenzen oder auch Kookkurrenzen berechnet werden. (Turney 2002)

Eine Erweiterung dieser Annahme erfolgt durch AUE und GAMON, indem die Hypothese hinzugefügt wird, dass Wörter mit gegensätzlicher semantischer Orientierung nicht in ein und demselben Satz auftauchen. (Aue und Gamon 2005)

2.3.2 Verfahren aus der Computerlinguistik

Dokumente, welche als Freitext vorliegen, müssen für die weitere Verarbeitung in die Form linguistischer Einheiten gebracht werden. Dies geschieht beispielsweise durch das Zerlegen in Sätze, Wortgruppen oder Phrasen. Hierfür spielt die sogenannte Tokenisierung eine wichtige Rolle. Das Erkennen eines Satzes ist trivialerweise an Satzzeichen, wie „!“, „?“ und natürlich dem „.“ möglich. Jedoch kommt es nicht selten vor, dass diese drei Zeichen auch im Zusammenhang mit Abkürzungen oder Eigennamen auftreten. Die einfachste Lösung hierfür sind manuell erstellte Listen, in denen alle möglichen Abkürzungen o. ä. aufgelistet sind. Jedoch wäre eine automatisierte Methode, welche auf statistischen Informationen beruht, vorteilhafter. Durch ersteren Ansatz können nur selten alle Möglichkeiten erfasst werden. (Carstensen, et al. 2010, 264ff.)

Die Satzverarbeitung spielt besonders bei der Analyse von Meinungen auf der Satzebene eine große Rolle. HU und LIU splitten in ihrer Arbeit alle Nutzerbewertungen in einzelne Sätze. Ihrer Meinung nach ist das Finden einer Gesamtaussage eines Dokumentes für viele Anwendungsfälle unbefriedigend. (Hu und Liu 2004)

Nachdem ein Text bzw. ein Dokument in linguistische Einheiten unterteilt wurde, ist es häufig von großem Nutzen, zu erkennen, was genau diese Einheiten darstellen. Hierbei kommen bereits erwähnte Verfahren, welche Wörter mit POS-Tags versehen zum Einsatz (Carstensen, et al. 2010, 271). Dies kann besonders für das Erstellen von Listen mit meinungsbehafteten Wörtern hilfreich sein, da, wie bereits erwähnt, u. a. Adjektive besonders wichtige Indikatoren für Subjektivität

sind. HU und LIU wendeten POS-Tags[20] für das Finden von Aspekten der gesuchten Produkte an, da diese meist durch Nomen beschrieben werden. (Hu und Liu 2004, 4)

Eine weitere hilfreiche Methode ist das Finden von Synonymen und Antonymen aus einem online verfügbaren Wörterbuch. Somit kann eine Startmenge von Adjektiven mit bekannten Polaritäten automatisch vergrößert werden. HU und LIU wenden das Verfahren erfolgreich in einer ihrer Arbeiten an. Der Prozess erfolgt iterativ, in dem jedes neue gefundene Wort der Startmenge hinzugefügt wird. Diese Methode kann gelegentlich bessere Ergebnisse liefern, als statistische Methoden. Letztere lernen einmalig aus großen Datenmengen und bedeuten immer einen höheren manuellen Aufwand, da diese Trainingsmengen vorher erstellt werden müssen. (Hu und Liu 2004, 5)

Ein weiteres Gebiet mit dem sich die Computerlinguistik beschäftigt, ist die Mehrdeutigkeit verschiedener Wörter in unterschiedlichen Kontexten. Weitere Bezeichnungen sind Polysemie- und Ambiguitätsprobleme. Hier schafft nur die Einbeziehung des Kontextes Abhilfe. (Ferber 2003, 46f.)

Korreferenzen sind, wie Synonyme, ebenfalls ein wichtiger Punkt bei der Analyse von Texten. Dies sind gleiche Referenzen für ein und dasselbe Objekt im Text. Beispielsweise „Deutsche Bahn AG“ und „Die Bahn“ ist ein solcher Fall. Diese Beziehungen müssen gefunden und aufgelöst werden (Carstensen, et al. 2010, 336). Auch beim Finden von Meinungen bezüglich dieser Aspekte tritt dieses Problem auf.

Auch die sogenannte Anaphern Resolution gehört dazu. Diese kommt beim Opinion Mining hinsichtlich der Aspekt Extraktion häufig vor. Auf das Objekt wird mit Hilfe eines Referenzausdruckes gezielt, beispielsweise mit „es“ o. ä. Aufgelöst wird dies durch Filter und Präferenzen, indem beispielsweise semantische Beziehungen oder syntaktische Restriktionen betrachtet werden. (Carstensen, et al. 2010, 399)[21]

Eine Abhilfe können hier lexikalisch-semantische Netze[22] schaffen. In ihnen werden unter anderem Relationen von Wörtern bzw. Konzepten zueinander beschrieben. Diese Verbindung kann durch Homonymie, das heißt, dass ein Wort auf mehrere Konzepte verweist oder auch durch Synonymie, welche dafür steht, dass

[20] Eines der bekanntesten deutschen Tagsets ist das Stuttgart-Tübingen- Tagset
[21] Weiterführende Literatur: (Carstensen, et al. 2010, 399ff.); (Roth 2006, 103ff.)
[22] Ein Beispiel hierfür ist WordNet

mehrere Wörter die gleiche Bedeutung haben, ermittelt werden. (Carstensen, et al. 2010, 378ff.)

Damit die richtige Bedeutung eines mehrdeutigen Wortes[23] gefunden werden kann, kommt die sogenannte Lesarten Disambiguierung (zu engl.: „word sense disambiguation) zur Anwendung (Roth 2006, 11). Hierfür gibt es unterschiedliche Ansätze. Einer davon basiert auf Lexika und Thesauren, der sogenannte Wissensbasierte Ansatz. Ein anderes Verfahren arbeitet mit statistischen Methoden und den Häufigkeiten, wie oft ein Wort in großen Dokumentsammlungen auftritt, der sogenannte Korpus-basierte Ansatz (Roth 2006, 77).[24] Hier ist eine Ähnlichkeit zu den Verfahren des Opinion Mining zu finden.

Auch beim Information Retrieval werden unter anderem Thesauren vorrangig für die Beseitigung von Polysemie und Synonymie verwendet. Thesauren beinhalten Wörter, Terme und Ausdrücke eines Sachgebietes und bilden zusätzlich die Beziehungen dieser untereinander ab. Im Allgemeinen, werden also hier für ein Wort neben der Definition, auch die Synonyme, Antonyme, verwandte Wörter, Oberbegriffe und speziellere Begriffe angegeben. Somit können mehrdeutige Wörter eindeutig erkannt werden. (Ferber 2003, 54f.)

Der Teilbereich Morphologie aus der Sprachwissenschaft bringt wichtige Regeln für das Bilden und Erkennen von Wörtern und Wortformen mit sich. Ein Wort kann in mehreren Formen vorkommen, beispielsweise durch Anpassung an die einzelnen Fälle oder an die Mehrzahl. Dies wird als Flexion bezeichnet.[25] Morpheme sind die kleinste Kette von Zeichen, welche noch eine Bedeutung oder grammatikalische Funktion haben. Sie bilden die Grundlage für die Bildung von Flexionen und Derivationen (z. dt.: Wortbildungen). Ein Wortstamm dagegen ist ein Morphem bzw. eine Kombination aus Morphemen, welche keine Flexionsendungen besitzt. (Carstensen, et al. 2010, 167f.)

Für die deutsche Sprache ergeben sich mehr Probleme als in der englischen Sprache hinsichtlich der maschinellen Analyse. Wörter können in starken Variationen auftreten. Unter anderem erfolgen in deutscher Sprache häufig Wortbildungen durch das Anhängen von Suffixen sowie das Voranstellen von Präfixen. Zusätzliche Schwierigkeiten bringen Präfixe mit sich, welche entweder bei bestimmten Beugungen abgetrennt werden müssen, oder eben nicht. Dies kann die Bedeutung von Wörtern erheblich verändern. Außerdem wird in der deutschen Sprache oft

[23] Beispiel: Hahn, der Vogel und Hahn, der Wasserhahn.
[24] Weiterführende Literatur: (Carstensen, et al. 2010); (Roth 2006)
[25] Beispiel: Der Mensch, die Menschen.

ein neues Wort aus mehreren Wörtern zusammengesetzt und dies nach den unterschiedlichsten Regeln. Somit ist es nicht möglich, allgemeingültige Regeln für die Wortbildung in der deutschen Sprache zu finden. Dementsprechend ist es für eine Vorverarbeitung von Daten wichtig, diese Unregelmäßigkeiten in einem Wörterbuch zu speichern. Da dies nur manuell umsetzbar ist, erfordert es einen sehr hohen Aufwand. (Ferber 2003, 44f.)

2.3.3 Verfahren aus dem Information Retrieval

Ein häufig genutztes Modell im Information Retrieval ist das Vektorraummodell. Bei dieser Methode werden extrahierte Dokumente in einem Vektorraum abgebildet. Da somit Dokumente mathematisch dargestellt werden, ist es möglich diese zu vergleichen und ähnliche Dokumente zu finden (Ferber 2003, 61ff.). Die Umwandlung in einen Vektor erfolgt durch das Wichten aller Terme eines Dokumentes und der darauf basierenden Einordnung in den Vektorraum. Gewichtungsmaß kann beispielsweise die TDF-IDF sein. Auf diese Vektoren können nun Ähnlichkeitsmaße angewendet werden, um beispielsweise Suchanfragen mit vorhandenen Dokumenten zu vergleichen. Ein häufig genutztes Maß dafür ist das Kosinus-Maß. (Carstensen, et al. 2010, 589f.)

Bevor Dokumente bzw. Texte verarbeitet werden können, sind einige Schritte notwendig, um diese dafür vorzubereiten. Text in natürlicher Sprache ist meist nicht von bester Qualität, besonders wenn dieser nicht aus offiziellen Dokumenten stammt Vor der Verarbeitung werden beispielsweise häufig vor sogenannte Stoppwörter entfernt. Dies sind Wörter, die in einem Dokument sehr oft vorkommen, aber keine große Bedeutung tragen. Ein Beispiel hierfür ist „ein“ oder auch „für“. (Liu 2011, 227)

Ein weiteres, oft verwendetes Verfahren im Information Retrieval ist das sogenannte Stemming (zu dt.: „Grundformreduktion). Damit werden Wortformen auf den gemeinsamen Wortstamm reduziert. Die Trefferquote, beispielsweise bei der Volltextsuche wird dadurch deutlich erhöht. (Carstensen, et al. 2010, 589)

Es gibt zwei Varianten für das Stemming. Entweder wird das Wort auf seine grammatikalische Grundform herunter gebrochen, was bei Substantiven der Nominativ Singular und bei Verben der Infinitiv ist. Oder es erfolgt eine Stammformreduktion, bei der das Wort auf seinen Stamm reduziert wird. Dieser taucht meist nicht als eigenständiges Wort in der Sprache auf (Ferber 2003, 40f.). Somit können im schlechtesten Fall (Overstemming) unterschiedliche Wortarten den

gleichen Wortstämme besitzen (Carstensen, et al. 2010, 589). Die Reduzierung auf die Grundform eines Wortes wird Lemmatisierung genannt.

Dieses häufig als Vorverarbeitungsschritt verwendete Verfahren, ebenso wie das Filtern von Stoppworten, bringt Vorteile bei der Verwaltung großer Datenmengen mit sich. Somit müssen weniger Terme in das Vektorraummodell übertragen werden, da eine Grundform gleichzeitig für mehrere Wortformen steht. Außerdem werden somit ggf. Auftrittswahrscheinlichkeiten für bestimmte Worte größer, da nun u. a. auch die gebeugten Wortformen gefunden werden können und nicht nur nach der Originalform eines Wortes in einem Text gesucht wird, welche im alltäglichen Sprachgebrauch eher selten vorkommt. (Ferber 2003, 41) HU und LIU nutzten das sogenannte Stemming in ihrer Arbeit, um alle Wortvarianten eines gewünschten Aspektes zu finden (Hu und Liu 2004, 4).

Ein Verfahren im Information Retrieval sowie auch der Informationsextraktion ist die Klassifikation von Objekten bzw. Dokumenten. Hier werden Dokumente nach bestimmten Kategorien - also systematisch - geordnet. Somit können beispielsweise Suchanfragen leichter verarbeitet werden, da nur in Dokumenten gesucht wird, welche das gesuchte Thema beinhalten. (Ferber 2003, 47f.)

Häufig wird die Textklassifikation auch als Themen-Klassifikation bezeichnet, da größtenteils Dokumente hinsichtlich ihrer Themen kategorisiert werden (Carstensen, et al. 2010, 591).

Verfahren, welche sich hierfür besonders gut eignen, sind auf statistischen Werten basierende Verfahren. Hierfür werden Trainingsbeispiele manuell nach den Regeln für die Klassenbildung markiert und die sogenannten überwachten Lernverfahren lernen daraus und berechnen aus den Trainingsdaten Wahrscheinlichkeiten für die neu einzuordnenden Dokumente. (Carstensen, et al. 2010, 592)

Unterstützt wird die Klassifikation meist durch das Finden der Attribute von Objekten. Anhand derer kann die Differenzierung vollzogen werden. Das heißt, dass alle Objekte mit einem gewissen Attribut in eine bestimmte Klasse gehören (Ferber 2003, 47f.). Das Vektorraummodell ist eine gute Grundlage für dieses Verfahren, da hier Dokumente anhand ihrer Merkmale abgebildet werden können (Carstensen, et al. 2010, 592). Die Textklassifikation findet u. a. häufig Anwendung in der Informationsextraktion und auch beim Opinion Mining (Carstensen, et al. 2010, 592).

2.3.4 Verfahren aus der Informationsextraktion

Die Extraktion von Eigennamen ist eine wichtige Methode der Informationsextraktion. Unter Eigennamen werden Bezeichnungen für Personen, Produkte, etc. verstanden. Eine einfache Liste mit allen möglichen Varianten ist oft nicht erstellbar, da Namen in unterschiedlichsten Formen auftauchen können. Auch hier müssen Vorverarbeitungsschritte wie Tokenisierung, morphologische Analyse und lexikalische Analyse vollzogen werden, jedoch bringen diese keine ausreichenden Ergebnisse. Aus diesem Grund kommen hier zusätzlich unter anderem maschinelle Lernverfahren für das Lösen dieses Problems in Frage. Neben dem überwachten Lernen, welches den Nachteil mit sich bringt, dass große Mengen an Trainingsbeispielen manuell erstellt werden müssen, kommen auch semi-überwachte Verfahren zum Einsatz. Bei mehrdeutigen Eigennamen ist die Erkennung schwieriger, da auch hier wieder der Kontext mit einbezogen werden muss. So steht beispielsweise Essen für einen Ort, aber auch für eine Mahlzeit. Ebenso ist das Vorkommen von Koreferenzen ein Problem der Eigennamenerkennung. Um diese Disambiguierung aufzulösen, können beispielsweise externe Wissensquellen in Form von Lexika[26] o. ä. genutzt werden. (Carstensen, et al. 2010, 596f.)

Neben der Eigennamenextraktion ist die Extraktion von Relationen zwischen Entitäten eine weitere Hauptaufgabe der Informationsextraktion. Auch hier kommen maschinelle Lernverfahren, davon vorwiegend überwachte, zum Einsatz (Carstensen, et al. 2010, 599f.).

Traditionelle Informationsextraktions-Systeme haben meist den Nachteil, dass sie manuell erstellte Trainingsmengen benötigen und an den jeweiligen Anwendungsfall stark angepasst sind. Die Entwicklung geht in die Richtung unüberwachter Lernverfahren, da verbesserte Methoden der Informationsextraktion mittlerweile immer häufiger in anderen Gebieten gebraucht werden, wie beispielsweise dem Opinion Mining. (Carstensen, et al. 2010, 603f.)

2.3.5 Verfahren speziell für Opinion Mining

Eine häufige Herangehensweise hierfür ist das Nutzen von Startmengen mit Wörtern, von denen die semantische Orientierung bereits bekannt ist.

[26] Wikipedia wird sehr häufig dafür genutzt, da es zu fast jedem Beliebigen Objekt über einen Eintrag verfügt.

YU und HATZIVASSILOGLOU bestimmen die semantische Orientierung weiterer Wörter, indem sie die Kookkurrenzen der Wörter aus einer Startmenge mit den neuen Wörtern betrachteten. Hier ist die Hypothese, dass positive Wörter häufig zusammen mit anderen positiven Wörtern auftauchen, die Grundlage. (Yu und Hatzivassiloglou 2003, 4)

DING, LIU und YU erstellten mit Hilfe von WordNet Listen wie bei (Hu und Liu 2004) aus Adjektiven, Nomen und Verben und berechneten mit Hilfe einer Summenfunktion die Orientierung bezügliche eine Produktaspektes. (Ding, Liu und Yu 2008, 4)

Negationen können die gefundenen Orientierungen eines Wortes in das Gegenteil umkehren. Im Information Retrieval ist die Beachtung dieser nicht so essentiell, wie beim Opinion Mining. (Pang und Lee 2008, 36)

DING, LIU und YU wendeten für das Erkennen von Negationen bestimmte Muster an. Beispielsweise wenn eine Negation vor einem negativen Wort gefunden wird, ergibt es insgesamt eine positive Aussage. (Ding, Liu und Yu 2008, 5)

DING, LIU und YU benutzten zusätzlich Regeln für Konjunktionen, welche innerhalb eines Satzes vorkommen. Wenn die Orientierung für ein bestimmtes unbekannt ist, dann wird diese mit Hilfe von schon bekannten Wörtern interpretiert. Ebenso wird dies auf die Satzebene übertragen. Wenn der vorherige Satz als Polarität positiv berechnet wurde, ist die Wahrscheinlichkeit hoch, dass der nachfolgende ebenso dieser Polarität angehört. (Ding, Liu und Yu 2008, 6).

Um in ihrer Arbeit auch infrequente Aspekte zu finden, wenden HU und LIU eine einfache Hypothese an. Wenn häufig vorkommende Aspekte bekannt sind und ebenso deren „Opinion Words“, so kann davon ausgegangen werden, wenn ein „Opinion Word“ in einem Satz vorkommt, in dem aber noch kein Aspekt bekannt ist, dass dort eins enthalten sein müsste. Das Nomen, welches am nahesten an dem meinungsbehafteten Wort im Satz steht, wird als Aspekt extrahiert. (Hu und Liu 2004, 6)

HU und LIU nutzten außerdem für das Erkennen der Polarität eines Satzes einfach die Anzahl der am meisten vorkommenden Adjektive einer Orientierung. Wenn beide Polaritäten gleich oft vorkommen, wird die Orientierung des effektiven Adjektivs genutzt, das Wort, welches am nahesten an dem Aspekt des Produktes steht. (Hu und Liu 2004, 7)

So genannte Valence Shifter (zu dt.: „Orientierungssänderer“) müssen im Bereich des Opinion Mining ebenso beachtet werden. Dies sind Wörter, welche die semantische Orientierung in einem Satz oder auch eines Wortes verändern. Beispiele hierfür können modale Hilfsverben sein, wie „sollte“ oder „könnte“, aber auch „dennoch“ oder „trotzdem“ können einen Satz einleiten, welcher eine andere Orientierung hat als der vorherige Satz. Diese Indikatoren werden auch für die Stimmungskonsistenz verwendet. (Liu 2011, 483)

2.4 Aktuelle Forschungsarbeiten

2.4.1 SentimentWortschatz

„SentimentWortschatz“ ist ein Wörterbuch in der Sprache Deutsch, welches für Verfahren der Meinungsanalyse eingesetzt werden kann. Es besteht aus Wörtern, mit der dazugehörigen Wortart, dem Gewicht hinsichtlich der Polarität und den vorhandenen Beugungen (Flexionen). Die aktuelle Version enthält 1.650 negative und 1.818 positive Wörter. Es enthält nur Wörter welche explizit oder implizit eine Meinung beinhalten. Erstellt wurde dieses Wörterbuch u. a. aus dem General Inquirer Lexikon. Die Wörter wurden mittels „Google Translate“ in die deutsche Sprache übersetzt.

2.4.2 SemaSuite der T-Systems Multimedia Solutions GmbH

Die SemaSuite ist ein Tool zur semantischen Analyse von Dokumenten. Hier werden Wissensquellen, wie beispielsweise die Lösungen zu häufig gestellten Fragen (FAQ), in Wissensmodelle übertragen. Das bedeutet, dass diese Informationen in eine übersichtsartige Struktur gebracht und bestehende Beziehungen dieser einzelnen Informationen untereinander abgebildet werden. Anschließend können neue Dokumente hinsichtlich des vorhandenen Wissens untersucht und analysiert werden. Ein Anwendungsbeispiel der SemaSuite ist das Zuordnen von Kundenanfragen auf vorhandene FAQs. Somit können gezielt Lösungsvorschläge vermittelt werden, ohne dass ein Mensch diese Anfragen lesen, verstehen und darauf reagieren muss.

Für die Umsetzung dieser semantischen Analyse bedient sich die SemaSuite unterschiedlichster Verfahren aus dem Bereich der Computerlinguistik bzw. der In-

formationsextraktion. Als erstes wird eine Wortschatzextraktion aus vorhandenen Wissensquellen vorgenommen. Hierfür dient als Input eine Menge an Dokumenten, welche die Fachterminologie für die zu analysierende Domäne umfasst.

Diese Dokumente werden in Sätze aufgeteilt („tokenisiert“). Zusätzlich werden alle vorhandenen Wörter gezählt sowie jedes Wort an einem Satzanfang kleingeschrieben und rechtschreibkorrigiert. Umgesetzt wird dies durch Verfahren, wie das Berechnen der Editierdistanz des aktuellen Wortes zu den vorhandenen Wörtern im eingebundenen Wörterbuch. Die Wortformen werden anschließend für die Erstellung eines semantischen Vektorraumes genutzt.

Darauf wird eine Terminologie-Extraktion durchgeführt. Das heißt, die Wörter und Phrasen, welche häufig in den Quelldokumenten vorkommen werden extrahiert. Zusätzlich werden die Wortarten der gefundenen Wörter bestimmt und es erfolgt eine optionale Lemmatisierung[27]. Stoppworte, werden ggf. mit Hilfe einer Stoppwortliste eliminiert. Es kann ausgewählt werden, welche Wörter für das Wissensmodell verwendet werden sollen, beispielsweise können nur Nomen dafür verwendet werden. Diese zulässigen Wörter und Phrasen werden anschließend extrahiert und gezählt (das Vorkommen insgesamt und in wie vielen Dokumenten). Daraufhin werden diese Wörter nach Kriterien gefiltert. Diese Kriterien sind u. a. die Termfrequenz (tf), das heißt ein Term muss mindestens n-mal vorkommen, und die relative Dokumentenhäufigkeit (rdf). Ein Wort, welches nur in wenigen Dokumenten vorkommt, ist beispielsweise oft ein wichtiges Wort für die Fachterminologie. Darüber hinaus wird die Normalized Mutual Information, bei der die Assoziationen zwischen Wörtern durch Hinzuziehen eines Vergleichskorpus gefunden werden. Abschließend wird mit den Ergebnissen dieser Phase ein weiterer semantischer Vektorraum erstellt. Das Ergebnis ist eine Liste aller relevanten Terme und Phrasen, mit den dazugehörigen Eigenschaften Wortart, Frequenz und Wortstamm.

Der nächste Schritt ist die Konzeptextraktion. Hier werden Terme und Phrasen anhand ihrer Kernbedeutung zusammengefasst. Unter anderem werden einem Konzept alle Worte mit demselben Wortstamm sowie Synonyme, welche mit Hilfe eines Thesaurus ermittelt wurden, hinzugefügt. Die extrahierten Konzepte können anschließend manuell verändert werden.

Die darauffolgende Phase vollzieht eine Taxonomie-Extraktion, das heißt es werden eventuell vorhandene Generalisierungen bzw. Spezialisierungen der Konzepte

[27] Eine Definition ist in Abschnitt 2.3.2 zu finden.

untereinander gefunden. Dies geschieht mit Hilfe des Erkennens von Wortmustern oder der semantischen Ähnlichkeit.

Anschließend werden Beziehungen zwischen den Konzepten bei der Assoziationsextraktion gefunden. Konzepte, welche häufig miteinander in einem bestimmten Textfenster auftauchen, welches also Kookkurrenzen sind, werden hierfür als Assoziation extrahiert.

Schlussendlich wird mit Hilfe dieser Informationen das Wissensmodell erstellt. Dokumente welche semantisch analysiert werden sollen, können nun darauf abgebildet werden. Diese werden dafür auf die gleiche Weise verarbeitet, wie die Wissensquellen aus denen das Modell besteht.

2.4.3 Andere aktuelle Veröffentlichungen

ZHANG und LIU versuchten in ihrer Arbeit Nomen zu identifizieren, welche ebenso wie Adjektive eine Polarität besitzen. Als Grundlage dient das Aspekt-basierte Modell des Opinion Mining (Hu und Liu 2004). Jedoch wird dabei kritisiert, dass häufig Meinungen, welche durch objektive Nomen und Sätze impliziert werden, nicht beachtet werden. Diese zu beachten stellt sich als schwierig heraus, aber auch als essentiell für effektives Opinion Mining. In ihrem Verfahren werden zuerst mit Hilfe von POS-Tags alle Aspekte in einem Dokument extrahiert und mit einem auf einem Lexikon basierenden Ansatz (Ding, Liu und Yu 2008) die zugehörige Polarität bestimmt. Anschließend werden die Nomen bestimmt, welche häufig in einem subjektiven Satz auftauchen. Ist beispielsweise kein Adjektiv in deren unmittelbarer Nähe, dann wird impliziert, dass es meinungsbehaftete Nomen sind. Zusätzliche Nomen werden mit einer Precision von 44% bestimmt. Dieses Verfahren hat also noch Potential für zukünftige Forschungsarbeiten. (Zhang und Liu 2011)

QIU, LIU et al. nutzen die sogenannte Double Propagation (zu dt.: „doppelte Verbreitung“) für das Erweitern der Menge an Wörtern mit Stimmungen und zur Extraktion von Zielen der Meinungen, den Aspekten. Es wird also davon ausgegangen, dass Meinungen immer auf ein Objekt zielen. Dementsprechend, kann wenn die Meinung bekannt ist, das zugehörige Ziel extrahiert werden. Es kann aber auch von einem Ziel ausgehend eine unbekannte Meinung gefunden werden. In dieser Arbeit wird mit einer Menge an Adjektiven, welche eine Polarität ausdrücken gestartet. Diese wird mittels Analyse der syntaktischen Relationen in Verbindung mit POS-Tags iterativ erweitert. Falsch gedeutete Ziele werden anhand

von deren Häufigkeiten in den Dokumenten eliminiert, da davon ausgegangen wird, dass Konsumenten von Produkten bei Bewertungen immer über die gleichen Aspekte sprechen (Hu und Liu 2004). Neue meinungsbehaftete Wörter und Ziele wurden in dieser Arbeit mit einer Precision von 88% und einem Recall von 83% gefunden. (Qiu, Liu, et al. 2010)

ZHAI et al. untersuchen in ihrer Arbeit das effektive Gruppieren der Aspekte von Produkten. Bisherige Arbeiten bringen laut den Autoren nicht die gewünschten Ergebnisse, da es für Eigenschaften von Produkten verschiedenste Synonyme gibt. Mit Hilfe von Synonymen in Thesauren werden selten alle Möglichkeiten abgedeckt. Ein Beispiel hierfür ist das Wort „Erscheinung“ und „Design“, welche keine Synonyme sind, jedoch häufig für die gleiche Produkteigenschaft genutzt werden. In ihrer Arbeit wurde versucht, dieses Problem mit Hilfe eines überwachten Modells umzusetzen. Das heißt, es wird anfangs eine kleine Startmenge an markierten Wörtern verwendet. Hierfür werden einfache Bedingungen genutzt, beispielsweise Aspekte, welche in einem Satz vorkommen, gehören meist unterschiedlichen Aspektgruppierungen an, da ein Nutzer kaum die Aspekte mehrmals in einem Satz beschreibt. Zusätzlich wird die lexikalische Ähnlichkeit zweier Aspekte mit Hilfe von WordNet ermittelt. Auch Wortteile, beispielsweise Kundenservice und Service werden zusammen gruppiert. Mit dieser markierten Startmenge wird dann ein Algorithmus, basierend auf dem Naive Bayesian Klassifikator, angewendet. Die Aspekte werden mit dem dazugehörigen Dokument „erlernt“, damit auch die Kontextinformationen gegeben sind. Zusätzlich können Aspekte welche einer Gruppierung zugeordnet sind, auch wieder geändert werden, sofern mit Hilfe von Kontextinformationen eine andere Einordnung gefunden wurde. (Zhai, et al. 2011)

3 Erstellung der Datensätze

In diesem Kapitel wird beschrieben, welche Daten und in welcher Art und Weise diese für die Goldstandards der Domänen Handy und Notebook extrahiert und anschließend hinsichtlich ihrer Polarität markiert wurden. Außerdem wird erläutert, wie ein weiterer größerer Datensatz für jeweils beide Domänen mit Hilfe von Produktbewertungen von Amazon erstellt wurde. In Abschnitt 3.4 wird eine Übersicht über die Eigenschaften der vier Datensätze gegeben. Anschließend folgen Schilderungen der Erkenntnisse und Schwierigkeiten während der Extraktion, sowie das Finden einer möglichen Begründung dafür in Abschnitt 3.5.

3.1 Allgemeine Herangehensweise

Um einen maschinellen Klassifikator für das Kategorisieren von Sätzen hinsichtlich ihrer Polaritäten zu trainieren, wird eine Menge an Beispielen benötigt, aus denen Regeln und Muster erlernt werden können. Diese Beispiele müssen mit der Kategorie, in welche sie gehören markiert sein. In dieser Arbeit diente ein Goldstandard dafür, welcher zusätzlich für zukünftige Forschungen verwendet werden kann. Hiermit könnten neuerarbeitete Methoden und Verfahren bezüglich des Opinion Mining getestet werden.

Der Goldstandard wurde mit Hilfe von Produktbewertungen aus den Domänen Handy und Notebook gebildet, wobei die Auswahl der Produkte zufällig erfolgte.

Eine besonders gute Quelle für das Gewinnen solcher Daten ist der Onlineshop Amazon[28]. Hier sind viele Nutzerbewertungen zu zahlreichen Produkten vorhanden. Zusätzlich werden diese automatisch in fünf Kategorien mit Hilfe der vom Nutzer gegebenen Sternbewertung eingeordnet. Dieser Umstand wurde bei der Erstellung des Goldstandards verwendet. Extrahierte Bewertungen mit vier bis fünf Sternen wurden als positiv, mit drei Sternen als objektiv und mit ein bis zwei Sternen als negativ markiert. Somit entstand automatisch eine Art Goldstandard, welcher auf Korrektheit überprüft werden musste.

Die Klassifikation wurde nicht auf der Ebene einer gesamten Bewertung umgesetzt, denn die Satzebene wurde für geeigneter befunden. In Sätzen ist oft nur eine Polarität enthalten, wohingegen bei einem kompletten Dokument dies seltener der Fall ist.[29] Zusätzlich ist es möglich, dass für den Klassifikator das Problem auftritt, dass sich erlernte Beispiele aus beiden Kategorien überschneiden könnten. In einer insgesamt positiven Bewertung könnte auch ein negativer Satz vorkommen, ebenso andersherum. Dies führt zu widersprüchlichen Lernbeispielen. Für die satzweise Klassifikation wurden die Dokumente in Sätze aufgeteilt und dementsprechend auch satzweise markiert. Anschließend mussten die Sätze manuell korrigiert werden. Es ist sehr selten der Fall, dass beispielsweise eine Ein-Stern-Bewertung ausschließlich negative Sätze enthält.

3.2 Erstellung des Goldstandards

Um die Goldstandards für die Domänen Handy und Notebook zu erstellen, wurden zuerst Reviews zu Produkten dieser Art von dem Online-Markt Amazon extrahiert. Für die Domäne Handy waren es 21 und für die Domäne Notebook 33 zufällig gewählte Produkte. Von diesen Geräten wurden jeweils die ersten 10 hilfreichsten Bewertungen extrahiert. Außerdem wurde die soeben beschriebene automatische Markierung durchgeführt, das heißt, Sätze entsprechend ihrer Sternbewertungen als positiv (fünf bis vier Sterne), objektiv (drei Sterne) oder negativ (ein bis zwei Sterne) markiert. Anschließend wurden die Markierungen der Sätze manuell kontrolliert und gegebenenfalls korrigiert. Die in Abbildung 4 dargestellte Verteilung der Sätze auf die einzelnen Kategorien entstand:

[28] www.amazon.de
[29] Im Abschnitt 2.2.2 auf Seite 19 werden die einzelnen Ebenen des Opinion Mining erläutert.

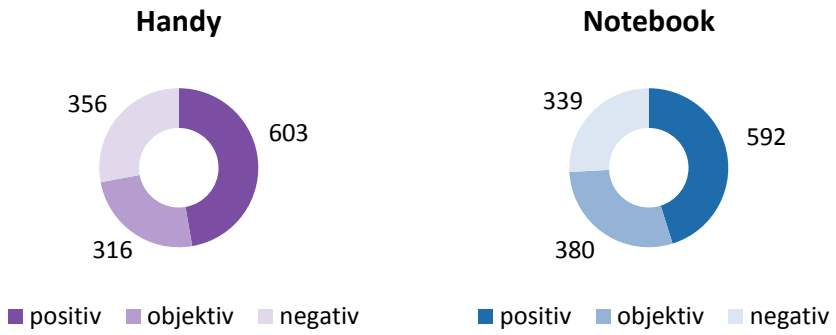

Abbildung 4: Verteilung nach manueller Markierung

Ursprünglich sollte jeder Satz der Bewertungen für die Goldstandards verwendet werden, jedoch wurde dieser Ansatz während der Erstellung des Datensatzes verworfen. Grund hierfür war u. a., dass Sätze, welche positive und negative Äußerungen enthielten, nicht praktikabel für den Klassifikator sind, da er somit widersprüchliche Regeln erlernen würde. Außerdem wäre eine eindeutige Markierung unter diesen Umständen auch nicht möglich gewesen.

Bei der manuellen Untersuchung der Sätze ist unter anderem aufgefallen, dass Bewertungen häufig in einer Art Auflistung geschrieben werden. Fehlerbeschreibungen dagegen folgen diesem Schema nur teilweise. Meist wird mit einer Gesamtaussage, wie beispielsweise „Das Display ist unheimlich schlecht.“ begonnen. Und anschließend folgt eine satzübergreifende Erklärung dafür. Diese Fälle könnten eventuell problematisch für die satzweise erfolgende Klassifikation sein, da diese Sätze keine Aussage enthalten und ohne Kontext keinen Sinn ergeben. Aus diesem Grund wurden diese ebenso weggelassen.

Bereits nach der Extraktion wurde deutlich, dass ein Ungleichgewicht zwischen positiven und negativen bzw. objektiven Bewertungen vorlag. Positive Äußerungen wurden wesentlich häufiger gefunden. Dies stellte die anfängliche Annahme, dass ein Klassifikator mit der realen Verteilung der Sätze auf die Kategorien trainiert werden sollte als nicht praktikabel dar. Nachdem jedoch wesentlich mehr positive Sätze gefunden wurden und bei einem ersten Testdurchlauf für das Erstellen eines Modells mit dem Klassifikator unbefriedigende Ergebnisse erzielt wurden, wurde ab einer Anzahl von ca. 1.000 gezielt versucht eher negative und objektive Sätze den Goldstandards hinzuzufügen. Dass dies sinnvoll war, wird in Kapitel 5 der Evaluation gezeigt. Zusätzlich barg das manuelle Überprüfen der Markierungen der Sätze einen hohen Zeitaufwand, was ein weiterer Grund für die Verwendung einer Teilmenge der extrahierten Bewertungen für die Goldstandards war.

Der Punkt der Subjektivität ist dringend anzuführen, da dieser auch in dieser Arbeit beim Erstellen des Goldstandards zum Ausdruck kam. Nicht jeder Mensch ordnet einen Satz in die gleiche Kategorie ein bzw. nimmt positive und negative Ausdrücke unterschiedlich stark wahr.

Die Kategorie objektiv wird für diese Goldstandards als das Fehlen von Meinungen in den Sätzen verwendet. Das heißt Aussagen, wie: „Ich startete erst mal das Notebook“ sind hier eingeordnet. Sobald dagegen ein subjektives Wort, welches eine Meinung ausdrückt, in einem Satz vorgekommen ist, wurde dieser in die jeweils passende Kategorie eingeordnet. Die neutrale Kategorie hätte auch für Sätze verwendet werden können, in denen sich die positiven und negativen Aussagen ausgleichen. Jedoch käme hier wieder das Problem der widersprüchlichen Trainingsbeispiele für den Klassifikator zum Tragen.

3.3 Erstellung eines größeren Datensatzes

Durch die bereits erwähnten Schwierigkeiten wurden die erstellten Goldstandards eher kleinere Datensätze. Darum wurde ein zweiter wesentlich größerer, automatisch markierter Datensatz für jede Domäne erstellt. Dafür wurden wieder Bewertungen von der Plattform Amazon extrahiert. Für diese Datensätze wurden nur Bewertungen verwendet, welche einen Stern und fünf Sterne besaßen. Hier gab es anschließend keine manuelle Überprüfung der Markierungen, da von der Hypothese ausgegangen wurde, dass in ihnen fast nur positive bzw. negative Sätze vorkamen. Da das maschinelle Lernen ein statistisches Verfahren ist, kam die Annahme hinzu, dass eine größere Anzahl an Trainingsbeispielen bessere Ergebnisse erzielen könnte, selbst wenn diese nicht korrekt markiert waren. Falls sich Trainingsbeispiele widersprechen sollten, könnte dies durch die Masse an Daten wieder korrigiert werden.

Für den zweiten Datensatz wurden 77 Produkte für die Domäne Handy und 69 Produkte für die Domäne Notebook zufällig gewählt. Es wurden jeweils die ersten zehn Seiten der fünf-Stern-Bewertung bzw. ein-Stern-Bewertung - sofern vorhanden - extrahiert. Begonnen wurde mit der Anordnung der Produkte nach Beliebtheit. Nachdem sehr wenig negative Bewertungen gefunden wurden, wurde gezielt nach Produkten mit solchen gesucht. Für die Domäne Notebook wurden aufgrund fehlender negativer Bewertungen zusätzlich Reviews über Netbooks extrahiert. Trotz der gezielten Suche nach negativ bewerteten Produkten wurde folgende in Abbildung 5 dargestellte Verteilung der Sätze erhalten.

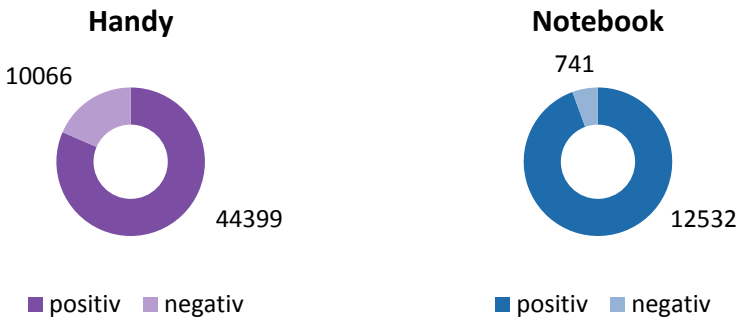

Abbildung 5: Verteilung der automatisch markierten Datensätze

Im Durchschnitt ist für beide Datensätze zu erkennen, dass für den Goldstandard aus Abbildung 4 eine Verteilung 8:1:1 (positiv-objektiv-negativ) bzw. für die automatisch markierten Datensätze aus Abbildung 5 eine Verteilung von 9:1 (positiv-negativ) zu finden ist.

3.4 Eigenschaften der Datensätze

In Tabelle 1 sind die Eigenschaften der entstandenen Datensätze dargestellt.

Datensätze	Anzahl der Sätze	Ø-Wörter pro Satz	Anzahl der Wörter	Ø - Länge der Wörter	Kategorien
Automatisch markierter Datensatz Domäne Handy	**54.465** Sätze	**16** Wörter	**897.525** Wörter	**5** Zeichen	**P, N**
Automatisch markierter Datensatz Domäne Notebook	**13.274** Sätze	**16** Wörter	**222.772** Wörter	**5** Zeichen	**P, N**
Goldstandard Handy	**1.275** Sätze	**15** Wörter	**19.671** Wörter	**5** Zeichen	**P, O, N**
Goldstandard Notebook	**1.311** Sätze	**15** Wörter	**20.164** Wörter	**5** Zeichen	**P, O, N**

Tabelle 1: Übersicht über die Anzahl der Worte und Sätze innerhalb der vier erstellten Datensätze

Ein in den Datensätzen enthaltenes Wort hat eine durchschnittliche Länge von fünf Zeichen. Dies kann eventuell hilfreich für die Einstellungen des Klassifikators bei der Durchführung der Evaluation[30] sein. Ein Satz besteht in den Goldstandards im Durchschnitt aus 15 Wörtern und in den automatisch markierten Datensätzen aus 16 Wörtern. Die extrahierten und gegebenenfalls korrigierten Da-

[30] Die Einstellungen des Klassifikators werden ausführlich in Abschnitt 4.1 erläutert.

tensätze wurden anschließend in Trainings- und Testmengen für den Klassifikator aufgeteilt.[31]

3.5 Erkenntnisse nach der Extraktion der Daten

Bei der Extraktion der Datensätze wurde festgestellt, dass wesentlich weniger Bewertungen für Notebooks als für Handys auf Amazon zu finden sind. Eine mögliche Begründung hierfür könnte sein, dass das Produkt Notebook von Kunden gezielter gekauft wird, da es eine größere Anschaffung und vorwiegend ein Produkt des Nutzens anstatt der Freizeitbeschäftigung ist. Eventuell informieren sich potentielle Kunden bei solch einer Anschaffung genauer über die einzelnen Eigenschaften und deren Auswirkungen auf den täglichen Gebrauch. Ebenso ist dies eine Produktkategorie, welche kaum essentielle Erneuerungen erlebt. Damit ist gemeint, dass allgemein bekannt ist, wie sich die Größe des Arbeitsspeichers auf die Arbeitsweise des Notebooks auswirkt, ebenso sind sich die meisten im Klaren darüber, welche Vorteile ein spiegelndes bzw. mattes Display hat. Ein weiteres Problem bei diesem Produkt ist, das häufig nur Bewertungen zu Produktlinien verfügbar sind, da die einzelnen Notebooks sich lediglich in Leistungskennzahlen unterscheiden, weswegen generell wesentlich weniger Bewertungen vorhanden sind.

Handys dagegen erleben derzeit einen Aufschwung. Sie dienen nicht nur zur Zweckerfüllung, sondern auch als freizeit-füllendes Gerät. Mit den Smartphones und der „neuen“ Technologie der Touchdisplays entstanden Geräte, welche komplett neue Eigenschaften in Handhabung und Leistung besitzen. Noch ist nicht allgemein bekannt, welche Auswirkungen ein Prozessor auf ein Android-Betriebssystem hat, weswegen diese Produkte ohne spezielles Vorwissen gekauft werden. Damit kann eventuell begründet werden, warum für diese Produkte wesentlich mehr Bewertungen und demnach auch mehr negative Bewertungen zu finden waren, da Nutzer solche Geräte eher aus Neugier kaufen.

Insgesamt stellte es sich jedoch als schwierig dar, negative Bewertungen zu den Produkten der Domänen Handy und Notebook zu finden. Dies könnte u. a. an der Nutzergruppe der Amazon-Plattform liegen. Menschen, die dort ein Produkt kaufen sind mit den Möglichkeiten des Internets vertraut. Sie informieren sich höchstwahrscheinlich ausgiebig vor einem Produktkauf über die Erfahrungen von

[31] In Abschnitt 4.2.2, Tabelle 3 sind diese Trainings- und Testmengen dargestellt.

anderen Nutzern und dementsprechend gibt es eventuell weniger Fehlkäufe. Negative Bewertungen hatten zudem häufig den Inhalt von Serienfehlern oder sogenannten „Montagsgeräten“. Dies sind jedoch nur Erklärungsversuche, weshalb so wenig negative Bewertungen verfasst werden.

3.6 Zusammenfassung

In diesem Kapitel wurde dargestellt wie die einzelnen Bewertungen für die Datensätze aus dem Web extrahiert und satzweise markiert wurden. Eine Beschreibung der Erstellung der Goldstandards der Domänen Handy und Notebook wurde ebenfalls gegeben. Es wurde festgestellt, dass am häufigsten positive Bewertungen zu finden waren. Zusätzlich wurde eine Übersicht über die einzelnen Eigenschaften der Datensätze gegeben. Sie besitzen eine durchschnittliche Satzlänge von 15 bis 16 Wörtern, welche durchschnittlich aus fünf Zeichen bestehen. In dem folgenden Kapitel werden die erstellten Datensätze als Trainings- und Testmengen für den Klassifikator angewendet und deren Ergebnisse ausgewertet.

4 Evaluation

In diesem Kapitel wird beschrieben, wie ein Klassifikator aus dem Bereich des maschinellen Lernens nicht nur zum Klassifizieren von Sätzen hinsichtlich ihrer Polaritäten verwendet wird, sondern wie dieser außerdem in seinen möglichen Einstellungen optimiert wird. Auch die einzelnen Datensätze, welche in dem vorherigen Kapitel erstellt wurden, sollen daraufhin untersucht werden, ob sie für diese Methode geeignet sind. Im ersten Abschnitt 4.1 wird ein kurzer Überblick über die vorhandenen Überlegungen vor der Umsetzung gegeben. Der zweite Abschnitt beschreibt den ersten Schritt zum Finden der besten Einstellung. Hier konnten einige Datensatzkombinationen für weitere Durchführungen ausgeschlossen werden. Im Abschnitt 4.2 werden mit Hilfe der bisher besten Trainingsmengen, der manuellen bzw. automatisch markierten Datensätze auf alle möglichen bzw. sinnvollen Einstellungsvarianten des Klassifikators durchgetestet. Die besten Einstellungen werden in Abschnitt 4.4 auf die anderen Datensätze ebenso angewendet und deren Ergebnisse miteinander verglichen. In Abschnitt 4.5 wird versucht mit Hilfe von Datensatzkombinationen noch bessere Ergebnisse zu erzielen. Der darauffolgende Abschnitt stellt eine kurze Übersicht der bisherigen Ergebnisse dar. Daraufhin wird in Abschnitt 4.6 ein weiteres Mal versucht die Genauigkeit zu verbessern indem ein Threshold für das Klassifizieren eingefügt wird. Im Abschnitt 4.7 wird die Übertragbarkeit der erstellten Modelle auf andere Domänen getestet. Abschließend wird die in diesem Kapitel evaluierte Methode mit einem Klassifikator, welcher auf einem Wörterbuch und nicht auf statistischen Daten basiert, verglichen.

4.1 Theoretische Herangehensweise

Um zu überprüfen ob ein auf Statistik basierter Klassifikator für das Klassifizieren von Sätzen in die Kategorien positiv und negativ geeignet ist, reicht es nicht aus nur geeignete Trainingsmengen zu bilden. Ebenso müssen die möglichen Einstellungen geprüft und die optimalen herausgefunden werden.

Für die Klassifikation wurde ein Klassifikator des Palladian Toolkit verwendet, welches einige Klassifikatoren, die mit N-Grammen[32] arbeiten, zur Verfügung stellt. Verwendet wurde ein wörterbuchbasierter Klassifkator. Dieser erlernt anhand der Trainingsbeispiele die Wahrscheinlichkeiten für die Zugehörigkeit der einzelnen N-Gramme zu den vorhandenen Kategorien. Dafür wird ein Wörterbuch durch zählen und normalisieren der Kookkurrenzen eines N-Gramms einer Kategorie in der Trainingsphase erstellt. Außerdem werden für jedes N-Gramm die Wahrscheinlichkeiten dafür gespeichert, zu wie viel Prozent es in die vorhandenen Kategorien gehört. Die Summe dieser muss eins ergeben, wie die weiter unten abgebildete Formel darstellt. Bei der Klassifikation eines neuen Dokumentes werden aus diesem ebenso N-Gramme generiert und mit den im Wörterbuch vorhandenen N-Grammen verglichen. Die Kategorie welche am wahrscheinlichsten für ein gefundenes N-Gramm ist, wird daraufhin auch dem neuen Beispiel zugeordnet. (Urbansky, et al. 2011)

$$CategoryProbability(category, document) = \sum_{n \in N\,url} relevance(n, category)$$

Die erstellten Datensätze, aus dem vorherigen Abschnitt wurden nun in unterschiedlichen Aufteilungen als Trainings- und Testmengen kombiniert. Genaue Zusammensetzungen sind im Abschnitt 4.2.2 aufgeführt. Diese Kombinationen sollten ursprünglich mit Hilfe des „Dataset Managers“ des Palladian Toolkits der TU Dresden umgesetzt werden. Jedoch konnten nicht alle damit erstellt werden. Das manuelle Zusammenfügen wurde aus diesem Grund für die Erstellung überschneidungsfreier Trainings- und Testmengen notwendig.

Die Evaluation wurde in fünf einzelne Abschnitte unterteilt. Zwischenergebnisse aus diesen Phasen wurden für die jeweils darauf folgenden verwendet. Eine Übersicht über den Ablauf ist in Abbildung 6 dargestellt.

[32] N-Gramme sind zusammenhängende Zeichenketten der Länge n.

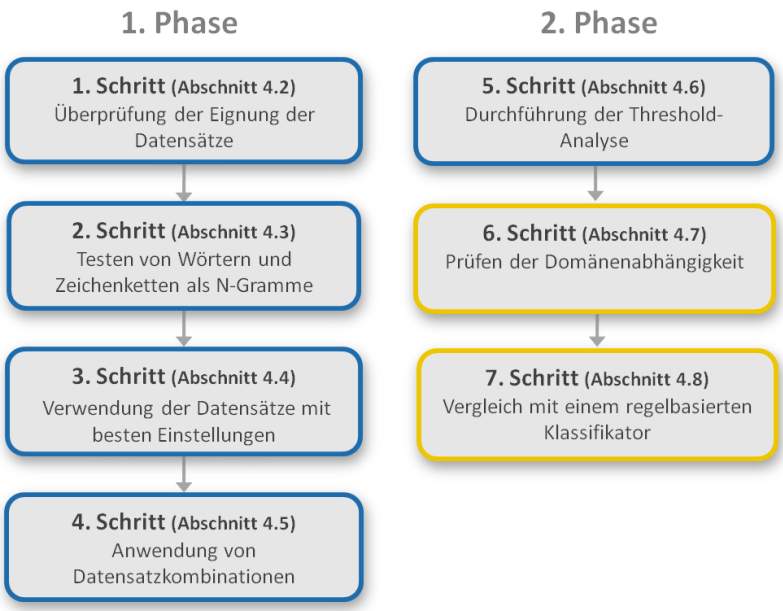

Abbildung 6: Die einzelnen Phasen der Evaluation

In der ersten Phase wurden unterschiedliche Kombinationen von Trainings- und Testmengen erstellt und auf deren Eignung zum Klassifizieren von Sätzen überprüft. Es wurde damit begonnen, die vier naheliegenden Einstellungen des Klassifikators zu testen, damit ein erster Überblick über die Möglichkeiten entsteht. Anschließend wurden in der zweiten Phase die besten Einstellungen für die Klassifikation gesucht. Dies unterteilt sich in zwei Abläufe, es werden ganze Wörter als N-Gramme bzw. Zeichenketten als N-Gramme verwendet. Im dritten Durchlauf der Evaluation wurden alle übrig gebliebenen Datensätze der ersten Phase auf die vier ermittelten Einstellungen der vorherigen Phase angewendet. Für das Erzielen von möglicherweise besseren Ergebnissen wurden im vierten Teil der Evaluation die unterschiedlichen Datensätze miteinander kombiniert und als Trainingsmengen verwendet.

Nachdem dieser Test durchgeführt wurde, stellte sich die Frage ob das Einfügen einer Grenze beim Klassifizieren der Sätze bessere Ergebnisse erzielen könnte. Ein Threshold wurde gesetzt, damit nur Sätze eingeordnet wurden, bei welchen sich der Klassifikator zu einer gewissen Konfidenz sicher war, dass der Satz auch dorthin gehörte. Dafür wurde die Methode welche die Evaluationsdaten des Klassifikators berechnet nachimplementiert und auf diese Anforderungen angepasst.

Nachdem die fünf ersten Phasen der Evaluation abgeschlossen waren, wurden die gefundenen Modelle auf Domänenabhängigkeit getestet. Ursprünglich sollten die Trainingsmodelle der beiden Domänen Handy und Notebook untereinander an-

gewendet werden. Da diese im Laufe dieser Untersuchung auch zusammengefügt als Trainingsmenge zum Einsatz kamen, wurde ein weiterer Datensatz benötigt. Hier erwies es sich als hilfreich, dass in einer weiteren wissenschaftlichen Arbeit ein ähnliches Thema bearbeitet wurde und ebenso einen Datensatz mit positiv und negativ markierten Sätzen erstellt hatte (Rybina 2012). Dieser Datensatz war technikfremd und entstammte aus Blogs, Wikipedia-Einträgen und Nachrichtenartikel.

Abschließend werden die Ergebnisse der besten Modelle mit den Ergebnissen eines wörterbuchbasierten Klassifikators verglichen. Hierfür wird als weiterer Klassifikator der „Sentiment Classifier“ verwendet. Dieser basiert auf dem Wörterbuch der Universität Leipzig[33].

4.2 Überprüfung der Eignung der Datensätze

Für die erste Phase der Evaluation wurden zufällig vier Einstellungsarten für den Klassifikator gewählt. Ziel für diese Phase war es, herauszufinden welche Kombinationen von Trainings- und Testmengen sinnvoll für das Kategorisieren von Sätzen sind. Zusätzlich sollte ein Überblick über die Möglichkeiten des Klassifikators erarbeitet werden.

4.2.1 Verwendete Einstellungen

Der verwendete Klassifikator hat zwei Einstellungsmöglichkeiten für das Anlegen des Wörterbuches. Als N-Gramme können sowohl Zeichenketten variabler Längen als auch einzelne Wörter bzw. zusammenhängende Wörter verwendet werden.

Um im ersten Durchlauf die Methodenreichweite des Klassifikators möglichst gut abzudecken, wurden folgende vier Einstellungen verwendet, welche in Tabelle 2 dargestellt sind:

Wort - N - Gramme	Zeichen - N - Gramme
Unigramm	3-10 Zeichen
1 - 3 Wörter	3-30 Zeichen

Tabelle 2: Übersicht über die zu Beginn verwendeten Einstellungen des Klassifikators

[33] Eine Beschreibung des Wörterbuches ist in Abschnitt 2.4.1 zu finden.

Unigramme wurden gewählt, da davon ausgegangen wird, dass bestimmte Wörter nur in positiven Sätzen und andere wiederum nur in negativen Sätzen vorkommen. Beispiele hierfür sind die allgemein bekannten Wörter „toll“ oder „schön“ für positive Sätze sowie „schlecht“ oder „blöd“ für negative Sätze. Die zweite Einstellung 1-3 Wörter wurde genutzt, um eventuell Phrasen wie, „nicht so gut“ oder „funktioniert nicht“ erkennen zu können. Für die Einstellung der Zeichenketten wurde mit der Annahme begonnen, dass ein Wort im Durchschnitt aus 3-10 Zeichen besteht.[34] Damit das Suchfenster nicht nur einzelne Wörter betrachtet, wurde diese Annahme zusätzlich auf 3-30 Zeichen erweitert, um ebenso eventuell 1-3 Wörter erfassen zu können.

4.2.2 Verwendete Datensätze

Während der ersten Durchführung der Evaluation kamen verschiedenste Kombinationen der vorher erstellten Datensätze zum Einsatz. Neben den realen Verteilungen der Sätze auf die Kategorien der Goldstandards, welche in Abbildung 4, Seite 41 zu finden sind, wurden zusätzlich Datensätze erstellt, welche in jeder Kategorie die gleiche Anzahl an Sätzen enthalten. Der Grund hierfür war, den Klassifikator so zu trainieren, dass eine Kategorie nicht automatisch eine höhere Wahrscheinlichkeit beim Einordnen von Testbeispielen erhält, nur weil der Klassifikator davon mehr Trainingsbeispiele erlernt hat. Die Basis für die Anzahl der Sätze pro Kategorie war die Anzahl der jeweils wenigstens Beispielsätze. Bei der Domäne Handy wurden dementsprechend auf der Basis der Satzanzahl der Kategorie objektiv mit 316 Sätzen, auch für die Gleichverteilung jeweils 316 Sätze je Kategorie verwendet. Bei der Domäne Notebook dagegen wurden 339 Sätze pro Kategorie genutzt, da die negativen Sätze mit einer Anzahl von 339 am wenigsten vorhanden waren. Analog wurde dies auf die Datensätze der automatischen Markierung übertragen, deren Verteilungen in Abbildung 5, Seite 43 zu finden sind.

Zusätzlich wurden Varianten der Goldstandards nur mit den Kategorien positiv und negativ erstellt, da die Datensätze mit automatischer Markierung nur diese beiden Satzkategorien enthalten, jedoch auf den Datensätzen der Goldstandards getestet werden.

Da die Goldstandards der Domänen Handy und Notebook eher kleinere Datensätze sind, wurden diese zu einem zusammengefügt und ebenso als Trainings- bzw.

[34] Die durchschnittliche Wortlänge von fünf Zeichen in den Datensätzen wurde hier in die Annahme mit einbezogen. Eine Übersicht der Eigenschaften der Datensätze ist in Abschnitt 3.4, Seite 46 zu finden.

Testmenge verwendet. Die Aufteilung dieses Datensatzes ist in Abbildung 7 dargestellt.

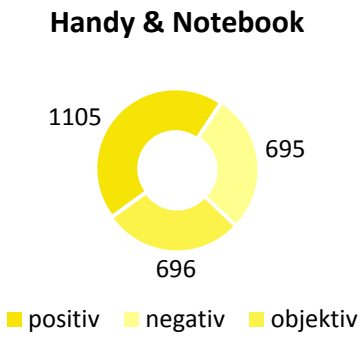

Abbildung 7: Reale Verteilung der Sätze auf die einzelnen Kategorien in einem beide Goldstandards enthaltenden Datensatz

Auch dieser Datensatz wurde mit einer gleichen Anzahl der Sätze pro vorhandene Kategorie verwendet. Als Basis wurde dafür die Kategorie mit den wenigsten Beispielsätzen der Bereiche Handy und Notebook verwendet. Dies war die Kategorie objektiv der Domäne Handy mit 316 Sätzen. Das heißt, es wurden aus den drei vorhandenen Kategorien jeder Domäne jeweils 316 Sätze für den zusammengesetzten Datensatz verwendet.

In Tabelle 3 ist eine Übersicht über die vorhandenen Datensätze gegeben. Zusätzlich werden in dieser Tabelle Namen für die einzelnen Varianten der Datensätze vergeben, welche sich aus den Anfangsbuchstaben der jeweiligen Eigenschaft bilden lassen.

Name des Datensatzes	Domäne	Art der Markierung	Anzahl Kategorien	Verteilung der Sätze auf die Kategorien	Anzahl Sätze
HA-2r	Handy	automatisch	2	realverteilt	Pos.: 44.399 Neg.: 10.066
HA-2g	Handy	automatisch	2	gleichverteilt	Pos.: 10.066 Neg.: 10.066
NA-2r	Notebook	automatisch	2	realverteilt	Pos.: 12.532 Neg.: 741
NA-2g	Notebook	automatisch	2	gleichverteilt	Pos.: 741 Neg.: 741
HG-3r	Handy	manuell (Goldstandard)	3	realverteilt	Pos.: 603 Obj.: 316 Neg.: 356
HG-3g	Handy	manuell (Goldstandard)	3	gleichverteilt	Pos.: 316 Obj.: 316 Neg.: 316
HG-2r	Handy	manuell (Goldstandard)	2	realverteilt	Pos.: 603 Neg.: 356
HG-2g	Handy	manuell (Goldstandard)	2	gleichverteilt	Pos.: 356 Neg.: 356
NG-3r	Notebook	manuell (Goldstandard)	3	realverteilt	Pos.: 592 Obj.: 380 Neg.: 339
NG-3g	Notebook	manuell (Goldstandard)	3	gleichverteilt	Pos.: 339 Obj.: 339 Neg.: 339
NG-2r	Notebook	manuell (Goldstandard)	2	realverteilt	Pos.: 592 Neg.: 339
NG-2g	Notebook	manuell (Goldstandard)	2	gleichverteilt	Pos.: 339 Neg.: 339
HNG-3r	Handy & Notebook	manuell (Goldstandard)	3	realverteilt	Pos.: 1.105 Obj.: 695 Neg.: 696
HNG-3g	Handy & Notebook	manuell (Goldstandard)	3	gleichverteilt	Pos.: 632 Obj.: 632 Neg.: 632
HNG-2r	Handy & Notebook	manuell (Goldstandard)	2	realverteilt	Pos.: 1.105 Neg.: 696
HNG -2g	Handy & Notebook	manuell (Goldstandard)	2	gleichverteilt	Pos.: 676 Neg.: 676

Tabelle 3: Übersicht über die vorhandenen Varianten der Datensätze

4.2.3 Durchführung

Für das Testen der vier Einstellungen des Klassifikators wurden unterschiedlichste Kombinationen von Trainings- und Testmengen aus den zuvor beschriebenen Datensätzen verwendet.

Die oben beschriebenen Goldstandards wurden in Trainings- und Testmengen unterteilt. Immer 80% dienten zum Anlernen des Klassifikators und 20% zum Prüfen der Datensätze. Es wurde davon ausgegangen, dass so viele Beispiele wie möglich für das Training verwendet werden sollten. Zur Bekräftigung dieser Hypothese wurde der zusammengefügte, gleichverteilte Goldstandard (HNG-3g) in unterschiedlich große Trainingsmengen aufgeteilt. Begonnen wurde mit 40% des Datensatzes als Trainingsmenge. Die größte Trainingsmenge bestand aus 80% des Datensatzes. Für einen möglichen Vergleich der Precision-. Recall- und F-Maß-Werte wurde durchgehend die gleiche Testmenge von 20% des Datensatzes verwendet.[35] Eine Verteilung von 90% für die Trainingsmenge und 10% als Testmenge, wurde nicht verwendet, da die Beispielmenge dann zu klein gewesen wäre. Die Genauigkeit des Modells eines Klassifikators wird von den Testbeispielen ebenso beeinflusst, weswegen auch diese möglichst viele sein sollten. In Abbildung 8 ist zu sehen, dass das F-Maß, bis auf bei der Größe der Trainingsmenge von 50% des Datensatzes, kontinuierlich ansteigt: Warum das F-Maß bei 50% abfällt ist aktuell nicht erklärbar. Eventuell sind in dieser Menge widersprüchliche Trainingsbeispiele vorhanden, mehr als bei 40% und diese werden bei einer Größe von 60% durch die anderen Daten eventuell wieder ausgeglichen.

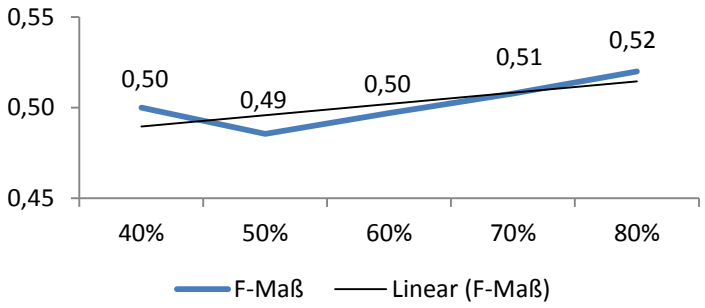

Abbildung 8: Anstieg des F-Maßes mit ansteigender Größe der Trainingsmengen

Diese Aufteilung eines Datensatzes in 80% Trainingsmenge und 20% Testmenge wurde ebenso auf die anderen Datensätze der Goldstandards übertragen.

In Tabelle 4 ist dargestellt, welche Kombinationen an Trainings- und Testmengen für die Einstellung Unigramm des Klassifikators verwendet wurden. Zusätzlich werden die durchschnittlichen, gewichteten Precision- und Recall-Werte, sowie das F-Maß angegeben.

[35] Wichtig ist hier anzumerken, dass sich nie die Beispiele der Trainingsmengen mit denen der Testmengen überschneiden dürfen.

		Trainingsmenge	Testmenge	Precision	Recall	F-Maß
Unigramme	1	**HA-2g**	**HG-2g**	0,70	0,68	0,67
	2	**HA-2g**	**HG-2r**	0,72	0,73	0,71
	3	**HA-2g**	**HG-3g**	0,45	0,67	0,53
	4	**HA-2g**	**HG-3r**	0,54	0,72	0,61
	5	**HA-2r**	**HG-2r**	0,77	0,63	0,49
	6	**NA-2g**	**NG-2g**	0,64	0,63	0,63
	7	**NA-2g**	**NG-2r**	0,64	0,64	0,64
	8	**NA-2g**	**NG-3g**	0,42	0,63	0,50
	9	**NA-2r**	**NG-2r**	0,40	0,64	0,49
	10	80% von **HG-3r**	20% von **HG-3r**	0,70	0,58	0,53
	11	80% von **HG-3g**	20% von **HG-3g**	0,57	0,55	0,54
	12	80% von **NG-3r**	20% von **NG-3r**	0,44	0,46	0,43
	13	80% von **NG-3g**	20% von **NG-3g**	0,52	0,52	0,52
	14	80% von **HNG-3g**	20% von **HNG-3g**	0,59	0,59	0,58

Tabelle 4: Verwendete Trainings- und Testmengen für die Einstellung Unigramme

Die Datensätze der automatischen Markierung wurden hier als Trainingsmenge mit Gleichverteilung (HA-2g und NA-2g), sowie mit der realen Verteilung (HA-2r und NA-2r) der Sätze pro Kategorie verwendet. Zusätzliche wurde die Trainingsmenge, welche die Gleichverteilung der Sätze enthält, auf unterschiedliche Variationen der Goldstandards getestet. Dies beinhaltete auch das Überprüfen wie die Ergebnisse durch eine unterschiedliche Kategorienanzahl beeinflusst wurden. Der Einfluss der Gleichverteilung der Sätze pro Kategorie in der Testmenge wurde ebenso überprüft, wobei dieser scheinbar marginal ist, da die Precision- und Recall-Werte anhand der Verteilung innerhalb der Testmenge gewichtet werden. Die Anwendung der realen Verteilung für die Trainingsmengen der Goldstandards wurde ebenso für die Domäne Handy und Notebook (HG-3r und NG-3r) getestet.

Für die anderen drei Einstellungen 1-3 Wörter, 3-10 Zeichen und 3-30 Zeichen- wurde ebenso eine Auswahl der Datensätze aus Tabelle 3, Seite 53 verwendet.[36]
Ergebnisse

[36] Eine genaue Übersicht über die verwendeten Trainings- und Testmengen und die entstandenen Ergebnisse sind im Anhang A zu finden

Bei der Anwendung der Datensätze auf die beschriebenen vier Einstellungen konnten einige Kombinationen von Trainings- und Testmengen ausgeschlossen werden. In diesem Abschnitt werden ausgewählte Ergebnisse ausführlich dargestellt. Ebenso werden Begründungen für das Eliminieren bestimmter Datensätze gegeben. Abschließend folgt eine Auflistung der besten Evaluationsmaße mit ihren dazugehörigen Trainings- und Testmengen sowie Methoden.

4.2.3.1 Reale Verteilung der Datensätze

Als erstes wurde überprüft ob es sinnvoll ist, die reale Verteilung der Sätze auf die drei Kategorien positiv, negativ und objektiv bzw. nur positiv und negativ auch für das Trainieren des Klassifikators beizubehalten.

In Abbildung 9 ist dargestellt, welche Precision- und Recall- Werte bei der Anwendung der automatisch markierten, real verteilten Datensätze als Trainingsmenge (HA-2r und NA-2r) erzielt wurden[37]. Die Testmenge war der jeweilige Goldstandard mit den Kategorien positiv und negativ (HG-2r und NG-2r). Die Precision- und Recall-Werte beziehen sich auf die einzelnen Kategorien. Das F-Maß dagegen ist über alle Kategorien und der Gewichtung gebildet worden. Diese Darstellung wurde gewählt, damit einerseits auf die Kategorien eingegangen werden, aber auch eine Gesamtbewertung für diese Trainingsmenge gegeben werden kann[38].

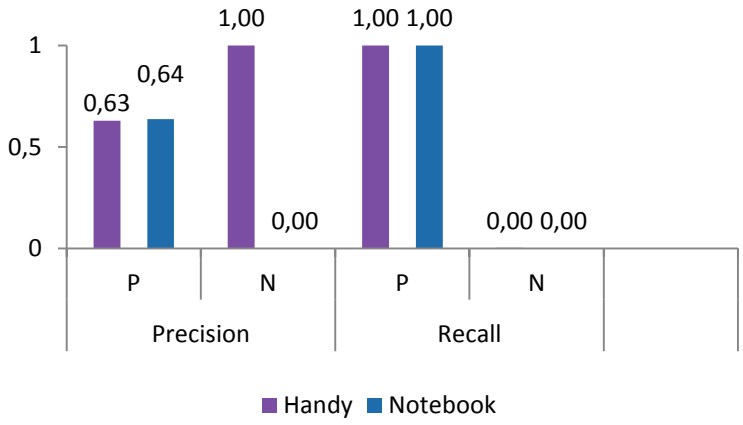

Abbildung 9: Ergebnisse der Verwendung der automatisch markierten Datensätze mit realer Verteilung der Sätze in den Trainings- und Testmengen mit der Methode Unigramme

Beim Trainieren des Klassifikators mit den Datensätzen der automatischen Markierung und dem anschließenden Testen auf den Goldstandards mit den Katego-

[37] Diese Kombination der Trainings- und Testmengen sind außerdem in Tabelle 4, Seite 63 dargestellt (Variante 5 und 9).
[38] Für folgende Abbildungen in dieser Form gilt dasselbe.

rien positiv und negativ, sowie der Einstellung Unigramme, wurde deutlich, dass es nicht sinnvoll ist mit dieser realen Verteilung zu arbeiten, da durch die Masse an positiven Sätzen, kein bzw. nur ein einzelner negativer Satz richtig erkannt wurde.

Bei dem Testen mit der realen Verteilung des Goldstandards (HG-3r und NG-3r), mit der Trainingsmenge zu je 80% und Testmenge zu je 20% und der Einstellung Unigramme[39] wurden nicht so schlechte Werte erzielt, da hier auch die Testmenge mit der realen Verteilung der Sätze pro Kategorie zusammengestellt wurde.

Dennoch wurde aufgrund dieser insgesamt schlechten Ergebnisse in den folgenden Durchführungen auf die reale Verteilung der Datensätze als Trainingsmenge verzichtet.

4.2.3.2 Gleichverteilung der Testmengen

Die Gleichverteilung der Sätze auf die einzelnen Kategorien in den Testmengen für die automatisch markierten Datensätze wurde für die folgenden Phasen ebenfalls ausgeschlossen. Es wird davon ausgegangen, dass die Gewichtung bei der Berechnung der Precision- und Recall-Werte einerseits die unterschiedlichen Verteilungen ausgleicht. Und andererseits kann im realen Anwendungsfall ebenso nicht davon ausgegangen werden, dass alle zu klassifizierenden Daten in einer gleichen Anzahl für die einzelnen Kategorien auftreten. Unterschiedliche Ergebnisse sind u. a. auch von den Beispielen innerhalb der Testmenge abhängig, das heißt, wie eindeutig diese einzuordnen sind.

4.2.3.3 Anzahl der Kategorien in den Trainings- und Testmengen

Eine weitere Kombination von Trainings- und Testmengen wurde ausgeschlossen. Eine unterschiedliche Anzahl von Kategorien in diesen beiden Mengen wurde für die folgenden Durchführungen nicht mehr verwendet. Dies war der Fall, wenn die Datensätze mit automatischer Markierung, welche nur die Kategorien positiv und negativ enthielten, auf den Goldstandards, welche wiederum die Kategorien positiv, negativ und neutral besaßen, der jeweiligen Domäne getestet wurden. Um zu überprüfen, ob dies einen Einfluss auf die Ergebnisse hat, wurde der Datensatz der automatischen Markierung der Domäne Handy (HA-2g) auf den Goldstandard der Domäne Handy (HG-2r) ohne die Kategorie objektiv angewendet. Und auf den Goldstandard der Domäne Handy mit allen drei Kategorien (HG-3r).[40] Die erziel-

[39] Diese Kombination an Trainings- und Testmengen ist in Tabelle 4, Seite 63 ebenso zu finden (Variante 10 und 12).
[40] Diese beiden Kombinationen von Trainings- und Testmengen sind ebenso in Tabelle 4, Seite 63 dargestellt (Variante 4 und 8).

ten Precision- und Recall-Werte unterschieden sich nicht. Jedoch ist das F-Maß mit 0,61 bei dem Durchgang mit einer ungleichen Anzahl an Kategorien wesentlich schlechter als bei dem Durchgang mit der gleichen Anzahl an Kategorien, mit 0,71. Dies liegt unter anderem an der unterschiedlichen Gewichtung bei der Berechnung des Wertes. Dieser wird anhand der Anzahl der Beispiele pro Kategorie im Verhältnis zu allen Beispielen berechnet. Es sind zusätzlich Sätze für die Kategorie objektiv vorhanden gewesen, was die Gesamtanzahl an Sätzen erhöht. Jedoch wurden diese bei der Einordnung nicht beachtet und somit verschlechterte sich das F-Maß. Es ist anzunehmen, dass es sich für die Domäne Notebook analog verhält. Im Folgenden werden nur noch Trainings- und Testmengen mit der gleichen Anzahl an Kategorien verwendet.

4.2.3.4 Verwendung der objektiven Kategorie

Eine offene Fragestellung im Bereich des Opinion Mining ist das Hinzunehmen einer objektiven Kategorie beim Klassifizieren nach Polaritäten. [41] Der Inhalt dieser ist u. a. auch umstritten. In einigen Forschungsarbeiten wird die objektive Kategorie für Inhalte ohne jegliche Polarität, das heißt für Fakten, verwendet. In anderen wiederum wird sie als neutrale Kategorie genutzt. Positive und negative Anteile gleichen sich aus, insgesamt ergibt sich eine neutrale Aussage. In dieser Arbeit wurde die objektive Kategorie dafür verwendet, Sätze ohne jegliche Polarität zu kategorisieren. Es gibt Forschungsansätze, die aussagen, dass das Hinzufügen einer objektiven Kategorie die Gesamtergebnisse der Kategorisierung positiv beeinflussen. Dies wird in den folgenden Ausführungen geprüft.

In Tabelle 5 sind die Ergebnisse der Evaluationsmaße dargestellt, welche für den zusammengefügten Goldstandard mit zwei (HNG-2g) bzw. drei Kategorien (HNG-3g) in den Trainings- und Testmengen entstanden. Mit einer durchschnittlichen[42] Precision von 0,74, einem Recall von 0,73 und einem F-Maß von 0,72 war der Goldstandard ohne die objektiven Sätze eine wesentlich bessere Trainingsmenge als mit der objektiven Kategorie.

[41] Weiteres dazu ist u. a. im Grundlagenkapitel im Abschnitt 2.2.1, Seite 11 zu finden.
[42] Über alle vier Haupteinstellungen, dies ist auch im Folgenden gemeint, wenn von Durchschnitt geschrieben wird.

Evaluationsmaße	ohne objektiv	mit objektiv
Precision	0,74	0,61
Recall	0,73	0,59
F-Maß	0,72	0,58

Tabelle 5: Vergleich der durchschnittlichen Evaluationsmaße des zusammengefügten Goldstandards mit und ohne der Kategorie objektiv in den Trainings- und Testmengen

In einem realen Anwendungsfall kann es aber dennoch häufig vorkommen, dass in zu klassifizierenden Sätzen nicht nur Meinungen ausgedrückt werden, sondern zwischendurch auch Fakten oder Umstände beschrieben werden. In Abbildung 10 sind daher die Ergebnisse eines Versuches zu sehen, in dem zwei unterschiedliche Trainingsmengen auf die gleiche Testmenge mit der Kategorie objektiv angewendet wurden. Die Trainingsmengen sind die gleichen, wie sie in dem Versuch, welcher in Tabelle 5 dargestellt wird, angewendet wurden. Die Trainingsmenge, welche nur auf die Kategorien positiv und negativ trainiert wurde, lieferte wesentlich bessere Precision-Werte für diese beiden Kategorien als die Trainingsmenge, welche zusätzlich objektive Sätze erlernt hat. Die gewichtete Precision über alle Kategorien dagegen fiel wesentlich schlechter aus, da diese im Verhältnis zu allen Beispielen der Testmenge berechnet wird. Auch das F-Maß wurde davon beeinflusst.

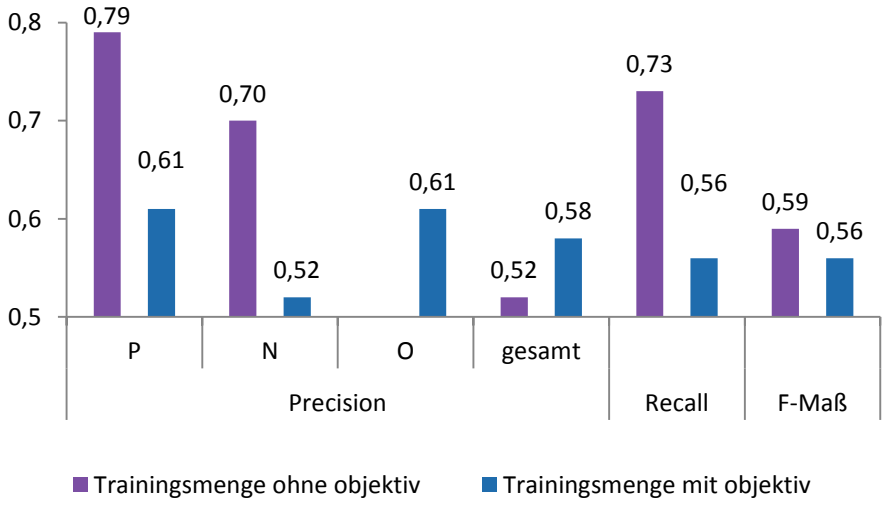

Abbildung 10: Vergleich der Ergebnisse der Trainingsmengen des zusammengefügten Goldstandards auf die Testmenge mit der Kategorie objektiv

Insgesamt kann jedoch gesagt werden, dass es eventuell trotzdem sinnvoller ist, mit nur zwei Kategorien zu trainieren und dennoch auf eventuell drei vorhandene anzuwenden.

4.2.3.5 Beste Ergebnisse

Die besten, erreichten Evaluationsmaße werden im Folgenden in Auszügen dargestellt.[43] In Tabelle 6 sind die höchsten erzielten Evaluationswerte für die Datensätze der automatischen Markierung zu sehen.[44] Der beste Precision-Wert ist 0,77 und wurde mit dem automatisch markierten Datensatz der Domäne Handy als Trainingsmenge (HA-2g) erzielt. Geprüft wurde das erstellte Modell des Klassifikators auf dem Handy-Goldstandard (HG-2r). Der beste Precision-Wert über alle Kategorien beträgt 0,73, ebenfalls mit der automatisch markierten Datensatz der Domäne Handy in Verbindungmit dem Goldstandard als Trainingsmenge. Das beste F-Maß betrug 0,73, ebenfalls mit dieser Kombination. Die besten Einstellungen im Zusammenhang mit den besten Gütemaßen sind N-Gramme aus 3-10 Zeichen sowie Unigramme.

		Wert	Trainings- und Testmenge	Einstellung
Precision	Positiv	0,77	HA-2g auf HG-2r	3-10 Zeichen
	Negativ	0,68	HA-2g auf HG-2r	Unigramme
	Gesamt	0,73	HA-2g auf HG-2r	3-10 Zeichen
Recall	Positiv	0,86	HA-2g auf HG-2r	Unigramme
	Negativ	0,57	HA-2g auf HG-2r	3-10 Zeichen
	Gesamt	0,73	HA-2g auf HG-2r	3-10 Zeichen
F-Maß	Gesamt	0,73	HA-2g auf HG-2r	3- 10 Zeichen

Tabelle 6: Die höchsten Precision- und Recall-Werte sowie F-Maße der automatisch markierten Datensätze

Die Durchführungen mit den Datensätzen der Goldstandards erzielten etwas schlechtere Werte als die automatisch markierten Datensätze als Trainingsmengen. In Tabelle 7 sind die besten Werte nach Kategorien aufgelistet. [45] Insgesamt konnten mit dem Goldstandard beider Domänen Handy und Notebook (HNG-3g) als Trainingsdatensatz die besten Precision-Werte für die Kategorien positiv und negativ erzielt werden. Auffällig ist, dass negative Sätze am schlechtesten kategorisiert werden mit einer Precision von 0,59, jedoch einem Recall von 0,76. Insge-

[43] Eine vollständige Übersicht aller Ergebnisse ist im Anhang A zu finden.
[44] Hier sind die Trainingsmengen HA-2g und NA-2g, sowie die Testmengen HG-2r und NG-2r gemeint.
[45] Hier wurden immer jeweils 80% als Trainingsmenge und jeweils 20% als Testmenge verwendet.

samt konnten positive Sätze am besten eingeordnet werden mit einer Precision von 0,67 und einem Recall von 0,79. Hinsichtlich des besten Recall erbrachte der Goldstandard der Domäne Handy (HG-3g) die besten Ergebnisse. Die Einstellung N-Gramme aus 3-10 Zeichen lieferte am häufigsten die höchsten Werte. Die besten Ergebnisse der manuell markierten Datensätze wurden mit den zusammengefügten Datensatz der Goldstandards (HNG-3g) erzielt.

		Wert	Trainings- und Testmenge	Einstellung
Precision	Positiv	0,67	HNG-3g	3-10 Zeichen
	Negativ	0,59	HNG-3g	3-30 Zeichen
	Objektiv	0,67	HG-3g	3-10 Zeichen
	Gesamt	0,63	HNG-3g	3-10 Zeichen
Recall	Positiv	0,79	NG-3g	3-10 Zeichen
	Negativ	0,76	HG-3g	1-3 Wörter
	Objektiv	0,60	HG-3g	3-10 Zeichen
	Gesamt	0,61	HG-3g	3-10 Zeichen
F-Maß		0,62	HNG-3g	3- 10 Zeichen

Tabelle 7: Die höchsten Precision- und Recall-Werte, sowie F-Maße der Goldstandards

Zusätzlich wurden einige Datensatzkombinationen für die folgenden Untersuchungen verworfen. Übrig blieben fünf Variationen von Trainings- und Testmengen, die in Tabelle 8 aufgelistet sind und in den folgenden Abschnitten verwendet werden.

Trainingsmenge	Testmenge
HA-2g	HG-2r
NA-2g	NG-2r
80% von HG-3g	20% von HG-3g
80% von NG-3g	20% von NG-3g
80% von HNG-3g	20% von HNG-3g

Tabelle 8: Kombinationen von Trainings- und Testmengen für die weiteren Schritte der Evaluation

4.3 Optimierung der Einstellungen des Klassifikators

In der zweiten Phase der Evaluation wurden die besten Einstellungen für den Klassifikator gefunden werden. Dafür wurde der zusammengefügte Goldstandard

als Datensatz (HNG-3g) genutzt. Davon wurden 80% für die Trainingsmenge und 20% für die Testmenge verwendet, da dieser in der ersten Phase der Evaluation die besten Ergebnisse erzielt hatte. Auch der automatisch markierte Datensatz der Domäne Handy erzielte hier gute Werte, jedoch ist dieser wesentlich umfangreicher und deswegen eventuell bei einigen Einstellungen sehr ressourcenaufwändig.

4.3.1 Zeichenketten als N-Gramme mit bestem Datensatz der manuellen Markierung

Begonnen wurde mit der Einstellung Zeichenketten als N-Gramme zu verwenden.[46] Hier wurde die Länge dieser Zeichenketten variiert und die gelieferten Ergebnisse verglichen. Begonnen wurde mit der minimalen Länge von einem Zeichen und der maximalen Länge von drei. Die minimale Länge wurde beibehalten und die maximale Länge von 3- 20 Zeichen verändert.

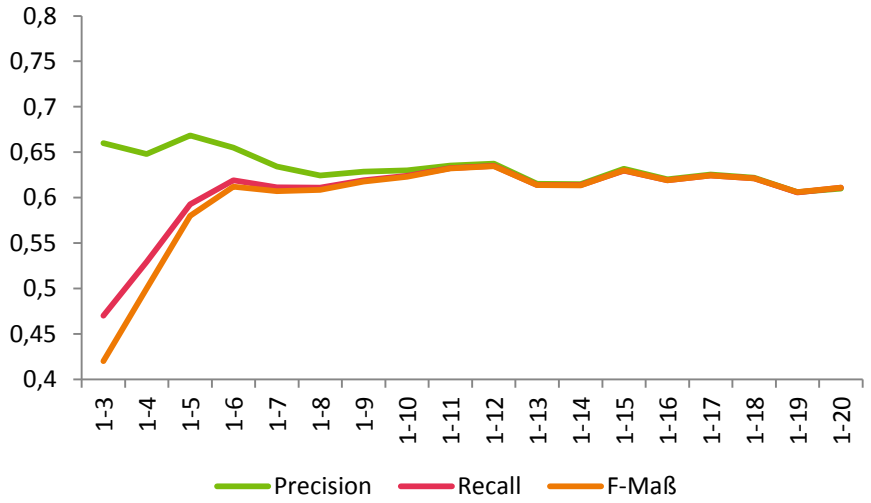

Abbildung 11: Entwicklung der Evaluationswerte innerhalb der Zeichenkettenlängen 1-20 Zeichen

In Abbildung 11 ist die Entwicklung der Evaluationswerte innerhalb der Spanne der minimalen und maximalen Zeichenkettenlängen zwischen 1-20 Zeichen abgebildet. Anfangs ist die Precision mit 0,66 wesentlich höher als der Recall mit 0,47 bzw. das F-Maß mit 0,42. Diese drei Werte näherten sich jedoch mit steigender maximaler Länge der Zeichenkette aneinander an und ab der Einstellung 1-11 Zeichen waren alle drei Werte ausgeglichen mit einer Precision von 0,63, einem Recall von 0,63 und einem F-Maß von ebenso 0,63. Geendet wurde bei 20 Zei-

[46] Eine ausführliche Übersicht aller Ergebniswerte ist im Anhang B.1 finden.

chen, da alle größeren Zeichenketten als Modell für den Klassifikator sehr speicherintensiv gewesen wären.

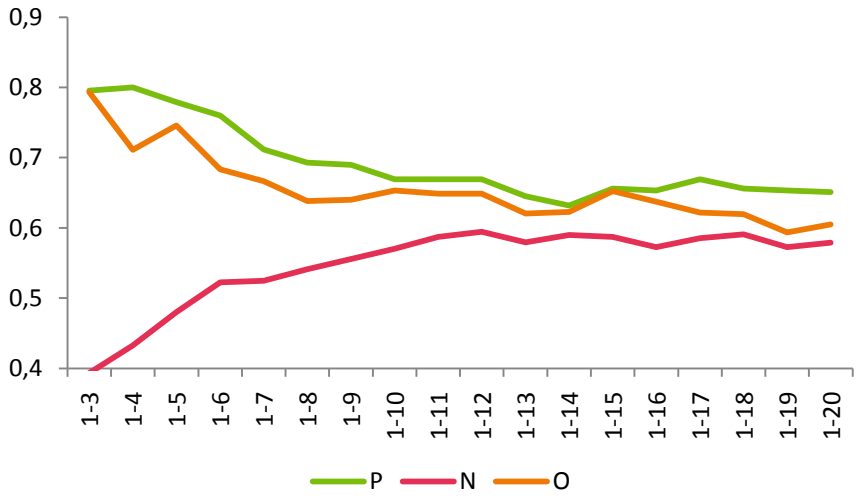

Abbildung 12: Entwicklung der Precision-Werte für die drei Kategorien innerhalb der Zeichenkettenlängen von 1-20 Zeichen

Bemerkenswert ist die Entwicklung der Precision-Werte für die einzelnen Kategorien. In Abbildung 12 ist dargestellt, dass anfangs die Sätze der Kategorien positiv und objektiv mit einer wesentlich höheren Genauigkeit (0,8 bzw. 0,79) durch das gebildete Modell des Klassifikators eingeordnet werden konnten als die negativen Sätze mit 0,39. Mit steigender maximaler Länge der Zeichenketten näherten sich diese Werte jedoch aneinander an. Die Precision für die negative Kategorie bleibt jedoch immer schlechter als die der anderen beiden Kategorien. Der beste Wert für die Precision mit 0,59 wurde u. a. bei der Länge von 1-11 und 1-12 Zeichen erreicht. Eventuell könnte das damit zusammenhängen, dass negative Meinungen nicht mit Hilfe weniger Wörter geäußert werden.

Da ein Zeichen in der deutschen Sprache nicht aussagekräftig ist, wurde anschließend die minimale Länge einer Zeichenkette auf zwei gesetzt und die maximale Länge zwischen 3 und 15 Zeichen verändert. Für diese Einstellung sind ähnliche Ergebnisse entstanden, wie in Abbildung 11. Ebenso für 3-15 und 4-15 Zeichen. Im Folgenden wird nur darauf eingegangen, welche Einstellung die besten Ergebnisse für die Evaluationsmaße erzielte.

Die Maße Precision und F-Maß werden als maßgebend beim Finden der besten Einstellung für den Klassifikator betrachtet. Auch die Genauigkeit des Einordnens von negativen Sätzen liegt im Mittelpunkt der Betrachtung, da dies für den Anwendungsfall dieser Arbeit wichtig ist. In Abbildung 13 sind jeweils die drei bes-

ten Ergebnisse für diese Werte und deren dazugehörigen Einstellungen für den zusammengefügten Goldstandard (HNG-3g) abgebildet. Der beste Wert für die Precision ist 0,67 für die Einstellung 1-5 Zeichen und das beste F-Maß ist 0,63 für die Einstellung 1-12 Zeichen. Zusätzlich beträgt die Genauigkeit für das Einordnen von negativen Sätzen hier 0,59. Dies ist der höchste Wert, welcher erreicht werden konnte. Aus diesem Grund wurden diese zwei Methoden für die nächsten Phasen der Evaluation weiterverwendet.

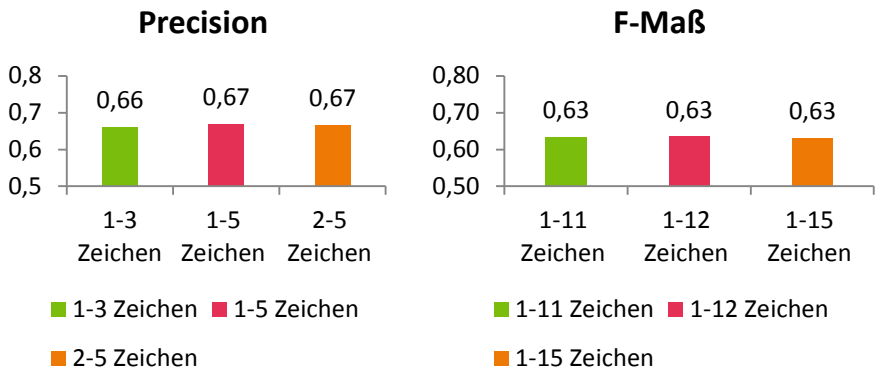

Abbildung 13: Die drei besten Werte für die Precision bzw. das F-Maß und die dazugehörigen Methoden

Um auch die wesentlich größeren Datensätze mit der automatischen Markierung bei dem Finden der besten Einstellungen zu berücksichtigen, wurde stellvertretend die Trainingsmenge HA-2g mit verschiedenen Einstellungen getestet. Die Einstellung mit den besten Ergebnissen bezüglich Precision, Recall und F-Maß wurde ebenso für die weiteren Untersuchungen hinzugenommen. Dies ist die Einstellung 6-10 Zeichen, da hier auch die Precision für negative Sätze mit 0,69 am höchsten war.

4.3.2 Wörter als N-Gramme mit bestem Datensatz der manuellen Markierung

Die Einstellung Wörter als N-Gramme wurde ebenso mit Hilfe des zusammengefügten Datensatzes der Goldstandards der Domäne Handy und Notebook (HNG-3g) in unterschiedlichen Variationen getestet. [47]

Begonnen wurde damit, die minimale Länge der N-Gramme als eins zu wählen und die maximale Länge von 2-8 variieren zu lassen. In Abbildung 14 ist zu sehen, dass alle drei Evaluationsmaße bei steigender Anzahl an Wörtern bei dem

[47] Eine ausführliche Darstellung aller Ergebnisse ist im Anhang B.2 zu finden

Wert 0,56 stagnierten. Lediglich bei der Einstellung 1-2 Wörter war die Precision ein wenig höher, mit 0,58 und der Recall sowie das F-Maß mit 0,57.

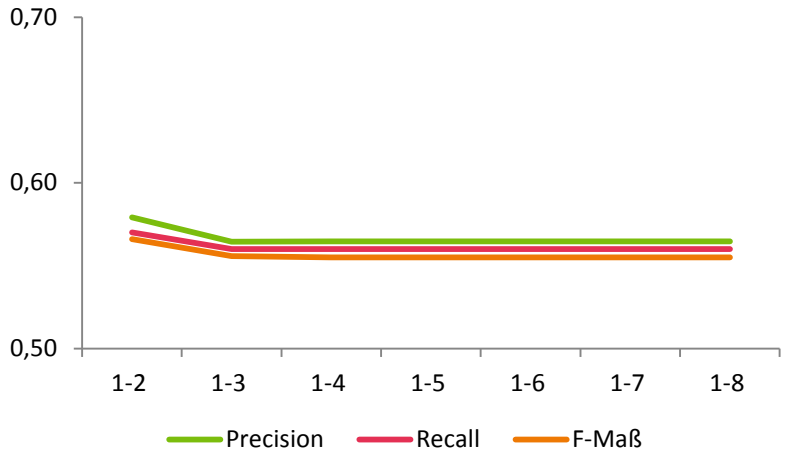

Abbildung 14: Entwicklung der Evaluationsmaße von 1-8 Wörter als N-Gramme

Anschließend wurden zusammenhängende Wörter getestet. Begonnen wurde mit genau einem Wort und dies wurde fortgeführt bis zu der Einstellung genau vier Wörter als N-Gramm. Wie in Abbildung 15 zu sehen ist, waren die Werte für die Evaluationsmaße bei einem Wort hoch (Precision 0,59, Recall 0,59 und F-Maß 0,58), sanken jedoch bei mehr zusammenhängenden Worten.

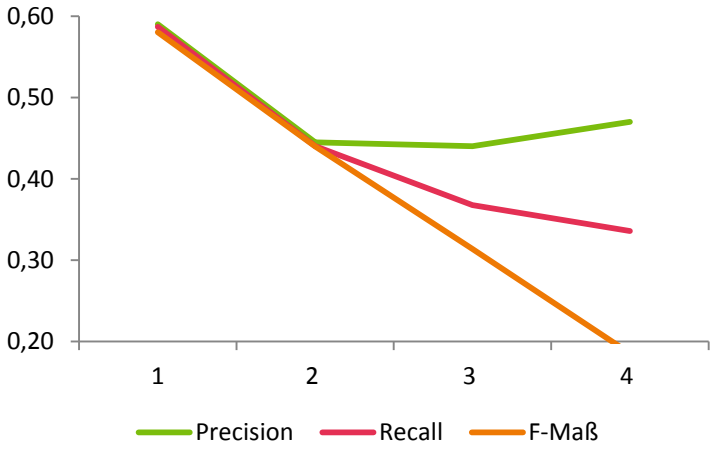

Abbildung 15: Entwicklung der Evaluationsmaß bei größer werdenden N-Grammen aus Wörtern

An dieser Stelle werden im Folgenden nur die besten Werte für die Precision bzw. für das F-Maß dargestellt, welchen in Abbildung 16 mit den dazugehörigen Einstellungen für das Erstellen eines Modells des Klassifikators, dargestellt sind.

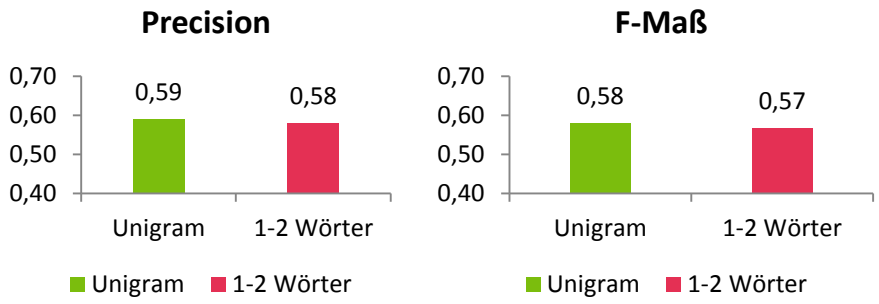

Abbildung 16: Beste Ergebnisse für Precision und F-Maß und die dazugehörigen Einstellungen

Der beste Wert für die Precision mit 0,59 wurde mit Unigrammen erzielt. Die gleiche Einstellung erzielte auch das beste F-Maß mit 0,58. Da für die folgenden Untersuchungen zwei Methoden verwendet werden sollten, wird die zweitbeste Einstellung ebenso weiterverwendet. Das Nutzen von 1-2 Wörtern als N-Gramme erzielte eine Precision von 0,58 und ein F-Maß von 0,57. Da der Hauptfokus auch auf dem Kategorisieren von negativen Sätzen liegt, sollte die Genauigkeit dafür auch besonders hoch sein. Bei Wörtern als N-Gramme stieg die Precision bei keiner Variante über 0,55. Dieser Wert wurde mit Unigramme erreicht. Nur bei vier Wörtern war dieser Wert wesentlich besser und zwar mit 0,75, jedoch wurden die Sätze der beiden anderen Kategorien sehr schlecht eingeordnet, so dass die Precision über alle drei nur 0,47 betrug. Aus diesem Grund wird diese Einstellung nicht weiterverwendet. Bei 1-2 Wörtern als N-Gramme konnten Sätze mit einer Genauigkeit von 0,54 eingeordnet werden, was im Vergleich zu den anderen Ergebnissen ein guter Wert ist.

Auch hier wurde stellvertretend für die automatische Markierung die Trainingsmenge HA-2g mit unterschiedlichen Varianten der Benutzung von Wörtern als N-Gramme getestet. Die Einstellung 1-4 Wörter wird für die weiteren Untersuchungen verwendet, da diese die besten Evaluationsmaße erreichte.

4.3.3 Ergebnisse

Es stellte sich in diesem Durchlauf der Evaluation heraus, dass die Einstellungen Unigramme bzw. 1-2 Wörter sowie 1-5 bzw. 1-12 Zeichen die besten Ergebnisse auf der Trainings- und Testmenge des zusammengefügten, gleichverteilten Goldstandards (HNG-3g) erzielten. Für den automatisch markierten Datensatz (HA-2g) sind die Einstellungen 1-4 Wörter und 6-10 Zeichen. Diese Einstellungen, welche in Tabelle 9 abgebildet sind, wurden für die folgenden Durchläufe verwendet.

Wort - N - Gramme	Zeichen - N - Gramme
Unigramm	1-5 Zeichen
1-2 Wörter	1-12 Zeichen
1-4 Wörter	6-10 Zeichen

Tabelle 9: Die sechs besten Einstellungen

4.4 Verwendung der besten Einstellungen für die Datensätze

In der dritten Phase der Evaluation wurden die Ergebnisse der ersten und zweiten Phase miteinander kombiniert. Das heißt, die übrig gebliebenen Datensätze aus Tabelle 8, Seite 61 der ersten Phase wurden mit den gefundenen besten Einstellungen der zweiten Phase getestet. Ziel für diesen Durchgang war zu überprüfen, welche Datensätze in Kombination mit einer der vier Einstellungen die besten Ergebnisse lieferten.

4.4.1 Automatisch markierte Datensätze

Die automatisch markierten Datensätze (HA-2g und NA-2g) wurden in dieser Phase auf alle vier Einstellungsvariationen, welche in der zweiten Phase der Evaluation ermittelt wurden angewendet. Testmenge war der jeweilige zur Domäne passende Goldstandard ohne die objektive Kategorie und mit realer Verteilung der Sätze (HG-2r und Ng-2r).[48]

In Abbildung 17 sind die Evaluationsmaße des Datensatzes der Domäne Handy (HA-2g) zu sehen. Die besten Werte wurden für die Einstellung 6-10 Zeichen erreicht. Hier waren die durchschnittlichen Precision- und Recall-Werte über alle drei Kategorien 0,74 bzw. 0,75. Ebenso erreichte das F-Maß einen Wert von 0,74. Für die anderen Methoden pendelten sich die Werte zwischen 0,70 und 0,73 ein.

[48] Eine detailliertere Darstellung der einzelnen Ergebnisse ist im Anhang C zu finden.

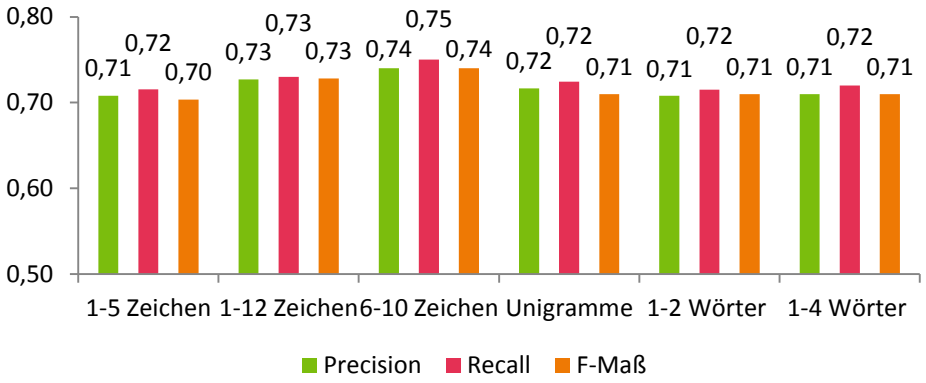

Abbildung 17: Ergebnisse des automatisch markierten Datensatz der Domäne Handy

Für die Kategorie positiv wurde die beste Genauigkeit mit der Einstellung 6-10 Zeichen mit einem Wert von 0,77 erlangt. Ebenso Negative Sätze wurden mit einer Precision von 0,69 richtig eingeordnet. Objektive Sätze konnten in dieser Domäne nicht bewertet werden, da die Kategorie für diesen Datensatz nicht existiert.

Der Datensatz der Domäne Notebook (NA-2g) als Trainingsmenge konnte die Sätze der Testmenge, dem Goldstandard Notebook (NG-2r), mit einer etwas schlechteren Genauigkeit einordnen. Die besten Werte wurden mit den Einstellungen 1-5 Zeichen und 6-10 Zeichen ermittelt. Hier betrug jeweils die Precision 0,65, der Recall 0,66 sowie das F-Maß 0,65. Die größte Genauigkeit von 0,72 beim Einordnen der positiven Satzbeispiele dieser Trainingsmenge wurde mit den Einstellungen 1-5 Zeichen und 6-10 Zeichen erreicht. Die negativen Sätze wurden in allen Einstellungen schlechter klassifiziert mit der besten Precision von 0,53.

4.4.2 Manuell markierte Datensätze

Auch die manuell markierten Datensätze, das heißt die Goldstandards, wurden mit allen vier Einstellungen aus der zweiten Phase verwendet. Hier wurden alle drei Datensätze (HG-3g, NG-3g und HNG-3g) mit einer Gleichverteilung verwendet. Davon wurden jeweils 80% für die Trainingsmengen und 20% für die Testmengen genutzt.[49]

Für den Goldstandard der Domäne Handy (HG-3g) ergaben sich sehr unterschiedliche Ergebnisse innerhalb der vier Einstellungen. Das schlechteste F-Maß mit 0,52 wurde für die Einstellung 1-5 Zeichen erreicht. Dagegen ist dort die Precisi-

[49] Eine ausführliche Übersicht über die Ergebnisse ist im Anhang auf Seite C zu finden.

on über alle drei Kategorien mit 0,63 die höchste aller vier Durchläufe. Die insgesamt besten Werte wurden mit der Einstellung 1-12 Zeichen erzielt. Die anderen Ergebnisse sind in Abbildung 18 zu sehen.

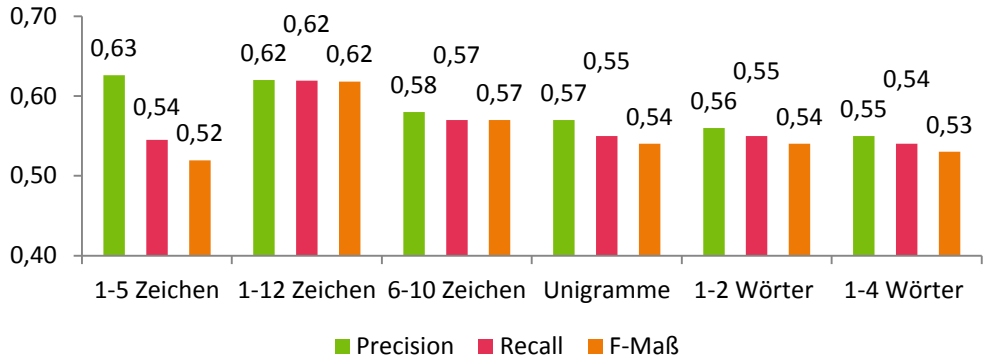

Abbildung 18: Ergebnisse des Goldstandards der Domäne Handy

Die beste Precision von 0,83 wurde für das Einordnen positiver Sätze mit der Einstellung 1-5 Zeichen erreicht. Dafür waren hier diese Werte für die negative und objektive Kategorie mit 0,46 bzw. 0,59 wesentlich schlechter. Am genauesten wurden negative Sätze mit der Einstellung 1-12 Zeichen klassifiziert, ebenso objektive Sätze.

Der Goldstandard der Domäne Notebook (NG-3g) hatte ebenso sehr unterschiedliche Werte für die Evaluationsmaße mit den vier Einstellungen. Generell ist zu sagen, dass die Recall-Werte durchgehend eher schlecht ausfielen. Die Precision-Werte dagegen waren besser. Eine Übersicht über die Werte ist in Abbildung 19 zu sehen.

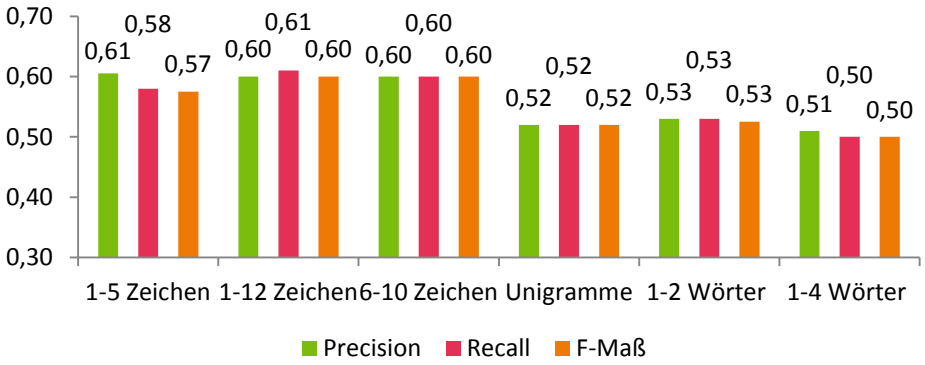

Abbildung 19: Ergebnisse des Goldstandards der Domäne Notebook

Die positiven Sätze wurden am besten mit der Einstellung von 1-5 Zeichen klassifiziert, das heißt mit einer Precision von 0,73. Ebenso objektive mit einer Genau-

igkeit von 0,6. Negative Sätze mit der Einstellung 6-10 Zeichen mit einer Genauigkeit von 0,61.

Bei dem zusammengefügten Goldstandard (HNG-3g) variierten die Werte nicht in einem so großen Maß wie bei den einzelnen Goldstandards. Die größte Genauigkeit beim Einordnen der Sätze wurde mit der Einstellung 1-5 Zeichen erlangt. Jedoch waren der Recall und das F-Maß hierbei wesentlich geringer. Eine Übersicht aller Ergebnisse ist in Abbildung 20 zu finden.

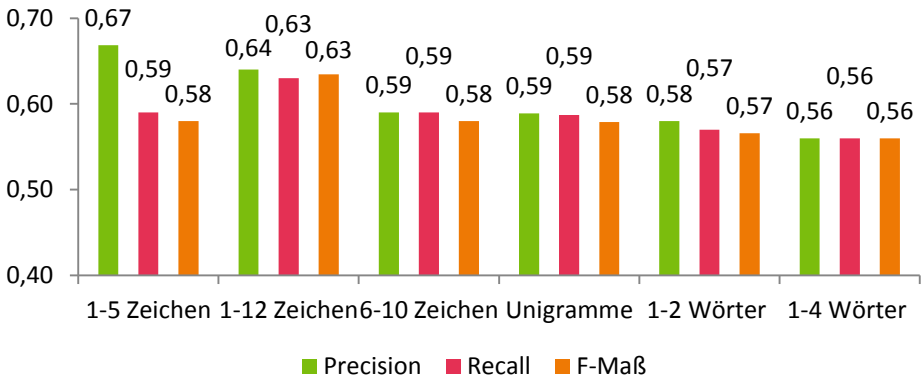

Abbildung 20: Ergebnisse des Goldstandards der Domänen Handy und Notebook

Positive Sätze wurden mit dem aus der Einstellung 1-5 Zeichen erstelltem Modell des Klassifikators am besten mit einer Precision von 0,78 eingeordnet. Ebenso objektive Sätze mit einer Präzision von 0,75. Negative Sätze dagegen mit N-Grammen aus 1-12 Zeichen mit einer Genauigkeit von 0,59.

4.4.3 Ergebnisse

Insgesamt kann gesagt werden, dass der zusammengefügte Goldstandard beider Domänen (HNG-3g) im Durchschnitt die beste Genauigkeit beim Klassifizieren der Sätze lieferte mit einer Precision von 0,61. Die Trainingsmenge aus dem Goldstandard der Domäne Handy (HG-3g) hatte eine Precision von 0,58 und die der Domäne Notebook (NG-3g) 0,56. Bessere Werte erzielten die Trainingsmengen aus den automatisch markierten Datensätzen. In der Domäne Handy (HA-2g) eine Precision von 0,72 und in der Domäne Notebook (HA-2g) 0,64.

Der Recall war im Durchschnitt ebenso für die manuell markierten Datensätze schlechter als für die automatisch markierten. Der Goldstandard der Domäne Handy (HG-3g) erlangte einen Wert von 0,56 und der der Domäne Notebook (NG-3g) 0,56. Der zusammengefügte Goldstandard (HNG-3g) war marginal bes-

ser mit einem Recall-Wert von 0,59. Die Trainingsmengen aus den automatisch markierten Datensätzen erlangten einen Recall von 0,72 für die Domäne Handy (HA-2g) und für die Domäne Notebook (NA-2g) 0,64.

Das F-Maß war dementsprechend ebenso für die automatisch markierten Datensätze im Durchschnitt höher als für die Goldstandards. Mit einem Wert von 0,72 für die Domäne Handy (HA-2g) und 0,64 für die Domäne Notebook (NA-2g).

In Tabelle 10 ist sind die Durchschnittswerte der Evaluationsmaße über alle Datensätze und alle vier Methoden zu sehen. Insgesamt kann gesagt werden, dass das maschinelle Lernen bzw. das Trainieren eines Klassifikators anhand von meinungsbehafteten Sätze durchaus möglich, wenn nicht sogar gut möglich ist. Abhängig sind die Ergebnisse einerseits von der Domäne aus der die Beispielsätze stammen und zusätzlich von den vorhandenen Kategorien.

	Precision	Recall	F-Maß
Durchschnitt	0,62	0,61	0,61

Tabelle 10: Übersicht über die durchschnittliche Evaluationsmaße aller Datensätze und aller Einstellungsmöglichkeiten

Auch die einzelnen Kategorien der Sätze weisen unterschiedliche Ergebnisse auf. In Tabelle 11 ist eine Übersicht gegeben über die durchschnittlichen Precision-Werte der einzelnen Kategorien aus allen Datensätzen und Methoden.

	Positiv	Negativ	Objektiv
Durchschnitt	0,68	0,55	0,59

Tabelle 11: Übersicht über die durchschnittlichen Precision-Werte der einzelnen Kategorien

Somit ist erkennbar, dass positive Sätze im Durchschnitt besser eingeordnet werden können, als negative und objektive. Dies spiegelte sich auch in den einzelnen Ergebnissen pro Domäne und pro Einstellung wider. Für die objektive Kategorie ist zusätzlich anzumerken, dass diese nur für die Trainingsmengen aus den Goldstandard-Datensätzen vorhanden war. Die Einstellungen, welche die besten Modelle für den Klassifikator erstellten, sind die Zeichenketten von 1-12 und 6-10 Zeichen. In Abbildung 21 sind die Werte der Evaluationsmaße in einer Übersicht dargestellt. Mit einer durchschnittlichen Precision von 0,64, einem durchschnittlichen Recall von 0,65 und einem durchschnittlichen F-Maß von 0,64 über alle verwendeten Datensätze wurden hier die besten Werte erreicht.

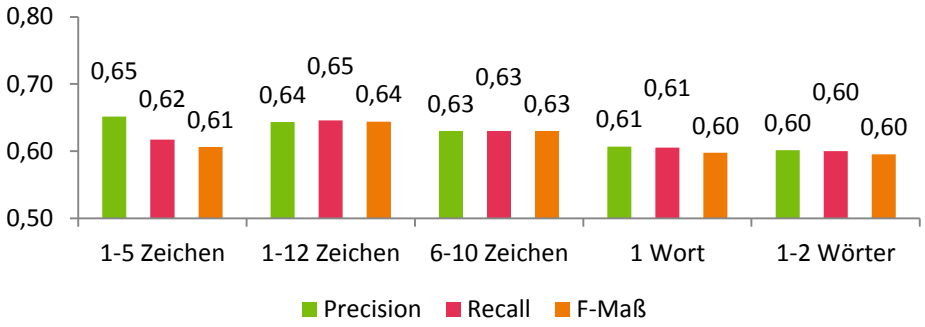

Abbildung 21: Übersicht der durchschnittlichen Evaluationsmaße der einzelnen Einstellungen

Die Trainingsmenge, welche mit diesen Methoden die höchsten Werte mit einer Precision von 0,74, einem Recall von 0,75 und einem F-Maß von 0,74 erzielte, war der automatisch markierte Datensatz der Domäne Handy (HA-2g).

4.5 Verwendung von Datensatzkombinationen

Für diese Phase der Evaluation wurden die Datensätze der Goldstandards mit denen der automatischen Markierung zusammengefügt, um so eventuell noch bessere Ergebnisse zu erzielen. Von den Goldstandards wurden dabei nur die Sätze der Kategorien positiv und negativ verwendet, da nur diese in beiden Datensätzen vorhanden waren. Ebenso blieb die Gleichverteilung der Sätze auf die Kategorien erhalten. Die Datensätze der Goldstandards wurden wieder in 80% und 20% unterteilt. Ersteres wurde der Trainingsmenge der automatischen Markierung hinzugefügt und auf letzterem wurden die erstellten Modelle getestet.

	Vorherige Trainingsmengen		Vorherige Testmengen	Neue Trainingsmengen	Neue Testmengen
1	**HA-2g**	80% von **HG-2g**	20% von **HG-2g**	**HA&HG-2g**	20% von **HG-2g**
2	**HA-2g**	80% von **HNG-2g**	20% von **HNG-2g**	**HA&HNG-2g**	20 % von **HNG-2g**
3	**NA-2g**	80% von **NG-2g**	20% von **NG-2g**	**NA&NG-2g**	20% von **NG-2g**
4	**HA-2g NA-2g**	80% von **HNG-2g**	20 von **HNG-2g**	**HA&NA&HNG-2g**	20 von **HNG-2g**

Tabelle 12: Übersicht der neuerstellten Kombinationen von Trainings- und Testmengen

4.5.1 Domänenweises Zusammenfügen der Datensätze

Die erste getestete Kombination, ist die der Domäne Handy (HA&HG-2g)[50]. In Abbildung 22 sind die erzielten Ergebnisse dargestellt. Die Werte der Evaluationsmaße fallen für alle sechs Einstellungen nie unter 0,7. Bisher sind dies die besten Ergebnisse eines Modells.

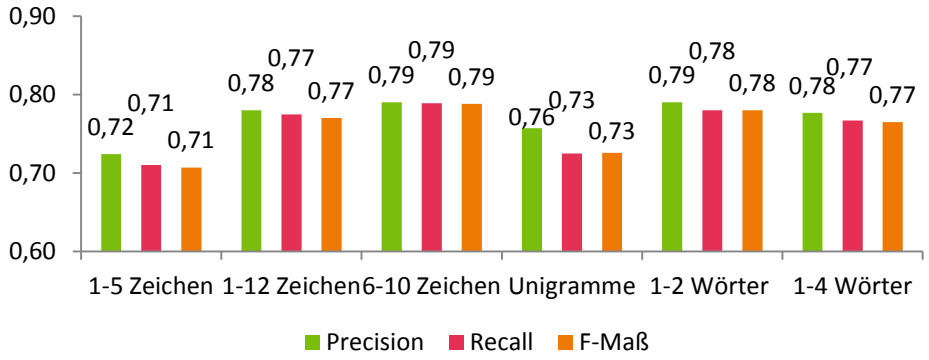

Abbildung 22: Übersicht der Ergebnisse der zusammengefügten Trainingsmengen der Domäne Handy

Negative Sätze werden mit dieser Kombination wesentlich genauer klassifiziert, mit einer Precision von 0,84 für die Einstellungen Unigramme und 1-2 Wörter als N-Gramme.

Insgesamt wurde mit dieser Kombination über alle Einstellungen ein durchschnittliches F-Maß von 0,76 erreicht. Für einen direkten Vergleich der Auswirkung des Kombinierens der beiden Datensätze, wurden beide einzeln ebenso auf der gleichen Testmenge[51] geprüft. Im Vergleich dazu lieferte der Goldstandard der Domäne Handy (HG-2g) mit der positiven und negativen Kategorie in einer Gleichverteilung ein durchschnittliches F-Maß von 0,70 und der automatisch markierte Datensatz der Domäne Handy (HA-2g) ein durchschnittliches F-Maß von 0,74.

Auch bei der der Datensatzkombination[52] aus der Domäne Notebook (NA&NG-2g) sind die Werte der Evaluationsmaße höher als mit der Trainingsmenge aus dem automatisch markierten Datensatz. Hier wiederum erstellen die Methoden 1-5 und 1-12 Zeichen das bessere Modell des Klassifikators. Auffällig ist auch, dass die negativen Sätze mit einer durchschnittlichen Precision von 0,73 besser klassifiziert werden, als die positiven Sätze mit 0,66.

[50] In Tabelle 12, Seite 86 ist dargestellt aus welchen Datensätzen diese Kombination entstand (Variante 1).
[51] Goldstandard Handy mit positiver und negativer Kategorie mit jeweils gleicher Anzahl an Sätzen
[52] Aus welchen Datensätzen diese Kombination besteht ist in Tabelle 12, Seite 86 abgebildet (Variante 3).

Für diese Kombination wurden ebenso zum Vergleich beide Datensätze auf der gleichen Testmenge überprüft. Hier erzielte die Trainingsmenge aus dem Notebook Goldstandard (NG-2g) ein durchschnittliches F-Maß von 0,72 und der automatisch markierte Datensatz der Domäne Notebook (NA-2g) ein F-Maß von 0,59. Somit ist für diese Domäne die Kombination dieser beiden Datensätze nicht sinnvoll, da damit nur ein durchschnittliches F-Maß von 0,69 erreicht wurde.[53]

4.5.2 Domänenübergreifendes Zusammenfügen der Datensätze

Eine weitere mögliche Kombination[54] ist, den automatisch markierten Datensatz der Domäne Handy (HA-2g) mit den Goldstandards beider Domänen (HNG-2g) zusammenzufügen. Da ersterer bisher das beste Modell erstellte, könnte dieses eventuell mit Hilfe aller manuell markierten Sätze noch weiter verbessert werden. Jedoch sind diese Ergebnisse, welche in Abbildung 23 dargestellt sind, mit einem durchschnittlichen F-Maß von 0,72 nicht so gut wie bei der gleichen Trainingsmenge (HA&HG-2g) nur ohne den Goldstandard der Domäne Notebook mit dem durchschnittlichen F-Maß von 0,76. Ebenso ist der kombinierte Goldstandard (HNG-2g) aus beiden Domänen eine bessere Trainingsmenge mit einem durchschnittlichen F-Maß von 0,72 besser als die Trainingsmenge mit den automatisch markierten Sätzen der Domäne Handy und den Goldstandards (HA&HNG-2g) zusammen.[55]

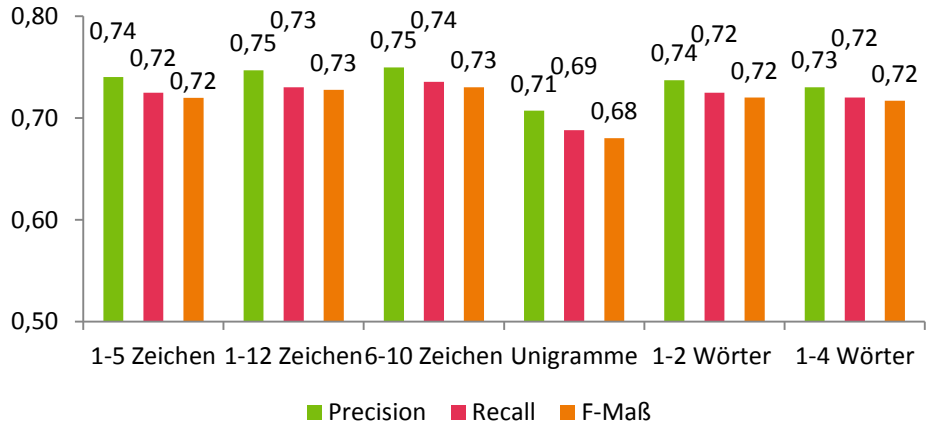

Abbildung 23: Übersicht der Ergebnisse der zusammengefügten Trainingsmengen der automatischen Markierung der Domäne Handy und beide Goldstandards

[53] Die ausführlichen Ergebnisse sind in Anhang D.1 und D.3 zu finden.
[54] Eine genaue Übersicht der Kombination ist in Tabelle 12, Seite 86 dargestellt (Variante 2).
[55] Die ausführlichen Ergebnisse sind in Anhang D.2 zu finden.

4.5.3 Alle Datensätze als Trainingsmenge

Eine weitere Variante für diesen Durchlauf der Evaluation war, alle Datensätze in einer Trainingsmenge zusammenzufügen und auf 20% der Goldstandards zu testen[56]. In Abbildung 24 sind die Ergebnisse dieser Kombination zusammengefasst. Auch hier ist zu bemerken, dass das durchschnittliche F-Maß dieser Kombination mit 0,72 schlechter war, als das der Trainingsmenge aus nur einer Domäne. Die Datensätze der Domäne Handy (HA&HG-2g) lieferten ein F-Maß von 0,76 und die der Domäne Notebook (NA&NG-2g) ebenso ein F-Maß von 0,69.[57]

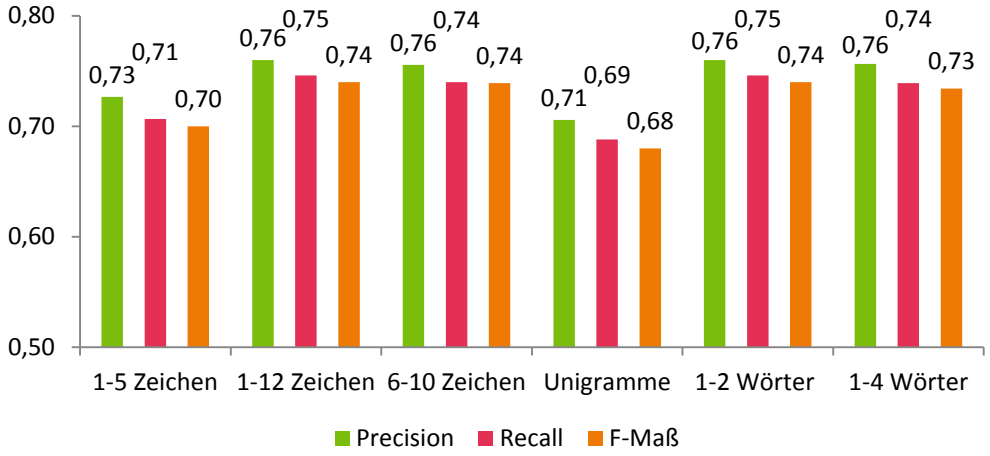

Abbildung 24: Ergebnisse der Trainingsmenge aus allen Datensätzen

4.5.4 Zusammenfassung

Insgesamt wird die Erkenntnis deutlich, dass die Kombination der Datensätze nicht immer bessere Ergebnisse erzielen als die Anwendung dieser einzeln. Einzig bei den Goldstandards liefert das Zusammenfügen eine bessere Trainingsmenge als die beiden Goldstandards separat. Für die automatisch markierten Datensätze gilt dies nicht. Hier liefert die Kombination der Sätze einer Domäne bessere Ergebnisse, als das Zusammenfügen mehrerer Domänen.

4.5.5 Ergebnisse der Datensatzkombinationen für die Einstellung 1-12 Zeichen

In diesem Abschnitt sind die Ergebnisse der kombinierten Datensätze für die beste Einstellung 6-10 Zeichen in Tabelle 13 abgebildet. Im Vergleich zu den Werten

[56] Eine genaue Übersicht der Kombination ist in Tabelle 12, Seite 86 dargestellt (Variante 4).
[57] Die ausführlichen Ergebnisse sind in Anhang D.4 zu finden.

aus den beiden vorherigen Abschnitten erzielen die Kombinationen bessere Werte als alle bisherigen Trainingsmengen. Die beste Kombination ist die der Domäne Handy (HA&HG-2g).

Trainingsmenge	Testmenge	Precision			Recall			F-Maß
		P	N	ø	P	N	ø	
HA&HG-2g	20% von HG-2g	0,76	0,83	0,79	0,85	0,73	0,79	0,79
HA&HNG-2g	20% von HNG-2g	0,69	0,81	0,75	0,86	0,62	0,74	0,73
NA&NG-2g	20% von NG-2g	0,65	0,73	0,69	0,79	0,57	0,68	0,68
HA&NA&HNG-2g	20% von HNG-2g	0,70	0,81	0,76	0,86	0,63	0,74	0,74

Tabelle 13: Ergebnisse der Datensatzkombinationen für die Einstellung 6-10 Zeichen

4.5.6 Die besten Evaluationsmaße der verschiedenen Einstellungen

In diesem Abschnitt werden die besten Evaluationsmaße der Datensätze übersichtlich dargestellt. Interessant hierbei sind die einzelnen Einstellungen, welche diese Werte erzielten.

Die besten Precision-Werte wurden für die folgenden in Tabelle 14 dargestellten Einstellungen mit den Datensätzen erreicht. Die Sätze wurden am häufigsten mit der Methode 1-5 Zeichen für Zeichenketten eingeordnet, was am genauesten war.

Methode	Precision	Datensatz
1-5 Zeichen	0,63	HG-3g
	0,61	NG-3g
	0,67	HNG-3g
	0,65	NA-2g
	0,74	NA&NG-2g
6-10 Zeichen	0,74	HA-2g
	0,65	NA-2g
	0,79	HA&HG-2g
	0,75	HA-HNG-2g
1-12 Zeichen	0,76	HA&NA&HNG-2g
	0,75	HA-HNG-2g
1-2 Wörter	0,79	HA&HG-2g
1-4 Wörter	0,76	HA&NA&HNG-2g

Tabelle 14: Die besten Precision-Werte nach Einstellungen

In Tabelle 15 sind die besten Recall-Werte und die Einstellungen mit denen diese erreicht wurden dargestellt. Hier wiederum lieferte die Einstellung 1-12 Zeichen am häufigsten die besten Ergebnisse, vorrangig für die Trainingsmengen aus den Goldstandards.

Methode	Recall	Datensätze
1-5 Zeichen	0,66	NA-2g
	0,74	NA&NG-2g
1-12 Zeichen	0,62	HG-3g
	0,75	HA&NA&HNG-2g
	0,61	NG-3g
	0,63	HNG-3g
6-10 Zeichen	0,75	HA-2g
	0,66	NA-2g
	0,79	HA&HG-2g
	0,74	HA&HNG-2g
1-2 Wörter	0,75	HA&NA&HNG-2g

Tabelle 15: Die besten Recall-Werte nach Einstellungen

In Tabelle 16 Sind die besten F-Maße abgebildet. Die Einstellung 1-12 Zeichen erzielte auch hier für die meisten Trainingsmengen die besten Werte.

Methode	F-Maß	Datensätze
1-5 Zeichen	0,65	NA-2g
	0,74	NA&NG-2g
1-12 Zeichen	0,62	HG-3g
	0,60	NG-3g
	0,63	HNG-3g
	0,73	HA&HNG-2g
	0,74	HA&NA&HNG-2g
6-10 Zeichen	0,74	HA-2g
	0,74	HA&NA&HNG-2g
	0,65	NA-2g
	0,79	HA&HG-2g
	0,73	HA&HNG-2g

Tabelle 16: Die besten F-Maß nach Methoden

Insgesamt ist die Einstellung, welche im Durchschnitt die besten Werte für die Evaluationsmaße hervorbringt, die Einstellung 1-12 Zeichen, wie in Tabelle 17 abgebildet ist.

	Precision	Recall	F-Maß	ø
1-5 Zeichen	0,69	0,66	0,66	0,67
1-12 Zeichen	0,69	0,69	0,68	0,69
6-10 Zeichen	0,68	0,68	0,67	0,68
Unigramme	0,66	0,65	0,64	0,65
1-2 Wörter	0,66	0,66	0,65	0,66
1-4 Wörter	0,65	0,65	0,64	0,65
ø	0,67	0,67	0,66	0,66

Tabelle 17: Die durchschnittliche Evaluationsmaße der Methoden über alle Datensätze

4.6 Durchführung einer Threshold-Analyse

In der sechsten Phase der Evaluation wurde betrachtet, ob es hilfreich ist, einen Threshold für den Klassifikator einzufügen. Dies ist eine Grenze für die Wahrscheinlichkeit, mit der der Klassifikator Sätze einer bestimmten Kategorie zuordnet. Das heißt, damit kann beeinflusst werden, dass nur Sätze klassifiziert werden, wenn der Klassifikator sich mindestens zu der gewählten Wahrscheinlichkeit sicher ist, dass der Satz in diese Kategorie gehört. Verwendet wurden hierfür die beste Trainingsmenge, mit dem durchschnittlich besten F-Maß (HA&HG-2g) von 0,76 der bisherigen Untersuchungen sowie die dazugehörige Testmenge. Als Einstellung wurde die durchschnittlich beste Einstellung von 1-12 Zeichen benutzt.

Die Grenze wurde zunächst auf 10% gesetzt. Die Evaluationsmaße zeigten bis zu der Grenze von 50% keine Veränderung. Danach stieg die Precision kontinuierlich an, wohingegen der Recall kontinuierlich gefolgt vom F-Maß sank. Das Verhalten der Ergebnisse ist in Abbildung 25 abgebildet.

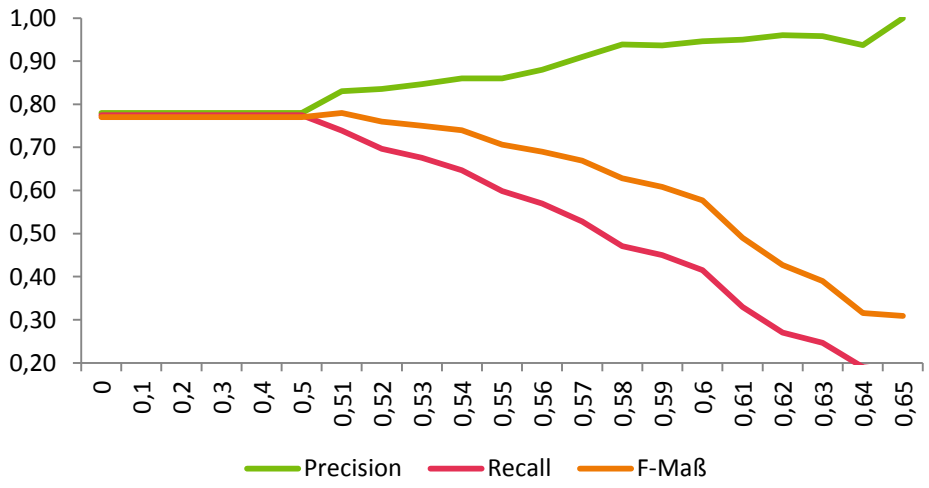

Abbildung 25: Veränderung der Evaluationsmaße mit steigenden Threshold

Auch die Anzahl der überhaupt eingeordneten Sätze sank mit einem höheren Threshold, ebenso die Gesamtanzahl an richtig eingeordneten Sätzen. Dafür wurde der Anteil der richtigen an den überhaupt eingeordneten Sätzen immer größer, bis er bei einem Threshold von 0,65 gleich wurde. Die Veränderung dieser Werte ist in Abbildung 26 abgebildet.

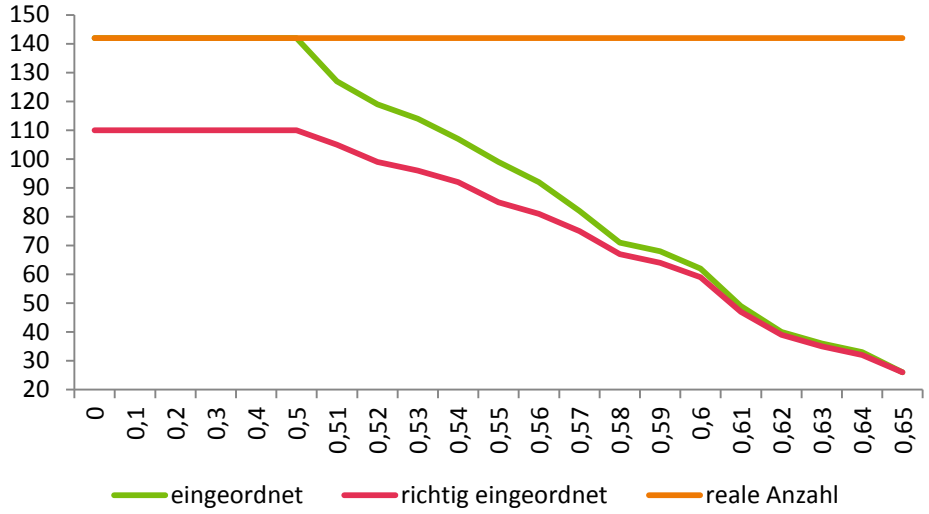

Abbildung 26: Veränderung der Anzahl an richtig und überhaupt eingeordneten Sätzen

Insgesamt zeigte sich bei der Threshold-Analyse, dass zwar mit steigender Grenze immer mehr eingeordnete Beispiele auch wirklich korrekt eingeordnet wurden. Jedoch sank gleichzeitig die Anzahl der überhaupt eingeordneten Beispiele. Eine immer größer werdende Zahl von Beispielen wurde ausgelassen. Bei einer großen Masse an Eingabedaten führt dies eventuell zu keiner Beeinträchtigung und der Hauptfokus kann somit auf die Genauigkeit der Ergebnisse gelegt werden. Bei einem kleinen Input jedoch ist diese Grenze eventuell nicht hilfreich, da noch weniger Sätze wirklich kategorisiert werden.[58]

4.7 Prüfen der Domänenabhängigkeit der einzelnen Modelle

Ein großes Problem des Opinion Mining ist, dass die meisten Verfahren nicht domänenübergreifend anwendbar sind.[59] Dies ist unabhängig von der Art des Verfahrens, sei es regelbasiert oder statistisch basiert. In diesem Abschnitt wird getes-

[58] Die ausführliche Übersicht der Ergebnisse ist in Anhang E abgebildet.
[59] Näheres zu der Domänenabhängigkeit ist im Grundlagenkapitel in Abschnitt 2.2.2 ab Seite 15 zu finden.

tet, ob sich trainierte Modelle des Klassifikators auf andere Domänen übertragen lassen.

4.7.1 Übertragung auf einen weiteren technischen Bereich

Als erstes wurde getestet, ob sich die Modelle der Trainingsmenge einer technischen Domäne auf eine weitere technische Domänen übertragen lassen. Hier liegt die Vermutung nahe, dass sich die Art der Sätze nicht vollkommen unterscheidet. Dass diese Datensätze auch kombinierbar zum Trainieren sind, wurde u. a. in Abschnitt 4.5 gezeigt.

Bei der Anwendung des Modelles der Domäne Handy (HA&HG-2g) auf den Datensatz der Domäne Notebook unterschreiten die Evaluationsmaße bei keiner Einstellung die Grenze von 0,60, jedoch sind die Ergebnisse mit der Trainingsmenge (NA&NG-2g) und der Testmenge aus der Domäne Notebook ein wenig besser. In Abbildung 27 sind die durchschnittlichen Evaluationsmaße im Vergleich zu sehen.

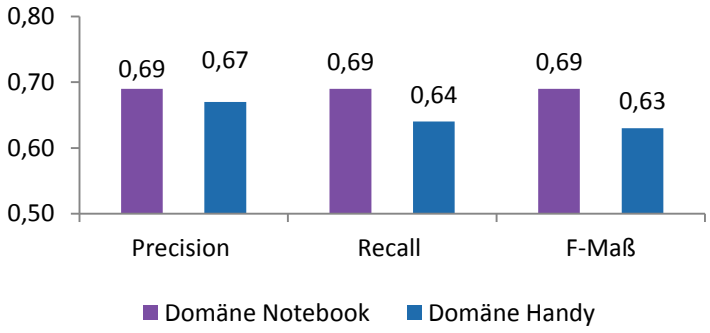

Abbildung 27: Vergleich der durchschnittlichen Evaluationsmaße des domänenübergreifenden Testens

In Abbildung 28 ist der Vergleich der durchschnittlichen Werte der Evaluationsmaße der beiden Modelle auf die Testmenge des Handy Goldstandards abgebildet. Wie bei der Domäne Notebook arbeitete die Trainingsmenge der Domäne Handy (HA&HG-2g) besser auf der Testmenge der gleichen Domäne als die Trainingsmenge der Domäne Notebook (NA&NG-2g). Jedoch sind auch hier die erzielten Werte mit der domänenfremden Trainingsmenge nie unter 0,60.

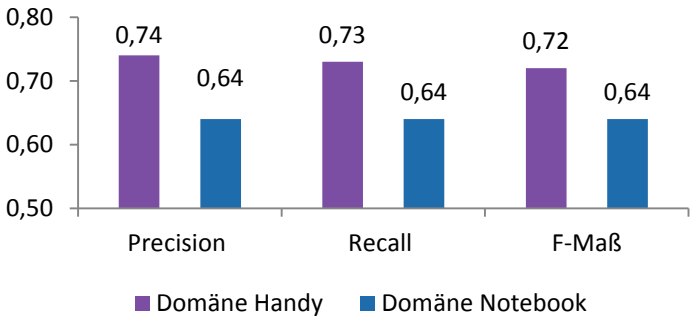

Abbildung 28: Vergleich der durchschnittlichen Evaluationsmaße des domänenübergreifenden Testens

4.7.2 Übertragung auf eine nicht technische Domäne

In der zweiten Durchführung der Modellübertragung wurde eine nicht technische Domäne verwendet. Hier handelt es sich um Sätze von Blogs, News oder auch Wikipedia-Einträgen. Der erste Test erfolgte mit den Datensätzen der Domäne Handy (HA&HG-2g) und mit einer gleichen Anzahl an Kategorien – nur positiv und negativ. Es zeigte sich, wie in Abbildung 29 dargestellt, dass die beiden Einstellungen 1-5, sowie 1-12 Zeichen bessere Werte erzielen als die anderen Einstellungen. Anzumerken ist, dass die Anzahl der Kategorien in beiden Mengen, also der Trainings- und der Testmenge, übereinstimmen.

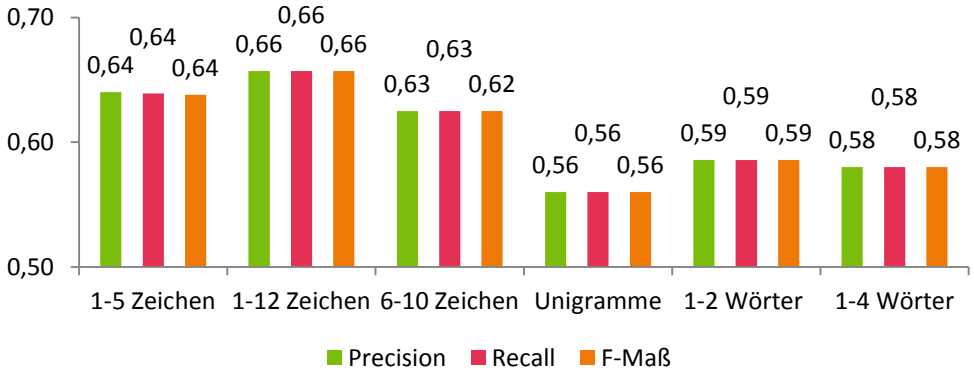

Abbildung 29: Ergebnisse der Anwendung des besten Modells der technischen Domänen auf eine technisch fremde Domäne mit gleicher Kategorienanzahl

In Abbildung 30 sind die Ergebnisse mit unterschiedlichen Kategorien zu sehen. Es zeigte sich wieder, dass eine ungleiche Anzahl an Kategorien in Trainings- und Testmengen schlechtere Ergebnisse lieferte, was bereits in Abschnitt 4.2.3.3, Seite 57, gezeigt wurde. Die eingezeichnete Linie ist die Grenze, welche die Werte der Precision mindestens annehmen müssten, da in der Trainingsmenge eine Gleichverteilung der Sätze auf die Kategorien vorliegt. Da der dargestellte durchschnitt-

liche Precision-Wert über alle Kategorien gebildet wurde und nach der Anzahl der Beispiele innerhalb der Testmenge gewichtet wurde, fällt dieser unter die Grenze.

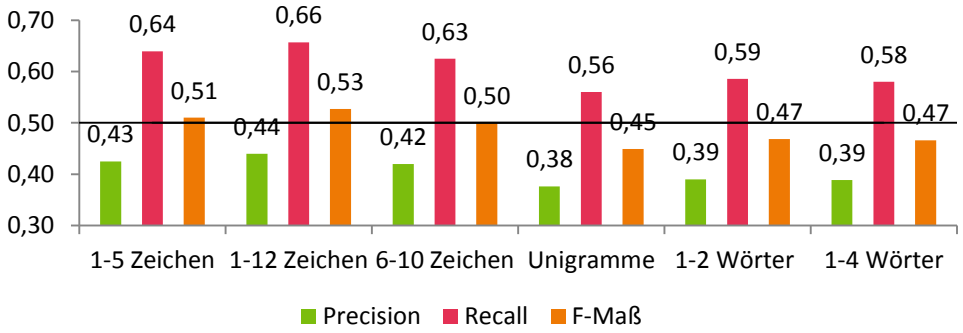

Abbildung 30: Ergebnisse der Anwendung des besten Modells der technischen Domänen auf eine technisch fremde Domäne mit ungleicher Kategorienanzahl

4.7.3 Ergebnisse

Es wurde gezeigt, dass es möglich ist, erstellte Modelle des Klassifikators mehr oder minder erfolgreich auf eine weitere Domäne anzuwenden. Bei einer Gleichverteilung der Beispielsätze der Kategorien, müssen die Ergebnisse der Evaluationsmaße größer als 0,5 sein, damit es überhaupt sinnvoll ist, einen Klassifikator anzuwenden. Die Ergebnisse sind bei fast allen Versuchen über diesem Wert. Bei dem Fall, wo dies nicht zutrifft, liegt es vermutlich an der dritten Kategorie in der Testmenge, auf welche der Klassifikator nicht trainiert wurde.[60]

4.8 Vergleich mit einem regelbasiertem Ansatz

Nachdem mit bisherigen Ergebnissen gezeigt wurde, dass deutsche, nicht vorverarbeitete Sätze mit Hilfe eines Klassifikators[61] aus dem Bereich des maschinellen Lernens in die Kategorien positiv, negativ und objektiv erfolgreich eingeordnet werden können, wird in diesem Abschnitt diese Methode mit einem wörterbuchbasierten (regelbasiertem) Ansatz verglichen. Wie im Grundlagenkapitel beschrieben gibt es für das Opinion Mining zwei unterschiedliche Ansätze.[62] Allgemein kann man diese als regelbasiert und als statistisch beschreiben. Ersterer arbeitet beispielsweise mit grammatikalischen Regeln oder Wörterbüchern und letzterer beispielsweise mit maschinellen Klassifikatoren.

[60] Die ausführlichen Ergebnisse sind in Anhang F dargestellt.
[61] In dieser Arbeit war es der „Dictionary Classifier“ des Palladian Toolkit.
[62] In Abschnitt 2.2.2, Seite 14 werden die unterschiedlichen Ansätze erläutert.

4.8.1 Durchführung des Vergleiches ohne Vorverarbeitung

Das Wörterbuch, welches hier zum Einsatz kam, ist das SentiWS der Universität Leipzig[63]. Es wurde mit Hilfe des Palladian Toolkits der TU Dresden als ein Klassifikator („Sentiment Classifier“) umgesetzt. Die zu klassifizierenden Sätze werden in Tokens, also Wörter, unterteilt und anschließend mit den vorhandenen Wörtern im Wörterbuch verglichen und deren positive bzw. negative Werte werden aufsummiert. Die Summe, welche entweder positiv oder negativ ist, stellt die Polarität des Satzes dar. Zusätzlich wird mit einbezogen, ob vor dem gefundenen Wort eine Negation oder ein bestärkendes Wort zu finden ist. Dies kann die Polarität ebenso beeinflussen.

Die Sätze für das Klassifizieren mit Hilfe des maschinellen Lernens wurden nicht vorverarbeitet, es gab keine Rechtschreibkorrektur oder Synonym-Findung. Auf die gleiche Art und Weise wurde der Wörterbuch-Klassifikator auf die Sätze angewendet werden. Als Datensatz zum Klassifizieren wurden die Testmengen des Goldstandards verwendet, was 20% des gleichverteilten Goldstandards jeder Domäne, und analog nur mit den Kategorien positiv und negativ ausmachte.

Der „Sentiment Classifier“ wurde auf alle drei Domänen angewendet und mit Ergebnissen des „Dictionary Classifier“ verglichen. Insgesamt sind die Werte der Evaluationsmaße mit dem Goldstandard der Domäne Handy (HG-3g) bzw. mit der Kombination des Goldstandards und der automatisch markierten Datensätzen der Domäne Handy (HA&HG-3g) höher als die Werte, welche mit dem Wörterbuch erzielt wurden, wie in Abbildung 31 dargestellt ist.

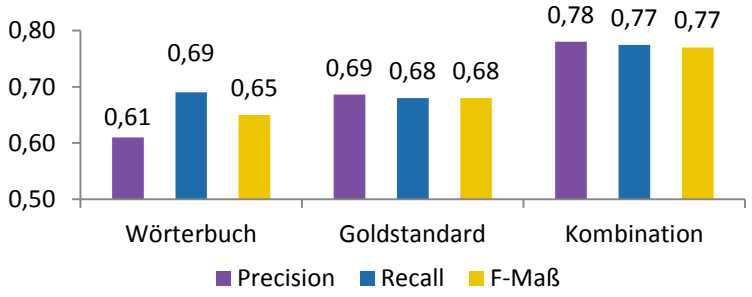

Abbildung 31: Vergleich der Ergebnisse des maschinellen Klassifizierens mit denen des Wörterbuches der Domäne Handy

Auch in der Domäne Notebook sind die Ergebnisse des maschinellen Lernens besser als die des Wörterbuches (siehe Abbildung 32). Für die Domäne Notebook

[63] Eine detaillierte Beschreibung ist in Abschnitt 2.3.1 zu finden.

(NG-3g) und beide zusammengefügten Domänen (NA-NG-3g) sind ähnliche Ergebnisse entstanden.

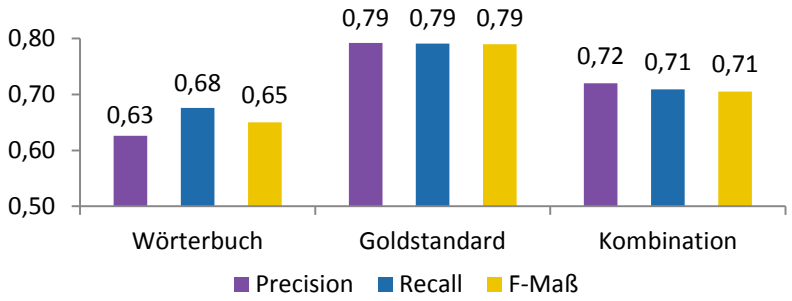

Abbildung 32: Vergleich der Ergebnisse des maschinellen Klassifizierens mit denen des Wörterbuches der Domäne Notebook

Schlussfolgerungen, welche daraus gezogen werden können sind, dass eine Klassifizierung mit Hilfe eines Wörterbuches ohne jegliche Vorverarbeitung nicht so gute Ergebnisse erzielt. Möglicherweise werden viele Wörter nicht richtig erkannt, da sie beispielsweise fehlerhaft geschrieben sein könnten. Besonders Datensätze, welche aus dem Internet extrahiert und nicht von Nachrichtenseiten oder auch Wikipedia-Einträgen erhalten wurden, sind meist unreine Daten. Sie sind oft mit Rechtschreibfehlern bestückt. Außerdem werden häufig umgangssprachliche Wörter benutzt, welche in dieser Form nicht in dem Wörterbuch auftauchen. Eine Vorverarbeitung ist somit dringend notwendig.

4.8.2 Vergleich mit Vorverarbeitung

Ein erster Schritt in Richtung Vorverarbeitung der zu analysierenden Sätze wurde mit Hilfe der SemaSuite bzw. der Funktion der Rechtschreibkorrektur dieser umgesetzt. Die gleichen Testmengen wie im vorherigen Abschnitt wurden also hinsichtlich ihrer Rechtschreibfehler korrigiert. Wenn durch die SemaSuite erkannt wurde, dass ein Wort falsch geschrieben ist, so wurde ein Korrekturvorschlag mit Hilfe des vorhandenen Wissensmodelles[64] bzw. dem eingebundenem Standardwörterbuch ermittelt. Der erste Eintrag aus der Liste der Vorschläge wurde verwendet und mit diesem das falsch geschriebene Wort in dem Satz ersetzt.

Anschließend wurde der „Sentiment Classifier“ des Palladian Toolkits auf diese korrigierten Sätze angewendet. In Abbildung 33 ist der Vergleich der Ergebnisse

[64] In diesem Fall waren es Produktbeschreibungen von Handys. Nähere dazu in dem folgenden Kapitel.

des „Sentiment Classifiers“ auf die originalen Sätze und die rechtschreibkorrigierten Sätze dargestellt.

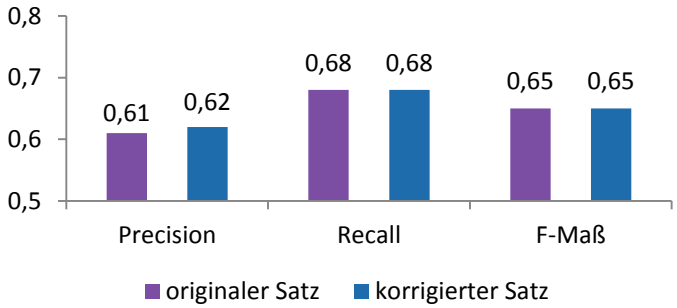

Abbildung 33: Vergleich der Testmengen des zusammengefügten Goldstandards für den „Sentiment Classifier“

Die Ergebnisse der beiden anderen Trainingsmengen, der Goldstandards der beiden Domänen Handy und Notebook, sind ähnlich ausgefallen. Die Rechtschreibprüfung hat nur kleine Verbesserungen der einzelnen Evaluationsmaße erzielt.

4.8.3 Ergebnisse

Abschließend kann gesagt werden, dass das Klassifizieren von Sätzen mit Hilfe eines Wörterbuches insgesamt schlechtere Ergebnisse erbracht hat, als der statistische Ansatz mit dem Klassifikator. Ursachen hierfür könnte weiterhin die nicht vollständige Vorverarbeitung sein. Möglicherweise wäre es hilfreicher, zusätzlich zu der Rechtschreibkorrektur, Synonyme der vorhandenen Wörter in den Sätzen zu ermitteln und diese bei der Klassifikation ebenso zu betrachten.[65]

4.9 Zusammenfassung

In diesem Kapitel wurden die einzelnen Phasen der Evaluation und deren Ergebnisse beschrieben. Die verwendeten Methoden und Datensätze wurden dargestellt. In Abschnitt 4.2 wurde unter anderem das Ergebnis erzielt, dass reale Verteilungen einer Stichprobe aus Sätzen für die jeweiligen Polaritätskategorien, für Trainingsmengen des maschinellen Lernens, ausgeschlossen werden können. Außerdem wurde gezeigt, dass die Anzahl der Kategorien in den Trainings- und Testmengen identisch sein sollten.

[65] Eine detaillierte Übersicht der einzelnen Ergebniswerte befindet sich in Anhang G.

Der Abschnitt 4.3 behandelte die Einstellungsmöglichkeiten des Klassifikators. Hier stellte sich heraus, dass für die Variante Wörter als N-Gramme ein Wort bzw. 1-2 Wörter die besten Werte erzielten. Dies galt ebenso für die Variante Zeichenketten als N-Gramme mit den Längen 1-5 und 1-12 Zeichen. Für die Durchführung der Tests wurde der zusammengefügte Goldstandard aus beiden Domänen mit einer Gleichverteilung verwendet, welcher die besten Evaluationsmaße in der ersten Durchführung erreichte.

Im dritten Teil der Evaluation, dem Abschnitt 4.4 wurden alle übrig gebliebenen Datensätze der ersten Phase mit den besten Einstellungen der zweiten Phase getestet. Aus diesen Ergebnissen gingen die Datensatzkombinationen aus Abschnitt 4.5 hervor. Hier wurde gezeigt, dass domänengleiche Datensätze erfolgreich zusammengefügt werden können. Besonders die Kombination der Domäne Handy erzielte gute Werte für die Evaluationsmaße.

Anschließend wurde eine Threshold-Analyse in Abschnitt 4.6 für die Trainingsmenge durchgeführt, welche das beste Modell in den vorherigen Abschnitten erstellte. In Abschnitt 4.7 wurde geprüft, ob die erstellten Modelle für den Klassifikator domänenabhängig sind. Es stellte sich heraus, dass es möglich ist, die Modelle zu übertragen, jedoch sind die Ergebnisse besser, wenn die Domänen sich nicht völlig im Inhalt unterscheiden.

Der letzte Schritt der Evaluation bestand aus einem Vergleich der Werte mit denen eines Klassifikators, der auf einem Wörterbuch basierte. Hier stellte sich heraus, dass das maschinelle Lernen bei Sätzen ohne Vorverarbeitung bessere Ergebnisse erzielte, als das Wörterbuch. Selbst mit einer Rechtschreibkorrektur der einzelnen Sätze konnten keine befriedigenden Ergebnisse erzielt werden.

5 Extraktion der Aspekte der beurteilten Produkte

5.1 Theoretisches Vorgehen der Aspekt-Extraktion

Nachdem die Sätze den Kategorien positiv und negativ zugeordnet wurden, musste anschließend herausgefunden werden, von welchen Themen[66] bzw. Aspekten[67] in diesen Sätzen die Rede war. Mit Hilfe der SemaSuite, welche in Abschnitt 2.4.2 ausführlich beschrieben wurde, wird nun versucht die Informationsextraktion durchzuführen.

Dabei wird davon ausgegangen, dass Aspekte vorrangig Nomen sind. Dementsprechend müssen diese aus den Sätzen extrahiert werden. Da die SemaSuite mit Hilfe von semantischen Netzen arbeitet, werden Daten benötigt, aus denen diese gebildet werden können. Hierfür wird erneut die Plattform Amazon verwendet. Gerätebeschreibungen von Handys und Notebooks werden extrahiert und für das Bilden von semantischen Netzen verwendet. Die extrahierten Nomen aus den Sätzen werden dann diesen Netzen zugeordnet.

Für das Bilden der Problemklassen reicht es jedoch nicht aus, nur zu wissen auf was genau die Äußerungen zielen. Hier stellte sich die Frage, ob eine weitere Berechnung notwendig wäre, welche zählt, wie häufig die einzelnen Aspekte in den positiven und negativen Sätzen vorkommen, um dann daraus Problemklassen bilden zu können.

[66] Themen sind in diesem Fall Produkte aus den Bereichen Notebook und Handy
[67] Aspekte sind hier Eigenschaften von Produkten aus den Bereichen Notebook und Handy

5.2 Erstellung des Prototyps

Der erste Schritt für die Erstellung des Prototyps war das Bilden eines für den Zweck der Aspekt-Erkennung passenden Wissensmodells. Hier wurden 20 Produktbeschreibungen für jeden Bereich (Notebook und Handy) von dem Online-Handel Amazon extrahiert. Ziel war es, mit diesen Beschreibungen ein ausreichendes Modell aus Nomen bzw. Produkteigenschaften zu erstellen. Dafür wurden nur Nomen aus den Dokumenten extrahiert. In Abbildung 34 ist ein Auszug des erstellten Wissensmodelles der Domäne Handy zu sehen.

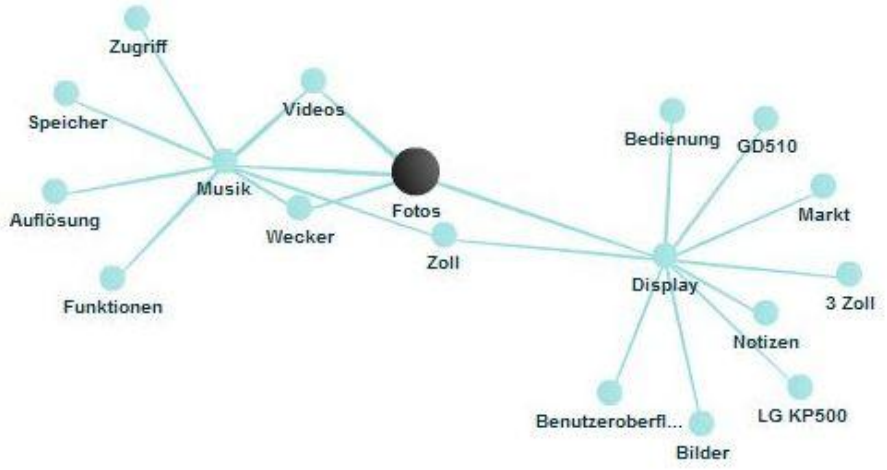

Abbildung 34: Auszug aus dem Wissensmodell der Domäne Handy

Hier sind verwandte und untergeordnete Konzepte des Aspektes Fotos dargestellt. Wie zu sehen ist, steht dieser Aspekt mit einer Vielzahl anderer Aspekte in Verbindung. Für die anschließende Analyse der klassifizierten Sätze sind diese Beziehungen jedoch nicht ausschlaggebend. Dafür ist es wichtiger, dass alle wichtigen Eigenschaften der Produkte überhaupt als Konzepte erkannt wurden und somit wiederum auch innerhalb der eingeordneten Sätze wiedererkannt werden können.

Nachdem passende Wissensmodelle mit Hilfe der SemaSuite erstellt wurden, war anschließend die Entscheidung zu treffen, ob die weiteren Berechnungen extern durchgeführt werden sollten, oder das Palladian Toolkit in die SemaSuite eingebunden werden sollte. Für letzteres wurde sich entschieden, weil das Klassifizieren der Sätze, die Wissensmodellbildung und auch das Erstellen der Problemklassen somit in nur einer Software umgesetzt sind.

Für das Einbinden des „Dictionary Classifiers“ des Palladian Toolkit wurde der SemaSuite ein weiterer Annotator hinzugefügt, welcher ein Trainingsmodell laden und anwenden kann. Werden Dokumente nun der Software für die Analyse übergeben, erfolgt neben der Tokenisierung etc. ein weiterer Verarbeitungsschritt und zwar die satzweise Klassifikation dieser Dokumente in positiv, negativ und objektiv. In der Applikation der Suche wurde die Batch-Suche den Bedürfnissen des Prototyps angepasst.

Es kann somit nun eine Datei hochgeladen werden, welche als Excel-Datei zurückgegeben wird. In dieser sind alle Sätze, mit der errechneten Polarität, sowie den erkannten Aspekt zu sehen. Zusätzlich wurde berechnet, wie oft ein gefundener Aspekt in Sätzen aus den Kategorien vorkam. Hier ist die Tendenz zum Bilden der Problemklassen zu finden. Jedoch wurde diese Berechnung generisch umgesetzt, das heißt es wird aus den Dokumenten erkannt, welche Kategorien vorhanden sind. Da nicht in jedem Fall negative Sätze dabei sein müssen, wurde sich dafür entschieden für jeden Aspekt die Anzahl der Vorkommen in den Kategorien aufzulisten. In Abbildung 35 ist ein Teil der Ausgabe abgebildet.

263		N	1
264	Bild	O	1
265		P	0
266		N	1
267	Foto	O	0
268		P	10
269		N	14
270	Handy	O	22
271		P	0
272		N	0
273	Musik	O	0
274		P	0
275		N	0
276	Speicher	O	0
277		P	0

Abbildung 35: Ausschnitt der Ausgabe des Prototyps

5.3 Zusammenfassung

In diesem Kapitel wurde beschrieben, wie die Aufgabenstellung prototypisch durch die Integration des „Dictionary Klassifikators“ in die SemaSuite der T-Systems Multimedia Solutions GmbH. Die Ausgabe des Prototyps umfasst die einzelnen Sätze mit ihrer erkannten Polarität sowie die gefundenen Aspekte. Zusätzlich wird das Vorkommen letzterer in den Sätzen der drei Kategorien gezählt und aufgelistet.

6 Zusammenfassung

Für das noch recht neue Fachgebiet Opinion Mining gibt es aktuell zahlreiche Lösungsansätze, welche jedoch noch keine vollkommen korrekte Ergebnisse liefern. Die Gemeinsamkeit dieser Lösungsansätze ist, dass die Stimmungsanalyse in mehrere Teilschritte zerlegt wird. Sie unterscheiden sich jedoch in den dafür verwendeten Verfahren. In Anlehnung an die Textklassifikation nach Themen, bestanden die ersten Ansätze aus der Suche nach Schlüsselwörtern für die einzelnen Polaritäten. Dabei wurde deutlich, dass das Opinion Mining wesentlich komplexer ist und einzelne Wörter oder Phrasen nicht alleinige Indikatoren für Meinungen sind. Eine Einbeziehung des Kontextes ist essentiell für gute Ergebnisse. Hierdurch werden jedoch die umgesetzten Lösungen von dem thematischen Bereich, welcher bei der Umsetzung betrachtet wird, abhängig. Damit sind sie nicht immer übertragbar auf andere Bereiche. Ebenso können Erfolg bringende Verfahren, welche in einer Sprache umgesetzt wurden, nicht zwangsläufig ebenso erfolgreich auf eine weitere Sprache übertragen werden. Die deutsche Sprache ist hinsichtlich des Opinion Mining weitestgehend unerforscht, und wurde in dieser Arbeit betrachtet.

Die unterschiedlichen Verfahren, welche für das Opinion Mining verwendet werden, stammen aus zahlreichen Fachgebieten, wie zum Beispiel der Computerlinguistik und der Informationsextraktion. Da diese ebenso nicht immer vollständig korrekte Ergebnisse liefern, gibt es mehrere Fehlerquellen, von denen die Qualität des Gesamtergebnisses für das Opinion Mining abhängt. Die größte Herausforderung in diesem Zusammenhang ist es, die zahlreichen Verfahren aus den unter-

schiedlichen Fachbereichen an die Stimmungsanalyse anzupassen und weiter zu verbessern.

In der vorliegenden Arbeit wurde die Umsetzung des Opinion Mining ebenfalls in mehrere Teilschritte zerlegt. Im ersten Schritt erfolgte das Ordnen der Sätze hinsichtlich ihrer Polarität, im zweiten Schritt die Untersuchung dieser auf ihren Inhalt. Allein bei diesen zwei Schritten können verschiedene Fehler auftreten. Dokumente, welche klassifiziert werden sollen, liegen in der Realität nicht in einzelnen Sätzen vor. Die Texte müssen vorher in Sätze aufgeteilt werden. An dieser Stelle können die ersten Fehler u. a. durch Abkürzungen oder Rechtschreibfehler entstehen. Der Klassifikator, für den ein Modell mit einer Trainingsmenge erstellt wurde, stellt eine weitere Fehlerquelle dar. Für die untersuchten Datensätze und Einstellungen konnte in dieser Arbeit insgesamt eine durchschnittliche Precision von 0,67 erreicht werden. Im Bereich des Opinion Mining ist dieser Wert ein oft mit anderen Methoden erreichter, aber selten übertroffener Wert. Die Klassifikation wurde in dieser Arbeit mit Hilfe eines maschinellen Lernverfahrens umgesetzt. Verwendet wurde ein Klassifikator, welcher aus einer Trainingsmenge die Wahrscheinlichkeiten für gewählte N-Gramme berechnet, mit welcher sie in die mit der Trainingsmenge festgelegten Kategorien gehören. Da die Qualität des Klassifikators von der gebildeten Trainingsmenge abhängt, wurde untersucht, mit welchen Datensätzen bessere Ergebnisse erzielt werden können. Im Laufe der Arbeit bestätigte sich die Eigenschaft von auf statistischen Werten basierenden Verfahren, dass weitestgehend korrekte, aber wenige Beispiele beim Trainieren schlechtere Ergebnisse liefern, als weniger korrekte, dafür jedoch viele.

Die verwendeten Datensätze wurden jeweils aus Produktbewertungen erstellt. Die Sätze der wesentlich kleineren Datensätze (ca. 1.200) erhielten manuell Markierungen bezüglich ihrer Zugehörigkeit zu den Kategorien positiv, negativ und objektiv. Die Sätze der beiden größeren Datensätze (ca. 10.000) wurden dagegen mit Hilfe der von den Rezensionsschreibern gegebenen Sternbewertungen automatisch markiert. Da nur die ein und fünf Sternbewertungen verwendet wurden, liegen bei diesen Datensätzen nur die Kategorien positiv und negativ vor. In der Literatur ist die Verwendung der objektiven Kategorie umstritten. In dieser Arbeit wurde ebenfalls untersucht inwiefern sich die Ergebnisse, mit oder ohne der Kategorie in den Trainingsmengen, unterscheiden, wenn sie dennoch auf eine realitätsnahe Testmenge mit ihr angewendet werden. Es zeigte sich, dass die beiden anderen Kategorien (positiv und negativ) genauer klassifiziert werden, wenn auch nur diese trainiert werden.

Die reale Verteilung der Sätze auf die einzelnen Kategorien wurde ausgeschlossen, da ein Klassifikator gleichmäßig auf alle Kategorien trainiert werden sollte um soviel wie mögliche Dokumente der Realität einordnen zu können. Ein Priorisieren einer einzelnen Kategorie durch eine größere Anzahl an Trainingsbeispielen wird nicht empfohlen, da nicht davon ausgegangen werden kann, dass eine bestimmte Kategorie in der Realität öfter vorkommt als eine andere. In der Stichprobe dieser Arbeit trägt es den Anschein als gäbe es mehr positive Sätze, als negative in Rezensionen. Doch erstere liegen nicht im alleinigen Fokus der Betrachtungen des Opinion Mining.

Die durchschnittliche Precision über alle verwendeten Einstellungen der manuell markierten Datensätze betrug 0,64 und für die automatisch markierten Datensätze 0,68. Das Zusammenfügen dieser verschiedenen Arten von Datensätzen erzielte mit einer durchschnittlichen Precision von 0,74 noch bessere Ergebnisse. Damit konnte die zweite Forschungsfrage beantwortet werden. Sätze in der Sprache deutsch können also mit Hilfe von maschinellen Lernverfahren erfolgreich hinsichtlich ihrer Polarität klassifiziert werden. Es ist kaum manueller Aufwand notwendig um gute Ergebnisse zu erzielen, wie an der erreichten Precision bei der Verwendung der automatisch markierten Datensätze erkennbar ist.

Die durchschnittliche Genauigkeit des Einordnens von negativen Sätzen, mit 0,66, ist schlechter als die von positiven Sätzen, mit 0,68. Außerdem fiel auf, dass unterschiedliche Einstellungen für beide Kategorien notwendig sind um jeweils die beste Precision zu erreichen. Positive Sätze konnten mit der Einstellung 1-5 lange Zeichenketten mit einer durchschnittlichen Precision von 0,72 am besten klassifiziert werden. Negative dagegen, mit einer Precision von 0,68 mit den Einstellungen 1-12 und 6-10 lange Zeichenketten klassifiziert werden. Dies könnte dadurch erklärt werden, dass wie schon bei der Erstellung des Goldstandards bemerkt wurde, positive Meinungen eher an einzelnen Schlüsselwörtern festgemacht werden können, als negative. Missstimmungen bezüglich der Produkte werden oft mit Hilfe von wünschenswerten Zuständen, also im Konjunktiv ausgedrückt. Da der Anwendungsfall dieser Arbeit den Fokus auf negativen Sätzen legte, um problematische Aspekte an Produkten erkennen zu können, wird dafür empfohlen die Einstellung 1-12 lange Zeichenketten zu verwenden. Hier liegen die Werte für die Precision beider Kategorien nah beieinander. Werden alle erreichten Evaluationsmaße der Datensätze mit den verschiedenen Einstellungen betrachtet, kann gesagt werden, dass diese empfohlene Einstellung ebenso im Durchschnitt die besten Ergebnisse im Vergleich zu den anderen Einstellungen erreichte.

Der größte Teil der Lösungsansätze für das Opinion Mining unterscheidet sich darin, dass entweder statistische oder regelbasierte Verfahren angewendet werden. In dieser Arbeit wurde ersteres analysiert und mit letzterem verglichen. Das Ergebnis davon ist, dass der Klassifikator welcher auf Wahrscheinlichkeiten für die Zugehörigkeit der N-Gramme zu den einzelnen Kategorien beruht, besser Sätze hinsichtlich ihrer Polarität ordnet, als der regelbasierte, welcher nach vorhandenen Schlüsselwörtern des Wörterbuches in den Sätzen sucht. Beide wurden auf die gleiche Testmenge angewendet. Ersterer erreichte eine durchschnittliche Precision von 0,75 und letzterer 0,62. Somit kann die dritte Forschungsfrage damit beantwortet werden, dass der Klassifikator, welcher auf statistischen Werten beruht, bessere Ergebnisse beim Erkennen von positiven und negativen Sätzen in der Sprache Deutsch erzielt, als der regelbasierte Klassifikator. Auch eine vorher durchgeführte Rechtschreibkorrektur der Testmenge für den regelbasierten Klassifikator erbrachte keine besseren Ergebnisse. Zu beachten ist dabei jedoch, dass dieser Klassifikator nicht domänenspezifisch ist, da die vorhandenen Schlüsselwörter im Wörterbuch keinen Kontext, außer Negationen und betonende Wörter, beachten. Die verwendeten Modelle des statistischen Klassifikators dagegen sind mit Sätzen aus der jeweiligen Domäne erstellt worden.

Ob dies einen maßgeblichen Einfluss auf die erzielten Ergebnisse hat wurde ebenso überprüft. Das erstellte Modell (die Kombination der manuell und automatisch markierten Sätze der Domäne Handy), welches die besten Ergebnisse lieferte, wurde auf eine technikfremde Domäne angewendet und erbrachte mit den getesteten Einstellungen eine durchschnittliche Precision von 0,61. Das Modell sowie die Testmenge besaßen die gleiche Anzahl an Kategorien, positiv und negativ. Damit bestätigt sich zwar die Annahme, dass die erstellten Modelle domänenabhängig sind, da die Precision-Werte innerhalb der Domänen im Durchschnitt 0,66 betragen. Jedoch sind diese Modelle dennoch nur marginal schlechter als der regelbasierte Klassifikator, was gleichzeitig die Beantwortung der vierten Forschungsfrage ist.

Die fünfte Forschungsfrage beinhaltet den zweiten gewählten Teilschritt dieser Arbeit für die Umsetzung des Opinion Mining. Hierbei stellte sich heraus, dass es sinnvoll ist, vorhandene Methoden aus dem Bereich der Informationsextraktion und des Text Mining dafür zu verwenden, die kategorisierten Sätze anschließend auf Inhalte zu untersuchen. Die meisten aktuellen Forschungen untersuchen Dokumente im ersten Schritt nach Entitäten und Aspekten und weisen diese Herangehensweise als ideal aus. Mit dieser Arbeit wurde gezeigt, dass dies nicht die

einzig ideale Herangehensweise ist, da das hier verwendete Vorgehen allgemeingültiger und auf vorher nicht detailliert festgelegte Produkte anwendbar ist.

Das Bilden von Problemklassen wird für nicht sinnvoll erachtet, da negativ klassifizierte Aspekte erst im Vergleich zu den positiven und gegebenenfalls objektiv beurteilten Aspekten sinnvoll beurteilt werden können. Somit enthält die prototypische Umsetzung des Verfahrens innerhalb der SemaSuite der T-Systems Multimedia Solutions GmbH als Ergebnis stattdessen eine Übersicht der Aspekte mit der dazugehörigen Anzahl des Vorkommens dieser in den vorhandenen Kategorien.

Für zukünftige Forschungen in dem Bereich Opinion Mining mit der Sprache Deutsch bleiben viele Problemstellungen ungelöst. So könnte zum Beispiel die Unterscheidung von subjektiven und objektiven Sätzen tiefgründiger betrachtet werden. In dieser Arbeit wurde versucht, dies mit Hilfe der objektiven Kategorie umzusetzen, jedoch wurden für die Klassifikation dieser Sätze keine befriedigenden Ergebnisse erzielt. Diese lagen bei einer durchschnittlichen Precision von 0,59. Der Grund dafür ist unter anderem das Fehlen dieser Kategorie im größten Teil der verwendeten Datensätze.

In dieser Arbeit wurde die Zuordnung der Meinungen zu den Aspekten ausschließlich über die Ebene des Satzes vollzogen. Dies könnte zukünftig detaillierter betrachtet werden, da mehrere Aspekte in einem Satz erwähnt werden können und dennoch die darauf bezogenen Meinungen nicht die gleichen Polaritäten besitzen müssen. Hier könnten Wörter welche die Polarität ändern, wie zum Beispiel „aber“, „und“ etc. in die Analyse einbezogen werden um Polaritätsänderungen zu erkennen.

Das Erkennen von negativen Sätzen wird mit Hilfe statistischer Verfahren besser umgesetzt als mit regelbasierten. Hier gibt es dennoch Forschungsbedarf, um insgesamt eine höhere Qualität zu erzielen. Eventuell ist es sinnvoll regelbasierte und statistische Verfahren miteinander zu kombinieren. Dies könnte zum Beispiel mit Hilfe von Synonymen umgesetzt werden. Die Sätze der Goldstandards aus der vorliegenden Arbeit könnten erweitert werden, indem die darin enthaltenen Wörter durch Synonyme aus einem Wörterbuch ersetzt werden. Daraus müssten neue Sätze gebildet und der Trainingsmenge zusätzlich hinzugefügt werden. Somit würden automatisch mehr Beispiele entstehen. Abschließend kann noch einmal betont werden, dass statistische Verfahren ein hohes Potential für das Klassifizieren von Sätzen hinsichtlich ihrer Polarität haben.

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

Alpaydin, Ethem. *Maschinelles Lernen.* Massachuseetts, 2004.

Aue, A., und M. Gamon. „Pulse: Mining Customer Opinions from Free Text.“ 2005.

Bethard, S., H. Yu, A. Thornton, V. Hatzivassiloglou, und D. Jurafsky. „Automatic extraction of opinion propositions and their holders.“ 2004.

Carstensen, K.-U., Ch. Ebert, C. Ebert, S. Jekat, R. Klabunde, und H. Langer. *Computerlinguistik und Sprachtechnologie - Eine Einführung.* Heidelberg: Spektrum Akademischer Verlag , 2010.

Choi, Y., C. Cardie, E. Riloff, und S. Patwardhan. „Identifying sources of opinions with conditional random fields and extraction patterns.“ *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), 355–362, Jg. (2005).* 2005.

Ding, X., B. Liu, und P. Yu. „A Holistic Lexicon-Based Approach to Opinion Mining.“ 2008.

Ensuli, A., und F. Sebastiani. „Determining Term Subjectivity and Term Orientation for Opinion Mining.“ *in: Proceedings of EACL-06, 11th conference of the european chapter of the association for computational linguistics.* 2006.

—. „Determining the Semantic Orientation of Terms through Gloss Classification.“ 2005.

Fellbaum, Christiane. *WordNet: an electronic lexical database.* 1998.

Ferber, Reginald. „Information Retrieval - Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web.“ Heidelberg: dpunkt.verlag GmbH, 2003.

Ganapathibhotla, M., und B. Liu. „Mining opinions in comparative sentences.“ *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), 241–248, Jg. 2008.* 2008.

Hatzivassiloglou, V., und K. McKeown. „Predicting the semantic orientation of adjectives.“ *In Proc. of ACL'97, 174-181, Jg. (1997).* 1997.

Heyer, G., U. Quasthoff, und T. Wittig. *Text Mining: Wissensrohstoff Text. Konzepte, Algorithmen, Ergebnisse.* 2006.

Hu, Y., und B. Liu. „Mining and Summarizing Customer Reviews .“ 2004.

—. „Mining Opinion Features in Customer Reviews .“ 2004.

Jindal, N., und B. Liu. „Identifying comperative sentences in text documents.“ 2006.

—. „Mining comparative sentences and relations.“ 2006.

Kamps, J., M. Maarten, R. Mokken, und M. Rijke. „Using WordNet to Measure Semantic Orientations of Adjectives.“ 2004.

Kanayama, H., und T. Nasukawa. „Fully automatic lexicon expansion for domain-oriented sentiment analysis.“ *in: Proceedings of Conference on Empirical Methods in Natural language Processing, 355-363, Jg. (2006).* 2006.

Klapdor, Marius, und Dr. Carsten Felden. „Eignung von Algorithmen zur Bereitstellung unstrukturierter Daten im Rahmen der Textklassifikation.“ Duisburg-Essen, Dezember 2005.

Klenner, Manfred, Stefano Petrakis, und Angela Fahrni. „A Tool for Polarity Classification of Human Affect from Panel Group Texts.“ Zürich University, 2009.

Koppel, M., und J. Schler. „The Importance of Neutral Examples for Learning Sentiment.“ 2005.

Li, S., C. Lin, Y. Song, und Z. Li. „Comparable entity mining from comparative questions.“ *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 650–658, Jg. (2010).* 2010.

Liu, Bing. „Sentiment Analysis and Subjectivity .“ *Handbook of Natural Language Processing, Second Edition, Jg. 2010.* 2010.

—. „Web Data Mining.“ Springer-Verlag Berlin Heidelberg, 2011.

Manning, C., P. Raghavan, und H. Schütze. *An Introduction to Information Retrieval.* 2009.

Mehler, Alexander. „Text Mining.“ *in: Hemmitzer, L.; Lobin, H. : Texttechnologie. Perspektiven und Anwendungen; 329-352; Jg. 2004.* 2004.

Mihalcea, R., C. Banea, und J. Wiebe. „Learning multilingual subjective language via cross-lingual projections.“ *in: Proceedings of the Association for Computational Linguistics (ACL), 976–983, Jg. (2007).* 2007.

Pang, B., und L. Lee. „A sentimental education Sentiment analysis using subjectivity summarization based on minimum cuts.“ 2004.

—. „Thumbs up? Sentiment Classification using Machine Learning Techniques.“ 2002.

Pang, Bo, und Lillian Lee. „Opinion Mining and Sentiment Analysis.“ 2008.

Qiu, G., B. Liu, J. Bu, und C. Chen. „Expanding Domain Sentiment Lexicon through Double Propagation.“ 2009.

—. „Opinion Word Expansion and Target Extraction through Double Propagation.“ 2010.

Remus, R., U. Quasthoff, und G. Heyer. „SentiWs - a Publicly Available German-language Resource for Sentiment Analysis.“ Leipzig, 2011.

Roth, S. „Lexikalisch-semantische Netze Anwendungsperspektiven für die Computerlinguistik .pdf.“ 2006.

Rybina, Kateryna. „Sentiment analysis of contextsaround query terms in documents.“ 2012.

Sasaki, F., und A. Witt. „Linguistische Korpora.“ In *in: Texttechnologie. Perspektiven und Anwendungen, 195-216, Jg. 2004,* von Lobin H. und L. Lemnitzer. 2004.

Sprejz, Michéle. „Meinungsanalyse zu Produkten oder Services in Social Media.“ 2011.

Tang, Huifeng, Songbo Tan, und Xueqi Cheng. „A survey on sentiment detection of reviews.“ *in: Expert Systems with Applications, 36. Jg. (2009), S. 10760-10773.* China, 2009.

Thielmann, K., und H. Paijmans. „Informationserschließung.“ In *Texttechnologie. Perspektiven und Anwendungen,* von Lobin H. und L. Lemnitzer. 2004.

Turney, P. „Thumbs up or thumbs down. semantic orientation applied to unsupervised classification of reviews.“ *in: Proceedings of Annual Meeting of the Association for Computational Linguistic, Jg. (2002).* 2002.

Turney, P., und M. Littman. „Measuring praise and criticism: Inference of semantic orientation from association.“ 2003.

Urbansky, David, Klemens Muthmann, Phillip Katz, und Sandro Reichert. „TUD Palladian.“ 2011.

Wilson, T., J. Wiebe, und R. Hwa. „Just how mad are you? Finding strong and weak opinion clauses.“ 2004.

Xia, Rui, Chengqing Zong, und Shoushan Li. „Ensemble of feature sets and classification algorithms for sentiment classification .“ *Information Sciences, 181, Jg. (2011), S. 1138-1152.* 2010.

Yu, H., und V. Hatzivassiloglou. „Towards Answering Opinion Question, Seperating Facts from Opinions and Identifying the Polarity of Opinion Sentences.“ 2003.

Zhai, Z., B. Liu, H. Xu, und P. Jia. „Clustering Product Features for Opinion Mining.“ 2011.

Zhang, L., und B. Liu. „Identifying noun product features that imply opinions .“ 2011.

Anhangsverzeichnis

Anhang

A Erster Schritt Evaluation

A.1 Einstellung 3-10 Zeichen

Trainingsmenge	Testmenge	Precision				Recall				F-Maß
		P	N	O	ø	P	N	O	ø	
HA-2g	HG-2r	0,77	0,67		0,73	0,83	0,57		0,73	0,73
HA-2g	HG-2g	0,66	0,78		0,72	0,84	0,57		0,71	0,70
NA-2g	NG-2r	0,72	0,54		0,65	0,76	0,49		0,66	0,66
NA-2g	NG-2g	0,61	0,70		0,65	0,79	0,49		0,64	0,63
80% von HG-3g	20% von HG-3g	0,64	0,54		0,62	0,54	0,68	0,60	0,61	0,61
80% von NG-3g	20% von NG-3g	0,65	0,55	0,59	0,59	0,79	0,60	0,40	0,60	0,59
80% von NG-3r	20% von NG-3r	0,53	0,57	0,57	0,55	0,86	0,18	0,37	0,54	0,50
80% von HNG-3g	80% von HNG-3g	0,67	0,58	0,64	0,63	0,67	0,69	0,52	0,62	0,62

A.2 Einstellung 3-30 Zeichen

Trainingsmenge	Testmenge	Precision				Recall				F-Maß
		P	N	O	ø	P	N	O	ø	
NA-2g	NG-2r	0,72	0,50		0,64	0,71	0,52		0,64	0,64
NA-2g	NG-2g	0,60	0,66		0,63	0,73	0,52		0,62	0,62
80% von HG-3g	20% von HG-3g	0,64	0,54	0,67	0,62	0,54	0,68	0,60	0,61	0,61
80% von NG-3g	20% von NG-3g	0,65	0,55	0,59	0,59	0,79	0,60	0,40	0,60	0,59
80% von HNG-3g	80% von HNG-3g	0,60	0,59	0,61	0,60	0,64	0,56	0,60	0,60	0,60

A.3 Einstellung Unigramme

Trainingsmenge	Testmenge	Precision				Recall				F-Maß
		P	N	O	ø	P	N	O	ø	
HA-2g	HG-2r	0,74	0,68		0,72	0,86	0,50		0,73	0,71
HA-2g	HG-2g	0,63	0,77		0,70	0,85	0,50		0,68	0,67
NA-2g	NG-2r	0,72	0,51		0,64	0,72	0,50		0,64	0,64
NA-2g	NG-2g	0,61	0,68		0,64	0,76	0,50		0,63	0,63
80% von HG-3g	20% von HG-3g	0,64	0,49	0,57	0,57	0,57	0,75	0,33	0,55	0,54
80% von NG-3g	20% von NG-3g	0,58	0,51	0,45	0,52	0,66	0,56	0,35	0,52	0,52
80% von NG-3r	20% von NG-3r	0,49	0,39	0,42	0,44	0,72	0,28	0,21	0,46	0,43
80% von HNG-3g	20% von HNG-3g	0,62	0,55	0,60	0,59	0,63	0,69	0,44	0,59	0,58
80% von HG-3r	20% von HG-3r	0,57	0,57	0,67	0,70	0,89	0,40	0,19	0,58	0,53
HA-2g	HG-3g	0,63	0,77		0,45	0,85	0,49		0,67	0,53
HA-2r	HG-2r	0,63	1,00		0,77	1,00	0,00		0,63	0,49
NA-2g	NG-3g	0,61	0,68		0,42	0,76	0,50		0,63	0,50
NA-2r	NG-2r	0,64	0,00		0,40	1,00	0,00		0,64	0,49
HA-2g	HG-3r	0,74	0,67		0,54	0,86	0,50		0,72	0,61

A.4 Einstellung 1-3 Wörter

Trainingsmenge	Testmenge	Precision				Recall				F-Maß
		P	N	O	ø	P	N	O	ø	
HA-2g	HG-2r	0,76	0,63		0,71	0,81	0,56		0,71	0,71
NA-2g	NG-2r	0,71	0,49		0,63	0,71	0,49		0,63	0,63
80% von HG-3g	20% von HG-3g	0,62	0,52	0,51	0,55	0,51	0,76	0,37	0,54	0,53
80% von NG-3g	20% von NG-3g	0,52	0,47	0,53	0,51	0,62	0,47	0,43	0,50	0,50
80% von NG-3r	20% von NG-3r	0,51	0,38	0,46	0,46	0,73	0,24	0,32	0,48	0,46
80% von HNG-3g	20% von HNG-3g	0,57	0,54	0,58	0,56	0,56	0,69	0,43	0,56	0,56
80% von Hg-3r	20% von HG-3r	0,60	0,52	0,59	0,57	0,88	0,40	0,20	0,58	0,54

B Zweiter Schritt Evaluation

B.1 Zeichenketten als N-Gramme

Datensatz HGN-3g

Einstellung	Precision				Recall				F-Maß
	P	N	O	ø	P	N	O	ø	
1-3 Zeichen	0,80	0,39	0,79	0,66	0,28	0,95	0,18	0,47	0,42
1-4 Zeichen	0,80	0,43	0,71	0,65	0,41	0,92	0,25	0,53	0,50
1-5 Zeichen	0,78	0,48	0,75	0,67	0,53	0,87	0,37	0,59	0,58
1 -6 Zeichen	0,76	0,52	0,68	0,66	0,60	0,83	0,43	0,62	0,61
1-7 Zeichen	0,71	0,52	0,67	0,63	0,63	0,76	0,44	0,61	0,61
1-8 Zeichen	0,69	0,54	0,64	0,62	0,63	0,73	0,48	0,61	0,61
1-9 Zeichen	0,69	0,56	0,64	0,63	0,63	0,71	0,51	0,62	0,62
1-10 Zeichen	0,67	0,57	0,65	0,63	0,64	0,71	0,65	0,62	0,62
1-11 Zeichen	0,67	0,59	0,65	0,64	0,66	0,67	0,57	0,63	0,63
1-12 Zeichen	0,67	0,59	0,65	0,64	0,66	0,67	0,57	0,63	0,63
1-13 Zeichen	0,65	0,58	0,62	0,62	0,63	0,63	0,57	0,61	0,61
1-14 Zeichen	0,63	0,59	0,62	0,61	0,63	0,65	0,56	0,61	0,61
1-15 Zeichen	0,66	0,59	0,65	0,63	0,65	0,64	0,60	0,63	0,63
1-16 Zeichen	0,65	0,57	0,64	0,62	0,66	0,63	0,57	0,62	0,62
1-17 Zeichen	0,67	0,59	0,62	0,63	0,66	0,63	0,59	0,62	0,62
1-18 Zeichen	0,66	0,59	0,62	0,62	0,65	0,62	0,60	0,62	0,62
1-19 Zeichen	0,65	0,57	0,59	0,61	0,64	0,60	0,58	0,61	0,61
1-20 Zeichen	0,65	0,58	0,61	0,61	0,65	0,61	0,57	0,61	0,61
2-3 Zeichen	0,79	0,40	0,77	0,66	0,30	0,95	0,19	0,48	0,44
2-4 Zeichen	0,80	0,44	0,73	0,66	0,44	0,92	0,26	0,54	0,52
2-5 Zeichen	0,78	0,48	0,74	0,67	0,53	0,87	0,37	0,59	0,58
2-6 Zeichen	0,75	0,52	0,69	0,65	0,60	0,83	0,42	0,61	0,61
2-7 Zeichen	0,71	0,52	0,67	0,63	0,63	0,76	0,44	0,61	0,61
2-8 Zeichen	0,69	0,54	0,65	0,62	0,63	0,72	0,48	0,61	0,61
2-9 Zeichen	0,68	0,56	0,63	0,62	0,63	0,71	0,51	0,62	0,61
2-10 Zeichen	0,68	0,56	0,64	0,63	0,64	0,71	0,52	0,62	0,62
2-11 Zeichen	0,67	0,58	0,66	0,63	0,66	0,68	0,55	0,63	0,63
2-12 Zeichen	0,66	0,59	0,64	0,63	0,65	0,67	0,56	0,63	0,63
2-13 Zeichen	0,64	0,58	0,62	0,61	0,63	0,63	0,58	0,61	0,61
2-14 Zeichen	0,63	0,59	0,63	0,61	0,63	0,63	0,58	0,61	0,61
2-15 Zeichen	0,64	0,59	0,64	0,62	0,64	0,64	0,58	0,62	0,62
3-4 Zeichen	0,75	0,45	0,72	0,64	0,50	0,87	0,30	0,56	0,54
3-5 Zeichen	0,78	0,49	0,69	0,65	0,56	0,86	0,37	0,60	0,59

3-6 Zeichen	0,73	0,52	0,69	0,64	0,59	0,81	0,43	0,61	0,60
3-7 Zeichen	0,71	0,53	0,67	0,64	0,63	0,76	0,45	0,62	0,61
3-8 Zeichen	0,68	0,55	0,64	0,62	0,65	0,72	0,47	0,61	0,61
3-9 Zeichen	0,67	0,56	0,65	0,63	0,65	0,71	0,49	0,62	0,62
3-10 Zeichen	0,67	0,58	0,64	0,63	0,67	0,69	0,52	0,62	0,62
3-11 Zeichen	0,65	0,56	0,65	0,62	0,64	0,68	0,52	0,62	0,62
3-12 Zeichen	0,66	0,58	0,64	0,63	0,64	0,66	0,57	0,62	0,62
3-13 Zeichen	0,65	0,57	0,61	0,61	0,65	0,63	0,54	0,61	0,61
3-14 Zeichen	0,63	0,57	0,62	0,61	0,63	0,63	0,56	0,61	0,61
3-15 Zeichen	0,64	0,58	0,61	0,61	0,64	0,63	0,56	0,61	0,61
4-5 Zeichen	0,75	0,51	0,67	0,64	0,59	0,83	0,40	0,61	0,60
4-6 Zeichen	0,75	0,52	0,68	0,65	0,61	0,82	0,42	0,62	0,61
4-7 Zeichen	0,71	0,53	0,68	0,64	0,62	0,79	0,44	0,62	0,61
4-8 Zeichen	0,68	0,55	0,65	0,63	0,63	0,75	0,46	0,62	0,61
4-9 Zeichen	0,66	0,56	0,65	0,62	0,64	0,71	0,49	0,61	0,61
4-10 Zeichen	0,66	0,58	0,63	0,63	0,66	0,70	0,51	0,62	0,62
4-11 Zeichen	0,64	0,56	0,63	0,61	0,64	0,67	0,50	0,61	0,60
4-12 Zeichen	0,63	0,55	0,62	0,60	0,63	0,64	0,52	0,60	0,59
4-13 Zeichen	0,64	0,58	0,64	0,62	0,63	0,65	0,56	0,62	0,62
4-14 Zeichen	0,63	0,57	0,61	0,60	0,63	0,62	0,55	0,60	0,60
4-15 Zeichen	0,62	0,57	0,60	0,59	0,60	0,62	0,56	0,59	0,59
5-6 Zeichen	0,72	0,53	0,67	0,64	0,61	0,79	0,43	0,61	0,61
5-7 Zeichen	0,68	0,54	0,65	0,62	0,61	0,77	0,44	0,61	0,60
5-8 Zeichen	0,66	0,55	0,63	0,61	0,62	0,73	0,45	0,60	0,60
5-9 Zeichen	0,65	0,56	0,63	0,61	0,63	0,72	0,47	0,61	0,60
5-10 Zeichen	0,63	0,56	0,63	0,61	0,62	0,70	0,48	0,60	0,60
5-11 Zeichen	0,62	0,56	0,61	0,60	0,63	0,66	0,50	0,60	0,59
5-12 Zeichen	0,61	0,54	0,59	0,58	0,63	0,63	0,48	0,58	0,58
5-10 Zeichen	0,63	0,56	0,63	0,61	0,62	0,70	0,48	0,60	0,60
6-10 Zeichen	0,62	0,55	0,61	0,59	0,63	0,66	0,47	0,59	0,58
7-10 Zeichen	0,61	0,56	0,62	0,59	0,65	0,67	0,44	0,59	0,59
8-10 Zeichen	0,62	0,55	0,60	0,59	0,65	0,65	0,47	0,59	0,59
9-10 Zeichen	0,60	0,52	0,54	0,55	0,64	0,60	0,42	0,55	0,55
5-20 Zeichen	0,64	0,58	0,59	0,61	0,67	0,60	0,55	0,61	0,61
10-20 Zeichen	0,54	0,49	0,48	0,50	0,63	0,50	0,40	0,51	0,50
15-20 Zeichen	0,38	0,46	0,41	0,42	0,75	0,26	0,19	0,40	0,37
5 Zeichen	0,74	0,52	0,66	0,64	0,62	0,79	0,41	0,61	0,60
6 Zeichen	0,66	0,53	0,64	0,61	0,60	0,75	0,44	0,60	0,59
7 Zeichen	0,61	0,54	0,60	0,58	0,62	0,71	0,40	0,57	0,57
8 Zeichen	0,60	0,53	0,62	0,59	0,64	0,68	0,40	0,58	0,57

9 Zeichen	0,62	0,54	0,59	0,58	0,67	0,63	0,44	0,58	0,58
10 Zeichen	0,57	0,54	0,54	0,55	0,63	0,59	0,42	0,55	0,54

Datensatz HA-2g

Einstellung	Precision			Recall			F-Maß
	P	N	ø	P	N	ø	
1-3 Zeichen	0,68	0,62	0,66	0,89	0,30	0,67	0,63
1-4 Zeichen	0,72	0,65	0,69	0,86	0,42	0,70	0,68
1-5 Zeichen	0,74	0,66	0,71	0,86	0,48	0,72	0,70
1 -6 Zeichen	0,75	0,67	0,72	0,85	0,51	0,72	0,71
1-7 Zeichen	0,75	0,68	0,73	0,85	0,53	0,73	0,72
1-8 Zeichen	0,76	0,67	0,73	0,84	0,55	0,73	0,73
1-9 Zeichen	0,77	0,67	0,73	0,84	0,57	0,74	0,73
1-10 Zeichen	0,77	0,67	0,73	0,83	0,57	0,74	0,73
1-11 Zeichen	0,77	0,66	0,73	0,82	0,58	0,73	0,73
1-12 Zeichen	0,77	0,66	0,73	0,82	0,58	0,73	0,73
1-13 Zeichen	0,78	0,68	0,74	0,84	0,60	0,75	0,74
1-14 Zeichen	0,77	0,67	0,73	0,83	0,59	0,74	0,73
2-3 Zeichen	0,69	0,61	0,66	0,88	0,32	0,67	0,64
2-4 Zeichen	0,72	0,64	0,69	0,85	0,43	0,70	0,68
2-5 Zeichen	0,74	0,66	0,71	0,85	0,49	0,71	0,70
2-6 Zeichen	0,75	0,66	0,71	0,84	0,52	0,72	0,71
2-7 Zeichen	0,75	0,67	0,72	0,85	0,53	0,73	0,72
2-8 Zeichen	0,76	0,67	0,73	0,84	0,56	0,74	0,73
2-9 Zeichen	0,76	0,67	0,73	0,84	0,56	0,74	0,73
2-10 Zeichen	0,77	0,67	0,73	0,83	0,57	0,74	0,73
2-11 Zeichen	0,77	0,67	0,73	0,83	0,57	0,74	0,73
2-12 Zeichen	0,76	0,66	0,72	0,83	0,57	0,73	0,73
2-13 Zeichen	0,78	0,68	0,74	0,83	0,60	0,75	0,74
2-14 Zeichen	0,77	0,67	0,74	0,83	0,59	0,74	0,74
3-4 Zeichen	0,74	0,66	0,71	0,85	0,49	0,72	0,71
3-5 Zeichen	0,75	0,66	0,72	0,84	0,52	0,72	0,72
3-6 Zeichen	0,75	0,66	0,72	0,84	0,53	0,73	0,72
3-7 Zeichen	0,76	0,67	0,73	0,84	0,54	0,73	0,72
3-8 Zeichen	0,76	0,67	0,72	0,84	0,54	0,73	0,72
3-9 Zeichen	0,76	0,67	0,73	0,83	0,57	0,73	0,73
3-10 Zeichen	0,77	0,67	0,73	0,83	0,57	0,73	0,73
3-11 Zeichen	0,77	0,67	0,73	0,83	0,58	0,74	0,73
3-12 Zeichen	0,77	0,66	0,73	0,82	0,58	0,73	0,73
3-13 Zeichen	0,78	0,67	0,74	0,83	0,60	0,74	0,74
3-14 Zeichen	0,77	0,67	0,74	0,83	0,59	0,74	0,74

4-5 Zeichen	0,76	0,66	0,72	0,83	0,56	0,73	0,72
4-6 Zeichen	0,76	0,65	0,72	0,83	0,55	0,72	0,72
4-7 Zeichen	0,75	0,66	0,72	0,83	0,54	0,73	0,72
4-8 Zeichen	0,77	0,68	0,73	0,84	0,56	0,74	0,73
4-9 Zeichen	0,76	0,67	0,73	0,84	0,56	0,74	0,73
4-10 Zeichen	0,77	0,66	0,73	0,83	0,58	0,73	0,73
4-11 Zeichen	0,77	0,67	0,73	0,83	0,57	0,73	0,73
4-12 Zeichen	0,77	0,67	0,73	0,83	0,58	0,74	0,73
4-13 Zeichen	0,77	0,67	0,74	0,83	0,59	0,74	0,74
5-6 Zeichen	0,76	0,66	0,72	0,83	0,55	0,73	0,72
5-7 Zeichen	0,76	0,67	0,73	0,84	0,55	0,73	0,73
5-8 Zeichen	0,76	0,66	0,72	0,83	0,56	0,73	0,72
5-9 Zeichen	0,77	0,68	0,73	0,84	0,57	0,74	0,73
5-10 Zeichen	0,77	0,68	0,73	0,84	0,57	0,74	0,73
5-11 Zeichen	0,77	0,68	0,74	0,84	0,59	0,74	0,73
5-12 Zeichen	0,77	0,68	0,74	0,84	0,59	0,74	0,74
5-10 Zeichen	0,77	0,68	0,73	0,84	0,57	0,74	0,73
6-10 Zeichen	0,78	0,69	0,74	0,85	0,58	0,75	0,74
7-10 Zeichen	0,77	0,66	0,73	0,83	0,57	0,73	0,73
8-10 Zeichen	0,77	0,67	0,73	0,83	0,57	0,73	0,73
9-10 Zeichen	0,77	0,68	0,74	0,84	0,59	0,74	0,74
5 Zeichen	0,76	0,66	0,72	0,83	0,55	0,73	0,72
6 Zeichen	0,76	0,67	0,73	0,84	0,56	0,73	0,73
7 Zeichen	0,76	0,66	0,72	0,83	0,56	0,73	0,72
8 Zeichen	0,76	0,67	0,73	0,84	0,56	0,73	0,73
9 Zeichen	0,77	0,68	0,74	0,84	0,58	0,74	0,74
10 Zeichen	0,77	0,66	0,73	0,82	0,58	0,73	0,73

B.2 Wörter als N-Gramme

Datensatz HNG-3g

Einstellung	Precision				Recall				F-Maß
	P	N	O	ø	P	N	O	ø	
Unigram	0,62	0,55	0,60	0,59	0,63	0,69	0,44	0,59	0,58
2 Wörter	0,42	0,45	0,47	0,44	0,48	0,48	0,36	0,44	0,44
3 Wörter	0,33	0,49	0,50	0,44	0,79	0,17	0,14	0,37	0,31
4 Wörter	0,33	0,75	0,33	0,47	0,98	0,02	0,01	0,34	0,19

2-3 Wörter	0,43	0,46	0,47	0,45	0,52	0,49	0,33	0,45	0,44
2-4 Wörter	0,43	0,46	0,47	0,45	0,52	0,49	0,33	0,45	0,45
2-5 Wörter	0,43	0,46	0,47	0,45	0,52	0,49	0,33	0,45	0,45
2-6 Wörter	0,43	0,46	0,47	0,45	0,52	0,49	0,33	0,45	0,45
1-2 Wörter	0,59	0,54	0,61	0,58	0,59	0,70	0,43	0,57	0,57
1-3 Wörter	0,57	0,54	0,58	0,56	0,56	0,69	0,43	0,56	0,56
1-4 Wörter	0,57	0,54	0,58	0,56	0,57	0,69	0,42	0,56	0,56
1-5 Wörter	0,57	0,54	0,58	0,56	0,57	0,69	0,42	0,56	0,56
1-6 Wörter	0,57	0,54	0,58	0,56	0,57	0,69	0,42	0,56	0,56
1-7 Wörter	0,57	0,54	0,58	0,56	0,57	0,69	0,42	0,56	0,56
1-8 Wörter	0,57	0,54	0,58	0,56	0,57	0,69	0,42	0,56	0,56
3-4 Wörter	0,33	0,49	0,50	0,44	0,79	0,17	0,14	0,37	0,31
3-5 Wörter	0,33	0,49	0,50	0,44	0,79	0,17	0,14	0,37	0,31
3-6 Wörter	0,33	0,49	0,50	0,44	0,79	0,17	0,14	0,37	0,31

Datensatz HA-2g

Einstellung	Precision			Recall			F-Maß
	P	N	ø	P	N	ø	
Unigram	0,74	0,67	0,72	0,86	0,50	0,72	0,71
2 Wörter	0,74	0,57	0,68	0,76	0,55	0,68	0,68
3 Wörter	0,70	0,56	0,65	0,80	0,44	0,66	0,65
4 Wörter	0,65	0,58	0,63	0,94	0,15	0,64	0,57
2-3 Wörter	0,75	0,58	0,69	0,75	0,58	0,69	0,69
2-4 Wörter	0,75	0,57	0,68	0,74	0,57	0,68	0,68
2-5 Wörter	0,75	0,57	0,68	0,74	0,57	0,68	0,68
2-6 Wörter	0,75	0,57	0,68	0,74	0,57	0,68	0,68
1-2 Wörter	0,75	0,63	0,71	0,81	0,55	0,72	0,71
1-3 Wörter	0,76	0,63	0,71	0,81	0,56	0,71	0,71
1-4 Wörter	0,76	0,63	0,71	0,81	0,56	0,72	0,71
1-5 Wörter	0,76	0,63	0,71	0,80	0,56	0,71	0,71
1-6 Wörter	0,76	0,63	0,71	0,80	0,56	0,71	0,71
1-7 Wörter	0,76	0,63	0,71	0,80	0,56	0,71	0,71
1-8 Wörter	0,76	0,63	0,71	0,80	0,56	0,71	0,71
3-4 Wörter	0,71	0,57	0,66	0,80	0,45	0,67	0,66
3-5 Wörter	0,71	0,57	0,66	0,80	0,45	0,67	0,66
3-6 Wörter	0,71	0,57	0,66	0,80	0,44	0,67	0,66

C Dritter Schritt der Evaluation

C.1 Einstellung 1-5 Zeichen

Datensatz	Precision				Recall				F-Maß
	P	N	O	ø	P	N	O	ø	
HG-3g	0,83	0,46	0,59	0,63	0,46	0,92	0,25	0,54	0,52
NG-3g	0,73	0,49	0,60	0,61	0,66	0,74	0,35	0,58	0,57
HNG-3g	0,78	0,48	0,75	0,67	0,53	0,87	0,37	0,59	0,58
HA-2g	0,74	0,66	-	0,71	0,86	0,48	-	0,72	0,70
NA-2g	0,72	0,53	-	0,65	0,75	0,49	-	0,66	0,65
ø	0,76	0,52	0,65	0,65	0,65	0,70	0,33	0,62	0,61

C.2 Einstellung 1-12 Zeichen

Datensatz	Precision				Recall				F-Maß
	P	N	O	ø	P	N	O	ø	
HG-3g	0,63	0,56	0,67	0,62	0,54	0,63	0,68	0,62	0,62
NG-3g	0,65	0,60	0,54	0,60	0,81	0,60	0,41	0,61	0,60
HNG-3g	0,67	0,59	0,65	0,64	0,66	0,67	0,57	0,63	0,63
HA-2g	0,77	0,66	-	0,73	0,82	0,58	-	0,73	0,73
NA-2g	0,71	0,51	-	0,63	0,74	0,47	-	0,64	0,64
ø	0,69	0,58	0,62	0,64	0,71	0,59	0,55	0,65	0,64

C.3 Einstellung 6-10 Zeichen

Datensatz	Precision				Recall				F-Maß
	P	N	O	ø	P	N	O	ø	
HG-3g	0,61	0,52	0,61	0,58	0,52	0,70	0,49	0,57	0,57
NG-3g	0,62	0,61	0,57	0,60	0,78	0,59	0,44	0,60	0,60
HNG-3g	0,62	0,55	0,61	0,59	0,63	0,66	0,47	0,59	0,58
HA-2g	0,77	0,69	-	0,74	0,85	0,58	-	0,75	0,74
NA-2g	0,72	0,53	-	0,65	0,76	0,48	-	0,66	0,65
ø	0,67	0,58	0,60	0,63	0,71	0,60	0,47	0,63	0,63

C.4 Einstellung Unigramme

Datensatz	Precision				Recall				F-Maß
	P	N	O	ø	P	N	O	ø	
HG-3g	0,64	0,49	0,57	0,57	0,57	0,75	0,33	0,55	0,54
NG-3g	0,58	0,51	0,45	0,52	0,66	0,56	0,35	0,52	0,52
HNG-3g	0,62	0,55	0,60	0,59	0,63	0,69	0,44	0,59	0,58
HA-2g	0,74	0,67	-	0,72	0,86	0,50	-	0,72	0,71
NA-2g	0,72	0,51	-	0,64	0,72	0,50	-	0,64	0,64
ø	0,66	0,55	0,54	0,61	0,69	0,60	0,37	0,61	0,60

C.5 Einstellung 1-2 Wörter

Datensatz	Precision				Recall				F-Maß
	P	N	O	ø	P	N	O	ø	
HG-3g	0,63	0,50	0,56	0,56	0,52	0,75	0,38	0,55	0,54
NG-3g	0,54	0,49	0,57	0,53	0,66	0,50	0,43	0,53	0,53
HNG-3g	0,59	0,54	0,61	0,58	0,59	0,70	0,43	0,57	0,57
HA-2g	0,75	0,63	-	0,71	0,81	0,55	-	0,72	0,71
NA-2g	0,71	0,50	-	0,63	0,72	0,49	-	0,64	0,63
ø	0,65	0,53	0,58	0,60	0,66	0,60	0,41	0,60	0,60

C.6 Einstellung 1-4 Wörter

Datensatz	Precision				Recall				F-Maß
	P	N	O	ø	P	N	O	ø	
HG-3g	0,61	0,52	0,51	0,55	0,52	0,76	0,35	0,54	0,53
NG-3g	0,52	0,47	0,53	0,51	0,62	0,47	0,43	0,50	0,50
HNG-3g	0,57	0,54	0,58	0,56	0,57	0,69	0,42	0,56	0,56
HA-2g	0,76	0,63	-	0,71	0,81	0,56	-	0,72	0,71
NA-2g	0,71	0,49	-	0,63	0,71	0,49	-	0,63	0,63
ø	0,63	0,53	0,54	0,59	0,65	0,59	0,40	0,59	0,59

D Datensatzkombinationen

D.1 HA&HG-2g

Einstellung	Precision			Recall			F-Maß
	P	N	ø	P	N	ø	
1-5 Zeichen	0,67	0,78	0,72	0,83	0,59	0,71	0,71
1-12 Zeichen	0,74	0,82	0,78	0,85	0,70	0,77	0,77
6-10 Zeichen	0,76	0,83	0,79	0,85	0,73	0,79	0,79
Unigramme	0,68	0,84	0,76	0,89	0,58	0,73	0,73
1-2 Wörter	0,74	0,84	0,79	0,87	0,69	0,78	0,78
1-4 Wörter	0,73	0,83	0,78	0,86	0,68	0,77	0,77
ø	0,72	0,82	0,77	0,86	0,66	0,76	0,76

D.2 HA&HNG-2g

Einstellung	Precision			Recall			F-Maß
	P	N	ø	P	N	ø	
1-5 Zeichen	0,68	0,80	0,74	0,86	0,59	0,72	0,72
1-12 Zeichen	0,69	0,81	0,75	0,86	0,61	0,73	0,73
6-10 Zeichen	0,69	0,81	0,75	0,86	0,62	0,74	0,73
Unigramme	0,64	0,77	0,71	0,84	0,54	0,69	0,68
1-2 Wörter	0,68	0,79	0,74	0,84	0,61	0,72	0,72
1-4 Wörter	0,68	0,79	0,73	0,84	0,60	0,72	0,72
ø	0,68	0,80	0,74	0,85	0,59	0,72	0,72

D.3 NA&NG-2g

Einstellung	Precision			Recall			F-Maß
	P	N	ø	P	N	ø	
1-5 Zeichen	0,72	0,77	0,74	0,79	0,69	0,74	0,74
1-12 Zeichen	0,67	0,77	0,72	0,82	0,60	0,71	0,71
6-10 Zeichen	0,65	0,73	0,69	0,79	0,57	0,68	0,68
Unigramme	0,66	0,73	0,69	0,78	0,60	0,69	0,68
1-2 Wörter	0,65	0,69	0,67	0,73	0,61	0,67	0,67
1-4 Wörter	0,64	0,66	0,65	0,69	0,61	0,65	0,65
ø	0,66	0,73	0,69	0,77	0,61	0,69	0,69

D.4 HA&NA&HNG-2g

Einstellung	Precision			Recall			F-Maß
	P	N	ø	P	N	ø	
1-5 Zeichen	0,66	0,79	0,73	0,86	0,56	0,71	0,70
1-12 Zeichen	0,70	0,83	0,76	0,87	0,62	0,75	0,74
6-10 Zeichen	0,69	0,81	0,76	0,86	0,63	0,74	0,73
Unigramme	0,65	0,77	0,71	0,83	0,54	0,69	0,68
1-2 Wörter	0,70	0,82	0,76	0,86	0,63	0,75	0,74
1-4 Wörter	0,69	0,82	0,76	0,87	0,61	0,74	0,73
ø	0,68	0,81	0,74	0,86	0,60	0,73	0,72

D.5 HG-2g

Einstellung	Precision			Recall			F-Maß
	P	N	ø	P	N	ø	
1-5 Zeichen	0,84	0,60	0,72	0,38	0,93	0,65	0,63
1-12 Zeichen	0,71	0,66	0,69	0,62	0,75	0,68	0,68
6-10 Zeichen	0,75	0,66	0,71	0,59	0,80	0,70	0,69
Unigramme	0,71	0,64	0,67	0,58	0,76	0,67	0,67
1-2 Wörter	0,73	0,67	0,70	0,62	0,77	0,70	0,70
1-4 Wörter	0,75	0,67	0,71	0,62	0,79	0,70	0,70
ø	0,75	0,65	0,70	0,57	0,80	0,68	0,68

D.7 NG-2g

Einstellung	Precision			Recall			F-Maß
	P	N	ø	P	N	ø	
1-5 Zeichen	0,86	0,72	0,79	0,66	0,90	0,78	0,77
1-12 Zeichen	0,77	0,81	0,79	0,82	0,76	0,79	0,79
6-10 Zeichen	0,71	0,78	0,74	0,81	0,67	0,74	0,74
Unigramme	0,73	0,73	0,73	0,73	0,73	0,73	0,73
1-2 Wörter	0,66	0,66	0,66	0,66	0,66	0,66	0,66
1-4 Wörter	0,66	0,67	0,66	0,67	0,66	0,66	0,66
ø	0,73	0,73	0,73	0,72	0,73	0,73	0,72

D.8 HNG-2g

Einstellung	Precision			Recall			F-Maß
	P	N	ø	P	N	ø	
1-5 Zeichen	0,86	0,64	0,75	0,47	0,93	0,70	0,68
1-12 Zeichen	0,76	0,73	0,75	0,72	0,78	0,75	0,75
6-10 Zeichen	0,78	0,73	0,76	0,70	0,81	0,75	0,75
Unigramme	0,76	0,69	0,73	0,65	0,80	0,72	0,72
1-2 Wörter	0,76	0,69	0,73	0,64	0,80	0,72	0,72
1-4 Wörter	0,77	0,69	0,73	0,64	0,81	0,72	0,72
ø	0,78	0,70	0,74	0,64	0,82	0,73	0,72

D.9 Trainingsmenge ohne objektiv auf Testmenge mit objektiv

Einstellung	Precision			Recall			F-Maß
	P	N	ø	P	N	ø	
1-5 Zeichen	0,86	0,64	0,57	0,48	0,92	0,70	0,56
1-12 Zeichen	0,77	0,74	0,51	0,72	0,79	0,75	0,61
6-10 Zeichen	0,79	0,73	0,51	0,70	0,81	0,75	0,61
Unigramme	0,76	0,70	0,51	0,66	0,79	0,73	0,59
1-2 Wörter	0,76	0,70	0,50	0,66	0,79	0,73	0,59
1-4 Wörter	0,77	0,70	0,51	0,65	0,81	0,73	0,59
ø	0,79	0,70	0,52	0,64	0,82	0,73	0,59

D.10 Trainings- und Testmenge mit objektiv

Einstellung	Precision				Recall				F-Maß
	P	N	O	ø	P	N	O	ø	
1-5 Zeichen	0,73	0,47	0,71	0,64	0,51	0,87	0,32	0,57	0,55
1-12 Zeichen	0,62	0,58	0,63	0,61	0,62	0,69	0,51	0,61	0,60
6-10 Zeichen	0,58	0,53	0,62	0,58	0,60	0,69	0,42	0,57	0,57
Unigramme	0,58	0,53	0,56	0,56	0,60	0,67	0,40	0,55	0,55
1-2 Wörter	0,57	0,51	0,58	0,56	0,56	0,70	0,38	0,55	0,54
1-4 Wörter	0,56	0,52	0,56	0,55	0,56	0,71	0,37	0,54	0,54
ø	0,61	0,52	0,61	0,58	0,57	0,72	0,40	0,56	0,56

E Threshold-Analyse

Thres-hold	Precision							Recall			F-Maß	Gesamtanzahl		
	Positiv			Negativ										
	P	Richtig einge-ordnet	ein ge-or d-net	N	Richtig einge-ordnet	ein-ge-ord-net	Ø	P	N	Ø		ein-ge-ord-net	Richtig einge-ordnet	Reale Anzahl
0	0,74	60	81	0,82	50	61	0,78	0,85	0,70	0,77	0,77	142	110	142
0,1	0,74	60	81	0,82	50	61	0,78	0,85	0,70	0,77	0,77	142	110	142
0,2	0,74	60	81	0,82	50	61	0,78	0,85	0,70	0,77	0,77	142	110	142
0,3	0,74	60	81	0,82	50	61	0,78	0,85	0,70	0,77	0,77	142	110	142
0,4	0,74	60	81	0,82	50	61	0,78	0,85	0,70	0,77	0,77	142	110	142
0,5	0,74	60	81	0,82	50	61	0,78	0,85	0,70	0,77	0,77	142	110	142
0,51	0,81	58	72	0,85	47	55	0,83	0,82	0,66	0,74	0,78	127	105	142
0,52	0,81	57	70	0,86	42	49	0,84	0,80	0,59	0,70	0,76	119	99	142
0,53	0,82	55	67	0,87	41	47	0,85	0,77	0,58	0,68	0,75	114	96	142
0,54	0,84	51	61	0,89	41	46	0,86	0,72	0,58	0,65	0,74	107	92	142
0,55	0,84	49	58	0,88	36	41	0,86	0,69	0,51	0,60	0,71	99	85	142
0,56	0,87	47	54	0,89	34	38	0,88	0,66	0,48	0,57	0,69	92	81	142
0,57	0,92	45	49	0,91	30	33	0,91	0,63	0,42	0,53	0,67	82	75	142
0,58	0,96	43	45	0,92	24	26	0,94	0,61	0,34	0,47	0,63	71	67	142
0,59	0,95	41	43	0,92	23	25	0,94	0,58	0,32	0,45	0,61	68	64	142
0,6	0,97	36	37	0,92	23	25	0,95	0,51	0,32	0,42	0,58	62	59	142
0,61	1,00	29	29	0,90	18	20	0,95	0,41	0,25	0,33	0,49	49	47	142
0,62	1,00	26	26	0,93	13	14	0,96	0,37	0,18	0,27	0,43	40	39	142
0,63	1,00	24	24	0,92	11	12	0,96	0,34	0,15	0,25	0,39	36	35	142
0,64	1,00	24	24	0,89	8	9	0,94	0,34	0,11	0,19	0,32	33	32	142
0,65	1,00	19	19	1,00	7	7	1,00	0,27	0,10	0,18	0,31	26	26	142

F Domänenabhängigkeit

F.1 Modell Domäne Handy auf Blogeinträge ohne objektiv

Einstellung	Precision			Recall			F-Maß
	P	N	ø	P	N	ø	
1-5 Zeichen	0,63	0,66	0,64	0,69	0,59	0,64	0,64
1-12 Zeichen	0,66	0,65	0,66	0,65	0,66	0,66	0,66
6-10 Zeichen	0,63	0,62	0,63	0,62	0,63	0,63	0,62
Unigramme	0,56	0,56	0,56	0,55	0,57	0,56	0,56
1-2 Wörter	0,59	0,58	0,59	0,57	0,60	0,59	0,59
1-4 Wörter	0,58	0,58	0,58	0,58	0,59	0,58	0,58
ø	0,61	0,61	0,61	0,61	0,61	0,61	0,61

F.2 Modell Domäne Handy auf Blogeinträge mit objektiv

Einstellung	Precision			Recall			F-Maß
	P	N	ø	P	N	ø	
1-5 Zeichen	0,63	0,66	0,43	0,69	0,59	0,64	0,51
1-12 Zeichen	0,66	0,65	0,44	0,65	0,66	0,66	0,53
6-10 Zeichen	0,63	0,62	0,42	0,62	0,63	0,63	0,50
Unigramme	0,56	0,56	0,38	0,55	0,57	0,56	0,45
1-2 Wörter	0,59	0,58	0,39	0,57	0,60	0,59	0,47
1-4 Wörter	0,58	0,58	0,39	0,58	0,59	0,58	0,47
ø	0,61	0,61	0,41	0,61	0,61	0,61	0,49

F.3 Modell Domäne Handy auf Domäne Notebook

Einstellung	Precision			Recall			F-Maß
	P	N	ø	P	N	ø	
1-5 Zeichen	0,60	0,84	0,72	0,93	0,39	0,66	0,63
1-12 Zeichen	0,60	0,73	0,67	0,84	0,45	0,64	0,63
6-10 Zeichen	0,59	0,71	0,65	0,82	0,43	0,63	0,61
Unigramme	0,58	0,68	0,63	0,81	0,42	0,61	0,60
1-2 Wörter	0,63	0,73	0,68	0,81	0,52	0,66	0,66
1-4 Wörter	0,63	0,73	0,68	0,81	0,52	0,66	0,66
ø	0,61	0,74	0,67	0,83	0,46	0,64	0,63

F.4 Modell Domäne Notebook auf Domäne Handy

Einstellung	Precision			Recall			F-Maß
	P	N	ø	P	N	ø	
1-5 Zeichen	0,73	0,63	0,68	0,52	0,80	0,66	0,66
1-12 Zeichen	0,67	0,64	0,66	0,61	0,70	0,65	0,65
6-10 Zeichen	0,66	0,65	0,66	0,63	0,68	0,65	0,65
Unigramme	0,70	0,68	0,69	0,66	0,72	0,69	0,69
1-2 Wörter	0,68	0,68	0,68	0,68	0,68	0,68	0,68
1-4 Wörter	0,67	0,69	0,68	0,70	0,66	0,68	0,68
ø	0,68	0,66	0,67	0,63	0,71	0,67	0,67

G Vergleich mit Wörterbuch-Ansatz

G.1 Domäne Handy

	Precision			Recall			F-Maß
	P	N	ø	P	N	ø	
Wörterbuch	0,746	0,4647	0,61	0,509	0,868	0,69	0,6446
Goldstandard	0,71	0,66	0,69	0,62	0,75	0,68	0,68
Kombination	0,74	0,82	0,78	0,85	0,70	0,77	0,77

G.2 Domäne Notebook

	Precision			Recall			F-Maß
	P	N	ø	P	N	ø	
Wörterbuch	0,73	0,522	0,63	0,538	0,81	0,67	0,65
Goldstandard	0,77	0,81	0,79	0,82	0,76	0,79	0,79
Kombination	0,67	0,77	0,72	0,82	0,60	0,71	0,71

G.3 Domänen Handy und Notebook

	Precision			Recall			
	P	N	ø	P	N	ø	F-Maß
Wörterbuch	0,739	0,49	0,61	0,52	0,839	0,68	0,6469
Goldstandard	0,76	0,73	0,75	0,72	0,78	0,75	0,75
Kombination	0,70	0,83	0,76	0,87	0,62	0,75	0,74

G.4 Domäne Handy mit Vorverarbeitung

Vorver-arbeitung	Precision			Recall			F-Maß
	P	N	ø	P	N	ø	
originaler Satz	0,75	0,46	0,61	0,51	0,87	0,69	0,64
korrigierter Satz	0,73	0,49	0,61	0,51	0,88	0,69	0,65

G.5 Domäne Notebook mit Vorverarbeitung

Vorver-arbeitung	Precision			Recall			F-Maß
	P	N	ø	P	N	ø	
originaler Satz	0,73	0,52	0,63	0,54	0,81	0,67	0,65
korrigierter Satz	0,73	0,54	0,63	0,54	0,82	0,68	0,66

G.6 Domäne Handy und Notebook mit Vorverarbeitung

Vorver-arbeitung	Precision			Recall			F-Maß
	P	N	ø	P	N	ø	
originaler Satz	0,74	0,49	0,61	0,52	0,84	0,68	0,65
korrigierter Satz	0,73	0,51	0,62	0,52	0,85	0,68	0,65