Greber, Andre:

Eine Modellbasierte Klassifikation von Kundenbewertungen am Beispiel von Hotelbewertungen aus dem Internet

Masterarbeit, Sommersemester 2017

Fakultät für Mathematik, Informatik und Statistik

Ludwig-Maximilians-Universität München

https://doi.org/10.5282/ubm/epub.41013

Masterarbeit

Eine Modellbasierte Klassifikation von Kundenbewertungen am Beispiel von Hotelbewertungen aus dem Internet

Andre Greber

Ludwig-Maximilians-Universität
München
Institut für Statistik
Lehrstuhl für Statistik Methodik und Anwendungen in
Wirtschafts-, Sozial- und Lebenswissenschaften

Masterarbeit

Thema: Sentiment Analyse

eingereicht von: Andre Greber
Matrikelnummer: 6041935
E-mail: andre.greber@gmx.de

eingereicht am: 9. Juni 2017

Betreuer: Herr Prof. Dr. Christian Heumann

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

Unternehmen nutzen verstärkt das Vorhandensein und die schnelle Abrufbarkeit von spezifischen Informationen, wie Produktbewertungen, im Internet. Diese werden verwendet um beispielsweise die Haltung der Kunden gegenüber bestimmten Produkten zu untersuchen, oder die Unternehmensstrategie nach der vorrherrschenden Meinung am Markt auszurichten. Privatpersonen informieren sich vor dem Kauf über die Vor- und Nachteile eines Produkts. Die im Netz verfügbaren Informationen können zur Entscheidungen der Konsumenten beitragen. Beispielsweise entscheidet sich ein potentieller Kunde aufgrund guter, oder schlechter, Bewertungen für, oder gegen, den Kauf eines Produkts, Pang and Lee, 2008.
Meinungen die der eigenen Grundeinstellung am nächsten sind, werden vom Nutzer oftmals stärker gewichtet, als Meinungen die zwar objektiv sind, aber nicht mit der eigenen Einstellung harmonieren, Liu and Zhang, 2012. Dadurch kann sich die subjektive Einschätzung einer bestimmten Thematik verfälschen. Aufgrund des großen Angebots im Internet ist es oftmals für den einzelnen Nutzer schwer sämtliche Information zu Überblicken, z.B. das Relevante vom nicht Relevanten zu unterscheiden. Daneben ist es für den einzelnen sehr zeitaufwendig die Daten so aufzubereiten, dass sich die Durchschnittsmeinung oder die Verteilung der Meinungen der Gesamten Bewertungen herauskristallisieren, Liu and Zhang, 2012.
Auf dieser Grundlage ist es sinnvoll eine objektive Hilfestellung anzubieten, welche die nützlichen Informationen herausfiltert und in aufbereiteter Form darstellt. Statistische Methoden wie beispielsweise Support Vector Machines, Random Forrests oder Neuronale Netzwerke können dafür ein Mittel zur Hilfestellung sein.

Als Basis der Untersuchungen dieser Arbeit dienen Hotelbwertungen des Hotelbewertungsportals *holidaycheck.de*.

Als Produkt wird das Hotel und die im Rahmen der Hotelerie angebotenen Dienstleistungen, wie Zimmerservice, oder Buffets betrachtet.

Die Arbeit ist wie folgt aufgebaut, im ersten Teil der Arbeit wird der theoretische Hintergrund thematisiert. Kapitel 3 definiert die Polarität einer Aussage und zeigt verschiedene Einflüsse auf die Polarität einer Aussage auf. Kapitel 4 beinhaltet die Beurteilung der Qualität einer Hotelbewertung und die Erkennung von Spam. Kapitel 5 und Kapitel 6 befassen sich mit der Erstellung von Sentiment Lexika und der semantischen Orientierung.
Kapitel 7 beinhaltet eine deskritive Beschreibung des Datensatzes. In Kapitel 8 wird einmal der Einfluss der Anzahl der Wörter auf die Polarität der Kritik untersucht. Zum zweiten der Einfluss der Polarität auf die Anzahl der Wörter einer Kritik. Die darauf folgenden Kapitel umfassen die überwachte Klassifikation von Hotelbewertungen mithilfe verschiedener Ansätze. Kapitel 11 dient als Einschub. Es werden Wortvektoren eingeführt, welche in den darauffolgenden Methoden verwendet werden. Kapitel 13 gibt einen Ausblick auf weitere Methoden im Bereich der Neuronalen Netzwerke. Das letzte Kapitel dient als Diskussion und Zusammenfassung.

Kapitel 2

Theoretischer Hintergrund

2.1 Definition und Anwendungsgebiete

Text Mining bezeichnet das Feld der computergestützten, auf Algorithmen basierten, Untersuchung von Texten um bestimmte Informationen zu extrahieren, Aggarwal and Zhai, 2012. Die Sentiment Analyse kann als Teilgebiet des Text Minings angesehen werden. Dabei wird das Ziel verfolgt aufgrund des Textes, z.B. einem Buch, einer Produktbewertung, oder einer SMS eine Aussage über die Emotionen oder Haltung des Autors zu einem bestimmten Thema zu treffen, Liu and Zhang, 2012.

Folgende Artverwandte Begriffe finden ebenfalls Verwendung, 'Opinion Mining', 'Subjectivity Analysis', 'Review Mining', oder 'Appraisal Extraction'. Dabei bezieht sich der verwendete Begriff auf das Forschungsgebiet. Beispielsweise zeigt die Verwendung des Begriffs, Sentiment Analyse an, dass die Forschung aus dem Gebiet der Computerlinguistik stammt.

2.2 Kundenbewertungen

Hotelbewertungen können genutzt werden um das Hotel bezüglich der eigenen Präferenzen auszuwählen, wie z.B. einen großen Pool oder ein opulentes Frühstücksbuffet. Zur Entscheidung tragen die Bewertungen vorheriger Gäste der Hotels bei.
Um die gewünschten Informationen zu erlangen kann ein potentieller Gast spezielle Onlineportale heranziehen, hier werden Bewertungen zu beispielsweise Produkten oder Dienstleistungen gesammelt und frei zugänglich zur Verfügung gestellt. Onlineportale lassen sich in Zwei Arten auftei-

len. Expertenbewertungsseiten, welche ausschließlich Meinungen von Experten veröffentlichen, wie `www.reviews.cnet.com`. Zum anderen Verbraucherportale wie `www.opodo.de`, die Bewertungen von privaten Verbrauchern veröffentlichen, vgl.Wolfgruber, 2015.

Alle Plattformen bieten ein ähnliches Punktesystem wie „`www.opodo.de`“ an, hierdurch können die verschiedenen Hotels bewertet werden. Darüber kristallisiert sich dann eine Hierarchie der Bewertungen oder sogar zwischen verschiedenen Hotels hinsichtlich deren Benotung heraus.
Zusätzlich wird die Möglichkeit angeboten die unterschiedlichen Aspekte, wie beispielsweise die Lage des Hotels, die Ausstattung oder das Hotelpersonal zu evaluieren.
Mit dem Verfassen einer Hotelbewertung gibt ein Gast, der Meinungsträger, seine subjektive Meinung wieder.

Eine Bewertung besteht aus (a) Dem Bewertungstitel, der den Gesamteindruck des Gastes in verkürzter Form angibt. (b) Der Gesamtnote des Gastes. (c) Dem Aspekt, dieser kann das Hotel oder auch nur einzelne Dienstleistungen darstellen. (d) Dem Bewertungskommentar, der Gast kann hier die Gründe der Bewertung detailliert erörtern. (e) Die Bewertungsqualität, die Angabe von anderen Nutzern, oder des Portals wie hilfreich diese Bewertung angesehen werden kann, Wolfgruber, 2015.

Die Informationen die aus den Bewertungen hervorgehen lassen sich nach Liu and Zhang, 2012 in übersichtlicher Form in Quintupel darstellen. Diese bestehen aus, (a) der Entität e, dem Zielobjekt der Bewertung, beispielsweise einem Produkt oder einer Dienstleistung, wie einem Frühstücksbuffet oder dem Zimmerservice. Jeder Entität wird ein Paar zugeordnet, (T, W). Wobei T die einzelnen Komponenten oder Teilkomponenten bezeichnet. Komponenten sind beispielsweise die Eierspeisen und Teilkomponenten die Rühreier. W fasst die zugehörigen Attribute zusammen. Attribute geben die Qualitätsmerkmale oder Eigenschaften der einzelnen Komponenten oder Subkomeponenten an. Beispiele hierfür sind, das Aussehen, die Temperatur, oder auch die Größe des Buffets. (b) Zur Vereinfachung der Darstellung des Quintupels wird das Paar (T, W) zusammengefasst betrachtet und als Aspekt a bezeichnet. (c) Die Orientierung der Meinung des Autors wird mit oo bezeichnet. Dies gibt an ob der Meinungsträger den Aufenthalt als positiv oder negativ empfunden hat. (d) Die Angabe des Autors der Kundenbewertung, h. Diese Information ist für einige Analysemethoden, wie die Spamanalyse elementar. Hierauf wird im späteren Verlauf der Arbeit noch einmal genauer eingegangen. (e) Der Zeitpunkt zu welchem die Kundenbewertung verfasst

wurde, diese Information kann z.B. einem Unternehmen Aufschluss darüber geben wie sich die Wahrnehmnung der Kunden eines Produkts über einen bestimmten Zeitverlauf verändert hat.
Eine Angabe der einzelnen Elemente wie der Entität oder der Komponenten, sowie der Orientierung ist für weitere Analysen essentiell. Fehlt beispielsweise die Entität oder die Orientierung ist die Information nicht weiter verwertbar.

2.3 Die Ebene der Betrachtung

Es herrschen drei Ebenen der Klassifikation vor, Wolfgruber, 2015. Zum einen die Dokumenten-Ebene, Satz-Ebene und Aspekt-Ebene.
Zusätzlich nehmen Liu and Zhang, 2012 an das eine Kundenbewertung nur eine Entität und einen Autor aufweist.

Aspekt-Ebene

Hierbei wird laut Kiritchenko et al., 2014 ein einziges Wort, oder eine Kombination aus Wörtern, wie Redewendungen, hinsichtlich deren Einfluss auf die Orientierung untersucht. Die Orientierung definiert die positive oder negative Ausrichtung der Haltung des Meinungsträgers bezogen auf eine Entität. Dabei ist zu erwähnen, dass Begriffe je nach Kontext eine andere Orientierung aufweisen können. Beispielsweise ist der folgende Satz, 'Ein alter Wein.' unter Umständen ein Qualitätsmerkmal und daher als positiv einzustufen. Hingegen ist dieser Satz in einem anderen Kontext als negativ zu werten.

Satz-Ebene

Im Gegensatz zur Aspekt Ebene wird der gesamte Satz hinsichtlich dessen Orientierung untersucht. Liu and Zhang, 2012 schlagen eine Identifikation in zwei Schritten vor.
Im ersten Schritt wird der vorliegende Satz als objektiv oder subjektiv eingestuft. Auf die Unterscheidung zwischen objektiven und subjektiven Sätzen wird im späteren Verlauf der Arbeit genauer eingegangen.
Im zweiten Schritt wird der subjektive Satz bezüglich dessen Orientierung untersucht. Es können jedoch komplexere Satzstrukturen vorliegen, wobei mehrere Komponenten einer Entität bewertet oder verglichen werden. Auch sind Sätze möglich die mehrere Orientierungen beinhalten. Ein Satz kann ebenfalls aus einem subjektiven und objektiven Teil bestehen, wobei nicht auszuschließen ist, dass aus dem objektiven Teil Rückschlüsse auf die Emotionen des Autors möglich sind.

Dokumenten-Ebene

Die dritte Ebene bildet die Klassifikation des gesamten Dokuments. Dabei müssen neben Satzstrukturen auch Zusammenhänge zwischen verschiedenen Aussagen oder Sätzen sowie einzelnen Kapiteln berücksichtigt und erkannt werden. Betrachtet man eine Kundenbewertung die mehr positiv als negativ ausgerichtete Sätze beinhaltet muss diese nicht unbedingt positiv sein. So kann die Bewertung mit dem Satz beginnen: 'Ich finde das Hotel extrem schlecht.', diese Aussage lässt den Schluss zu dass die Bewertung negativ ausfällt. Werden an diesen Satz eine Reihen von positiven Sätzen angehängt, wie 'Obwohl die Höflichkeit des Personals fabelhaft war.', ändert sich die Orientierung der Bewertung nicht. Diese sehr komplexen Strukturen müssen erkannt und richtig gedeutet werden.

Aufgrund der Komplexität der Klassifikation auf der Dokument-Ebene wird in dieser Arbeit die Aspekt-Ebene und Satz-Ebene für eine Hotelbewertung angenommen.

2.4 Filtern der Informationen

Kiritchenko et al., 2014 und Pang and Lee, 2008 weisen auf eine weitere grundlegende Aktion hin die bei der Klassifikation von Texten durchzuführen ist. Das Filtern von relevanten Informationen aus relevanten Texten. Dabei ist es nötig für die Bildung der richtigen Cluster die Entitäten, die Komponenten und die Teilkomponenten sowie alle Synonyme oder artverwandten Begriffe zu definieren und sinnvoll zusammenzufassen. Beispielsweise gehören zum Personal, unter anderem die Begriffe Kellner, -in, Rezeptionist, -in, etc. Daneben ist auch die Beachtung der Wörter die eine falsche Rechtschreibung aufweisen von Bedeutung. Auch Abkürzungen, Spitznamen oder Neologismen sind nicht zu vernachlässigen. Desweiteren müssen bewusste Änderungen der Rechtschreibung, z.B. 'groooooooßartige Aussicht' richtig erkannt und als relevant eingestuft werden.

Aus Gründen der Einfachheit wird in dieser Arbeit bei späteren Modellen die Annahme getroffen dass bei einer relevanten Kundenbewertung keine Rechtschreibfehler oder bewussten Änderungen der Rechtschreibung, keine Neologismen o.ä. vorliegt.

2.5 Domain

Die Domain bildet den thematischen Rahmen der Aussage, das heißt ob man sich beispielsweise im Themengebiet der Film-, Buch-, Hotelbewertungen oder im Bereich der Beurteilung einer politischen Rede etc. befindet. Das Vokabular, sowie die Bedeutung einzelner Phrasen oder Redewendungen kann je nach Domain variieren, die Orientierung ist somit abhängig von der Domain, Pang and Lee, 2008. Der Ausdruck 'Da gehe ich lieber Zelten.' ist bei der Bewertung eines Campingplatzes positiv zu werten, im Bereich der Hotelbewertung als negativ.
Nicht jede Domain muss zwangsläufig ein fachspezifisches Vokabular aufweisen, artverwandte Domains können durchaus die selben Begriffe mit der selben Orientierung beinhalten. 'Das Buffet war gut.', ist im Bereich der Campingplatz-, sowie der Hotelbewertungen als positiv zu werten. Wohingegen zu erwähnen ist, dass auch bei artverwandten Domains der selbe Ausdruck zu unterschiedlichen Ergebnissen kommen kann, beispielsweise der Ausdruck, 'Der Geruch war schon in der Lobby zu bemerken.' kann für ein Buffet positiv sein, für die Sauberkeit der Hotelzimmer kann das eine negative Bewertung widerspiegeln.

Kapitel 3

Die Polarität einer Aussage

Die Orientierung, eines Wortes wird durch die Richtung der Abweichung von der Norm bestimmt. Hierdurch wird die Verwendung und die Bewertung eines Wortes beschränkt. Hatzivassiloglou and McKeown, 1997 Die Polarität st das Gefühl das mit der Wotwahl verbunden ist, Kiritchenko et al., 2014. Hierbei ist auf eine Reihe von sprachwissenschaftlichen Punkten einzugehen. Im Mittelpunkt stehen Adjektive, die als Indikator angesehen werden.

3.1 Die Meinung

Die Meinung wird im Duden als, 'eine persönliche Ansicht, Überzeugung, Einstellung o. Ä., die jmd. in Bezug auf jmdn., etwas hat (und die sein Urteil bestimmt.)' definiert, Wolfgruber, 2015. Daraus lässt sich ableiten, dass die Meinung auf die individuell gewonnenen Eindrücke einer Person, beispielsweise während eines Hotelaufenthalts zurückzuführen sind. Die Meinung einer Person, bezieht sich auf ein Objekt. Diese Objekte können Dienstleistungen, Produkte o.Ä. sein. Beispielsweise die Sauberkeit eines Hotels, oder die Qualität eines kürzlich gekauften Laptops.

Wolfgruber, 2015 definiert unterschiedliche Kategorien der Meinungsäußerung.
(a) Direkt und subjektiv, dabei wird zu einem Objekt eine explizite Meinung geäußert, welche positiv oder negativ sein kann. Subjektive Sätze enthalten ein Adjektiv welches sich auf das zu bewertende Objekt bezieht. Z.B. 'Ich finde die Auswahl der Speisen des Frühstücksbuffet qualitativ hochwertig.'. Hierbei ist das Objekt, oder die Entität das 'Frühstücksbuffet' und 'hochwertig' das Adjektiv. Die Richtung des Satzes ist positiv. Jedoch sind nicht alle Adjektive als Indikatoren für die Orientierung eines Satzes einzustufen.

Dies istvom Kontext und dem Zielobjekt abhängig.
Liu and Zhang, 2012 definieren diese zwei Arten von Sätzen folgendermaßen. Durch subjektive Sätze werden Emotionen, Wünsche, Spekulationen, Ängste, Befürchungen o.ä. zum Ausdruck gebracht. Objektive Sätze beinhalten nur Informationen und lassen zumeist keine Rückschlüsse auf die Polarität zu. Objektive Sätze sind durch die Abwesenheit von Adjektiven und das vermehrte Auftreten von Nomen zu erkennen. 'Ich habe heute morgen am Frühstücksbuffet teilgenommen.', ist ein objektiver Satz und kann als neutral angesehen werden.
Weiter definiert Wolfgruber, 2015 die Kategorie (b) Implizit, dabei wird über eine rhetorische Frage oder einem Modalverb der Meinung Ausdruck verliehen. 'Wäre ein heisser Kaffee morgens nicht etwas schönes?', oder 'Der Umgangston mit Gästen sollte grundsätzlich verändert werden.'. Diese beiden Beispiele geben einen Hinweis darauf, dass der Gast mit dem Kaffee, bzw. dessen Temperatur nicht zufrieden war. Im zweiten Beispiel war der Gast mit dem Umgangston des Hotelpersonals nicht zufrieden. Beispielsweise bei der Beschwerde über den nicht zufriedenstellenden Kaffee, könnte die Antwort oder die Reaktion des Kellners den Unmut des Gastes provoziert haben.
(c) Vergleichend, dabei verleiht der Meinungsträger seiner Meinung über einen Vergleich Ausdruck. Liu and Zhang, 2012 und Wolfgruber, 2015 arbeiten heraus, dass ein Vergleich neben dem Komparativ und Superlativ, auch über andere Verben wie 'vorziehen' oder 'übertreffen', also deren Infinitiv, bestehen kann. Beispiele hierfür sind, 'Hotel 1 ist besser als Hotel 2.' 'Hotel 1 ist das beste.' 'Der Service in Hotel 1 hat den Service in Hotel 2 übertroffen.' In allen drei Beispielen bevorzugt der Meinungsträger das Hotel 1.
Desweiteren können nach Liu and Zhang, 2012 Vergleiche in vier Haupttypen eingeteilt werden.
(1) Vergleiche die eine Hierarchie zulassen, das heißt man kann die angesprochenen Entitäten in Bezug auf eine Komponente ordnen. 'Die Zimmer in Hotel 1 sind sauberer als in Hotel 2.' Hieraus lassen sich Präferenzen ableiten.
Wiederum lässt sich diese Art von Vergleich in Untergruppen teilen. Zum einen in die Gruppe, in der eine Steigerung angesprochen wird. Beispielsweise durch die Verben 'übertreffen' oder die Adjektive 'länger', 'weiter', 'höher'. Zum anderen in die Gruppe der absteigenden Hierarchie, durch Verben wie 'unterbieten' oder Adjektive wie 'kürzer', 'kleiner', 'dunkler'.
(2) Vergleiche die Gleichheit ausdrücken. Es werden gleiche Komponenten unterschiedlicher Entitäten verglichen. 'Die Sauberkeit der Zimmer in Hotel 1 unterscheidet sich nicht von der Sauberkeit der Zimmer in Hotel 2.'
(3) Vergleiche der Superlative. Eine Entität wird besser als alle anderen be-

wertet. 'Hotel 1 ist das beste.'
(4) Nicht hierarchische Vergleiche. Hierbei können die verglichenen Entitäten, oder Komponenten nicht geordnet werden und somit keine Aussage über die Haltung des Autors getroffen werden. Es lassen sich drei Gruppen bilden.
(4.1) Zwei Ähnliche oder auch unterschiedliche Entitäten werden hinsichtlich gleicher Komponenten verglichen. 'Der Service in Hotel 1 ist anders als der Service in Hotel 2.'
(4.2) Entität 1 hat eine Komponente a_1 und Entität 2 besitzt eine Komponente a_2, wobei die Komponenten a_1 und a_2 Substitute definieren. 'Hotel 1 bietet ein à la carte Frühstück und Hotel 2 ein Frühstücksbuffet an.'
(4.3) Entität 1 besitzt eine Komponente die Entität 2 nicht aufweist. 'In Hotel 1 gibt es einen Indoor Golfplatz, Hotel 2 hat so etwas nicht.'.
Bei dieser Art von Vergleich ist die Gesinnung des Autors nur durch den Kontext zu erkennen.

Satzeichen spielen ebenfalls eine Rolle und können die Aussage eines Satzes ändern. Der Satz, „Abitur ist einfach anders als ein Studium.“ kann als objektiver Satz betrachtet werden. Hingegen durch das einfügen eines Kommas wird die Aussage des Satzes geändert. „Abitur ist einfach, anders als ein Studium.“. Hierdurch wird ein subjektiver Satz generiert und eine Aussage über das Empfinden bezogen auf das Abitur im Vergleich zu einem Studium getroffen.

3.2 Die Polarität

Die Orientierung einer Meinung lässt sich grundsätzlich in drei Grundrichtungen einteilen, negativ, neutral und positiv. Diese Grundrichtung kann dann wiederum in feinere Klassen unterteilt werden, wie z.B. Bewertungskategorien in Onlineportalen.
Die Richtung wird durch bestimmte Wörter innerhalb der Aussage bestimmt. Beispiele für solche Wörter sind nicht ausschließlich Adjektive wie 'gut' oder 'schön'. Auch Nomen wie 'Urlaub', erwecken eine positive Emotion. Wörter wie 'Hausaufgaben' geben eine negative Polarität an. Die neutrale Polarität wird zumeist objektiven Sätzen zugeordnet.
Die Polarität einer Aussage wird durch Sprachwissenschaftliche Effekte beeinflusst, beispielsweise durch die Verbindung eines Adjektives und dem negierendem Wort 'nicht'. Liu and Zhang, 2012 stellen fest, dass die Polarität der Aussage hierdurch umgekehrt werden kann. Betrachtet man den Satz, 'Das Wetter ist schön.', liegt eine positive Polarität vor. Fügt man zusätzlich ein negierendes Wort ein, wie 'Das Wetter ist nicht schön.', wird die Polarität

des Satzes umgekehrt und ist negativ.

Liu and Zhang, 2012 betrachten diesen Punkt detaillierter und schlagen sechs Regeln vor, durch die eine Aussage hinsichtlich deren Polarität bewertet werden kann. Diese Regeln können jedoch nicht verallgemeinert werden und unterscheiden sich je nach Kontext.
(a) Polaritätsanzeigende Wörter oder Phrasen sind Wörter wie Adjektive.
(b) Es wird ein erwünschter oder unerwünschter Umstand angesprochen. 'Die Sitze im Bus waren hart.', hier wird ein unerwünschter Umstand angesprochen und ist daher als negativ zu werten.
(c) Eine hohe oder geringe Menge an potentiell negativen, oder positiven Items. 'Lange Akkulaufzeit.' ist als positiv zu werten. Die Akkulaufzeit ist ein potentiell positives Item (PPI). Hingegen der Ausdruck 'Die Kosten für das Handy sind hoch.' ist negativ behaftet. Das Wort Kosten ist ein potentiell negatives Item (PNI). Die Kombination solcher Items mit einem Ressourcen verzehrendem Wort oder einem Ressourcen gewinnendem Wort ist von Situation zu Situation unterschiedlich. Wörter wie 'hoch' oder 'viel' in Kombination mit einem PNI ergeben einen negativen Ausdruck. Die Kombination mit einem PPI, hat einen positiven Ausdruck zur Folge.
(d) Eine ansteigende oder sinkende Menge in Kombination mit einem Wort welches eine bestimmte Polarität anspricht. Der Schmerz ist ein negatives Wort, in Verbindung mit 'verringern' ergibt sich ein positiver Ausdruck. Die Verbindung des Wortes Schmerz mit 'erhöhen' ergibt einen negativen Ausdruck.
(e) Abweichung von einem bestimmten Spektrum, wie z.B. die Erhöhung des Blutdrucks ist als negativ zu betrachten, wenn diese Erhöhung des Blutdrucks eine bestimmte Grenze übersteigt. Liegt man außerhalb eines bestimmten Bereichs, welcher als positiv eingestuft wird ist dieser Sachverhalt als negativ zu betrachten.
(f) Generieren oder verzehren von Ressourcen und Abfallprodukten. Verbraucht ein Auto viel Benzin oder generiert viel Abgas ist dies negativ. Generiert eine Solaranlage viel Strom, ist dies positiv. Wird hingegen wenig Strom erzeugt, so ist dies negativ

Wolfgruber, 2015 hingegen teilt die Polarität in drei unterschiedliche Ebenen ein.
(1) Die lexikalische Polarität. Hierbei werden Synonyme der selben Polarität zugeordnet und Antonyme der gegenteiligen. Ein Beispiel hierfür ist das Adjektiv 'schön' mit einer positiven Orientierung. Ein mögliches Synonym ist 'entzückend' und demnach ebenfalls positiv. Das Antonym 'hässlich' besitzt eine negative Orientierung.

(2) Die kontextbedingte Polarität, in Verbindung mit einer Negation. Durch eine Negation in Bezug auf das Polaritätsgebende Wort wird eine Umkehrung der Orientierung bewirkt.
(3) Die kontextbedingte Polarität in Verbindung mit Ironie oder Sarkasmus. Je nach Kontext ergibt sich für die Aussage eine andere Orientierung. 'Dank der tollen Bedienung ist der Abend ein voller Erfolg geworden.' Das Wort toll besitzt hier eine positive Orientierung, auf den Kontext bezogen ergibt sich für die gesamte Aussage eine positive Orientierung. Hinegen im Kontext des folgenden Satzes ist die Orientierung negativ. 'Dank der tollen Bedienung ist jede warme Speise kalt am Tisch serviert worden.'

Wolfgruber, 2015 erklärt das die Polarität ebenfalls vom Satzbau abhängt. Die Kombination, 'Nicht nur ich mochte das Essen.' oder die Aussage, 'Nur ich mochte das Essen nicht.' besitzen je nach Blickwinkel eine unterschiedliche Richtung. Ebenfalls ist zu sehen das daneben auch die Entfernung zwischen zwei Wörtern die Orientierung einer Aussage beeinflusst.
Redewendungen besitzen ebenfalls eine Orientierung, 'Das Essen entsprach der Kategorie unterste Schublade.' impliziert eine negative Polarität. Hingegen deutet der Satz, 'Man hatte das Gefühl man speist wie Gott in Frankreich.' auf eine positive Haltung gegenüber dem Buffet hin.
Zur Vervollständigung der Auflistung möglicher Indikatoren seien die Hashtags bei Tweets sowie die Emoticons erwähnt. Ein lachender Smiley bedeutet eine positive Haltung des Autors, wohingegen ein grimmiger Smiley eine negative Haltung des Autors impliziert, Kiritchenko et al., 2014 und Wolfgruber, 2015.
Aus Gründen der Einfachheit wird für spätere Analysen angenommen, dass weder Hashtags noch Emoticons in Hotelbewertungen zu beobachten sind. Desweiteren wird der Fall der Ironie und Sarkasmus ausgeschlossen.

3.3 Intensität der Polarität

Liu and Zhang, 2012 erwähnen, dass sich Emotionen in sechs Hauptgruppen einteilen lassen. Beispielsweise in Liebe, Freude, Verwunderung, Zorn, Traurigkeit und Angst. Jede Emotion kann zusätzlich in hierarchisch angeordnete Untergruppen unterteilt werden.
Wie Angst, Fürchten oder Panik u.s.w. sind unterschiedliche Ebenen einer ähnlichen Emotion. Diese Emotionen besitzen unterschiedliche Intensitäten, beispielsweise ist die Panik eine intensivere Form der Angst. Polanyi and Zaenen, 2004 sprechen in diesem Zusammenhang die Valenz an, diese repräsentiert die Stimmung die aus einem lexikalischen Begriff hervorgeht. Das heißt

das für den selben Umstand durch unterschiedliche Wortwahl verschiedene Emotionen geweckt werden können. Desweiteren definieren Polanyi and Zaenen, 2004 Wörter die für eine Verschiebung der Valenz verantwortlich sind. Diese ändern die Orientierung oder die Intensität der Aussage. Wolfgruber, 2015 spricht in diesem Fall von Partikeln. Beispielsweise besitzt die Aussage, 'Das Essen ist geniessbar.' eine positive Polarität. Wird der Partikel 'kaum' eingefügt, also 'Das Essen war kaum geniessbar.' kehrt sich die Polarität ins leicht negative um.
Durch das hinzufügen anderer Wörter kann sich nicht nur die Richtung ändern, sondern auch die Intensität der Aussage. Die beiden Sätze 'Das essen war äusserst geniessbar.', oder 'Das essen war sehr geniessbar', dienen als Beispiel.

Kapitel 4

Spam und Qualität der Beurteilungen

Aus der Arbeit von Liu and Zhang, 2012 geht hervor, dass Produktbewertungen unter Umständen einen großen Einfluss auf die Konsumenten besitzen. Es gibt einen bestimmten Personenkreis der ein großes Interesse daran haben könnte ein Produkt sehr gut darzustellen. Denkbar wäre das Hotelangestellte die Konkurrenz schwächen wollen, oder sich über sehr positive Bewertungen eine größere Anzahl an Gästen erschleichen möchten. Dieser Personenkreis schreibt Pseudo-Beurteilungen, den sogenannten Spam, um ihr Ziel zu erreichen.
Im folgenenden wird basierend auf der Arbeit von Liu and Zhang, 2012 der Umgang mit Spam kurz dargestellt. Als Ausgagngspunkt für einen tieferen Einblick sei auf die Arbeit von Liu and Zhang, 2012 verwiesen.

4.1 Erkennung von Spam Beurteilungen

Arten von Spamverfassern

Zum einen gibt es die individuellen Verfasser von Spam. Diese melden sich mit einem, oder mehreren Accounts bei einer, oder mehreren Plattformen an. Mithilfe der / des Accounts wird der Spam bezogen auf ein Produkt, oder mehrere Produkte verbreitet.
Zum anderen gibt es die Mitglieder einer Gruppe die gemeinschaftlich ein Ziel verfolgen, z.B. möchten diese ein Konkurrenzprodukt mit falschen Bewertungen schwächen, oder ihr eigenes Produkt besser darstellen. Das Vorgehen der Gruppe ist dem Handeln einer Einzelperson ähnlich. Diese unterscheiden sich lediglich durch die Gruppendynamik.

Methoden zur Erkennung von Spam

Spam ist in vielen Fällen schwer zu erkennen, ein möglicher Grund ist, das sich der Verfasser von Spam Bewertungen an anderen bereits bestehenden Beurteilungen orientiert und den Schreibstil kopiert, um durch diese Tarnung nicht aufzufallen.

Die Erkennung von Spam ist prinzipiell ein Problem der Klassifikation, in die Klassen 'Spam' oder 'Kein Spam'. Trainingsdaten zu generieren ist schwer da es aus beispielsweise dem oben genannten Grund nicht leicht ist Pseudo Beurteilungen von ernsthaften Kundenbewertungen zu unterscheiden. Eine Methode Trainingsdaten zu erstellen ist es, sogenannte Duplikate zu suchen. Dabei werden Kritiken gesucht die bereits identifizierten Spam Bewertungen möglichst ähnlich oder diesen gleich sind. Diese Kritiken werden dann ebenfalls als Spam klassifiziert. Die restlichen Bewertungen werden der Klasse 'Kein Spam' zugeteilt.
Hierzu werden die Trainingsdaten in drei Gruppen mit bestimmten Eigenschaften eingeteilt.
(a) Beurteilungsbezogen, das bedeutet die Inhalte der bestehenden Beurteilungen werden berücksichtigt. Unter anderem wird die Wortwahl, die Länge des Textes oder die Häufigkeit mit der ein bestimmter Markenname in der Beurteilung erwähnt wird betrachtet.
(b) Autorenbezogen, dabei werden Informationen die über den Autor vorliegen verwendet. Beispielsweise die durchschnittlichen Bewertungen oder die Standardabweichung der abgegebenen Bewertungen betrachtet. Weicht eine Bewertung stark von der 'Norm' ab, so ist dies ein Anzeichen für möglichen Spam. Desweiteren werden z.B. Relationen gebildet, wie die Anzahl der Bewertungen zu einem bestimmten Produkt im Verhältnis zur Gesamtzahl der Bewertungen die der betreffende Autor abgegeben hat.
(c) Produktbezogen, Daten wie die durchschnittlichen Bewertungen die zu dem bestimmten Produkt abgegeben wurden. Weicht eine Bewertung zu stark von den mittleren Bewertungen ab, so wird diese als Spam klassifiziert.

Zur Modellierung werden logistsche Regressionsmodelle verwendet.

Automatische Methoden bergen das Risiko, dass eine falsche Einteilung in die zwei Klassen vergenommen werden kann. Das heißt eine ernst gemeinte Kritik kann fälschlicherweise als Spam klassifiziert werden, und umgekehrt. Bewertet eine Person ein Produkt sehr schlecht da diese mit dem Produkt sehr unzufrieden ist, obwohl von dieser sonst immer sehr gute Bewertun-

gen abgegeben werden, wird die Bewertung fälschlicherweise als potentieller Spam eingeteilt. Eine weitere Gefahr ist, dass ein Duplikat fälschlicherweise als Spam markiert wird, obwohl rein zufällig eine starke Ähnlichkeit zu einer als Spam gekennzeichneten Kritik besteht. Auch kann ein nicht Duplikat eine Spam Bewertung sein, da der Autor diese nur getarnt hat.

Eine weitere Methode ist die Erkennung von Spam basierend auf der Analyse des Verhaltens der Gesamtheit der Kritiken. Sticht eine Bewertung aus der breiten Maße heraus kann das ein Zeichen für Spam sein. Beispielsweise bewertet die Allgemeinheit ein Produkt gut, jedoch bewertet eine Person dieses Produkt extrem schlecht kann das auf Spam hindeuten. Hierzu existieren zwei Methoden.
Bei der ersten Methode wird jedem Autor ein Score zugeteilt, der das Spam Potential wiedergibt.
Bei der zweiten Methode werden unerwartete Bewertungen als Spam deklariert. Hierzu dienen vier Definitionen als Basis. (a) Konfidenzerwartungen, ist die Allgemeinheit mit einem bestimmten Produkt sehr zufrieden und nur einige wenige sehr unzufrieden kann dies auf Spam hindeuten. (b) Förderung des Produkts, eine Person schreibt mehr als nur eine Bewertung zu einem Produkt, während andere nur eine Bewertung verfassen. (c) Die Verteilung der Zuordnung, hierbei stammen viele positive Bewertungen von einem Verfasser, obwohl die restlichen Bewertungen anderer Verfasser meist negativ sind. (d) Zuordnung, von einem Verfasser werden die Produkte eines Herstellers äußerst positiv und Produkte anderer Marken sehr negativ bewertet.

Erkennung von Gruppen basiertem Spam

Um ein Produkt oder eine Dienstleistung zu unterstützen oder zu schwächen arbeiten einige Verfasser von Bewertungen in Form einer Gruppe zusammen. Zur Identifikation von Gruppen basiertem Spam wird die folgende Methode verwendet.
Die Zusammenfassung aller Kundenbewertungen eines Produkts oder einer Dienstleistung wird nach Auffälligkeiten untersucht. So werden Bewertungen die ein ähnliches Muster aufweisen zusammengefasst und die Verfasser der Bewertungen als eine potentielle Gruppe betrachtet. Potentielle Gruppen bewerten die gleichen Produkte immer in sehr ähnlicher Form.
In einem zweiten Schritt wird die potentielle Gruppe hinsichtlich deren verhalten untersucht, dazu werden verschiedene Indikatoren herangezogen,
diese sollen gruppendynamsiches Verhalten aufdecken. Beispielsweise wird die Zeitspanne in der die Beurteilungen verfasst wurden betrachtet, je klei-

ner die Zeitspanne in der Personen die Bewertungen 'gleichzeitig' abgegebenen haben, desto höher ist das Potenzial das eine Gruppe existiert. Ein weiterer Indikator ist z.B. das Verfassen von Nachrichten kurz nach einer Produktveröffentlichung, in der eventuell das Produkt gar nicht genügend genutzt werden konnte um sich eine ernsthafte Meinung bilden zu können. Dieser Umstand kann auch zur Erkennung von Spam einer einzelnen Person genutzt werden.
Hierzu können die von Liu and Zhang, 2012 erwähnten Quintupel herangezogen werden.

4.2 Qualität einer Beurteilung

Neben der Erkennung von Spam, spielt die Einschätzung der Qualität einer Bewertung eine Rolle. Liu and Zhang, 2012 arbeiten den Punkt heraus, dass es für die Angabe einer Rangordnung der Bewertungen notwendig ist die Nützlichkeit einer Beurteilung festzulegen. Dadurch erscheinen die als am besten erachteten Beurteilungen zuerst. Um dies zu ermöglichen bieten Plattformen wie `https://www.amazon.de` eine Schaltfläche am Ende einer Kritik an, mit welcher der Leser die Beurteilung hinsichtlich deren Nützlichkeit bewerten kann. Weiter wird bei jeder Beurteilung angegeben wie viele Personen diese als Nützlich erachten. Hierdurch wird ein Bewertungsschema geschaffen, das die einzelnen Kritiken hinsichtlich deren Qualität einstuft.

In späteren Modellen wird die Annahme getroffen, dass kein Spam vorliegt. Weiter werden keine Unterschiede in der Qualität zwischen einzelnen Kritiken gemacht.

Kapitel 5

Sentiment Lexika

Frühe Methoden der Klassifikation von Texten basieren auf der Anwendung von sogenannten Sentiment Lexika. Diese Sentiment Lexika werden im folgenden beschrieben.
Ein Sentiment Lexikon ist die Zusammenfassung von Wörtern, Phrasen und deren Polarität. Zusätzlich kann der Sentiment Score, welcher die Intensität einzelner Einträge angibt angegeben werden. Auch die Angabe der Wortart oder der Part-of-Speech Tags ist möglich. Part-of-Speech Tags sind Kürzel die Wörter genauer beschreiben,vgl. Schiller et al., 1999. Auf diese wird im Rahmen dieser Arbeit nicht näher eingegangen
Sentiment Lexika können in zwei Hauptgruppen unterteilt werden.
(1) Lexikon basierte Ansätze, diese lassen sich in die Untergruppen (a) Manuelle, (b) Wörterbuch basierte und (c) Korpus basierte Ansätze unterteilen.
(2) Die Gruppe der maschinellen Lernverfahren in Ansätze aus dem (d) unüberwachtem und (e) überwachtem Lernen.

5.1 Manueller Ansatz

Bei einem manuellen Ansatz wird die Zusammenfassung einzelner Wörter und Phrasen per Hand erstellt und die entsprechende Polarität der Einträge vom Ersteller selbst oder ausgewählten Personen annotiert. Dieses Verfahren ist sehr zeitaufwendig und kostenintensiv. Neben dem manuellen extrahieren und dem eintragen der Begriffe in Listen müssen die einzelnen Polaritäten festgestellt werden. Dieses Verfahren hängt sehr vom Ersteller ab, da die Klassifizierung der Aussagen je nach Person und Anzahl der Klassen variieren kann. Je mehr Klassen zur Einteilung wählbar sind desto größer werden, je nach Ersteller, die Unterschiede in der Einteilung einzelner identischer Begriffe. Diese Form der Erstellung eines Lexikons bietet sich jedoch an wenn

ein bereits bestehendes Lexikon verbessert oder Domainspezifisch modifiziert werden soll.
Als ein Vertreter dieser Art sei der *General Inquirer* (GI) erwähnt. Dieses englisch sprachige Lexikon wurde von Philip J. Stone in den 1960er Jahren manuell entwickelt, Wilson et al., 2005 und Züll and Mohler, 1989. Der GI sollte zur computergestützten Inhaltsanalyse in verschiedenen Wissenschaften wie der Psychologie dienen. Heutzutage ist der GI ein in der Sentiment Analyse weit verbreitetes Hilfsmittel. Unter folgendem Link ist die Homepage des GI zu erreichen. `http://www.wjh.harvard.edu/~inquirer/`. Im Rahmen dieser Arbeit wird nur kurz auf den GI verwiesen da dieser für den englischen Sprachgebrauch entwickelt wurde und in weiteren Analysen nicht verwendet wird.

5.2 Wöterbuch-basierter Ansatz

Zur Erstellung eines Wörterbuch-basierten Lexikons werden Wörterbücher oder andere Hilfsmittel wie z.B. der online Duden verwendet. Dabei wird zuerst eine Referenzliste aus positv und negativ polarisierten Wörtern erstellt, um danach Antonyme und Synonyme zu identifizieren. Diese werden in das Lexikon aufgenommen, Wolfgruber, 2015.

Sentiment Wortschatz

Der Sentiment Wortschatz (SentiWS) ist ein öffentlich zugängliches deutschsprachiges Sentiment Lexikon der Universität Leipzig. Dieses Lexikon wurde ursprünglich generiert um die Effekte von Zeitungsartikeln in Finanzzeitungen und DAX betreffende Blog Einträge zu untersuchen, Remus et al., 2010. SentiWS beinhaltet deutsche Wörter und deren Polarität. Daneben werden die Part-of-Speech Tags (POS) angegeben, sowie die Flexion und der Sentiment Score, falls diese vorhanden sind. Zusätzlich wird die semantische Orientierung (SO) berechnet um die unterschiedlichen Intensitäten zu berücksichtigen. Die semantische Orientierung wird Normalisiert (dem Gewicht) und ist in einem Intervall von $[-1; 1]$ ausgewiesen. Auf die semantische Orientierung wird im späteren Verlauf der Arbeit nocheinmal eingegangen. Im folgenden wird dieses Lexikon kurz erläutert.

Das Lexikon besteht nicht nur aus polaritätsgebenden Adjektiven, und deren Flexion, sondern auch aus Adverben, Nomen und Verben. Aus diesen vier Wortarten ergibt sich eine Gesamtzahl von 1650 positiven und 1818 negativen

Wörtern, unter Berücksichtigung der Flexionen beinhaltet das Lexikon 15649 positive und 15632 negative Wörter.

Wort	POS	Gewicht	Flexion			
Genuss	NN	0.0701	Genüssen	Genuss	Genüsse	Genusses
Genuß	NN	0.073	Genußes	Genüßen	Genüße	Genußs

Tabelle 5.1: Beispielhafte Einträge SentiWS

Die Tabelle 5.1 gibt beispielhaft zwei Einträge des Wortes Genuss (Genuß) an, entsprechend der neuen und der alten Rechtschreibung. Es ist erkennbar dass sich durch die Änderung der Schreibweise das Gewicht verändert, jedoch ist dies bei z.B. einer Rundung auf zwei Dezimalstellen nach dem Komma vernachlässigbar. Weiter wird nicht zwischen der maskulinen und femininen Form unterschieden, beispielsweise wird für das Wort Gewinner nicht zusätzlich Gewinnerin aufgeführt.

Zur Erstellung des Lexikons wurde von Remus et al., 2010 (a) der GI, (b) das gemeinsame Auftreten von verschiedenen Wörtern und (c) das *German Collocation Dictionary* verwendet.
Die als positiv und negativ gekennzeichneten Wörter aus dem GI wurden semi automatisch mit Google Translate übersetzt. Die Übersetzungen wurden mit der des GI zugewiesenen Polarität in das Lexikon aufgenommen. Im Anschluss daran bearbeiteten die Autoren die Liste manuell um Wörter ohne, oder nicht passender, Polarität zu entfernen. Einige ausgewählte Begriffe aus der Domäne des Finanzmarktes wurden ebenfalls manuell von den Erstellern des SentiWS hinzugefügt.
Als zweites wurde das gemeinsame Auftreten einiger Wörter beobachtet. Dazu wurden Pseudo Wörter, sog. Marker, in positiv oder negativ annotierte Bewertungen eingefügt. Remus et al., 2010 beobachten dadurch welche Wörter signifikant oft mit den Markern auftreten, und gewinnen daraus neue Wörter sowie deren Polarität. Diese neuen Informationen werden ebenfalls in das Lexikon aufgenommen. Ein Vorteil dieser Methode ist, dass eine Domänenspezifische Terminologie generiert werden kann. Da je nach Domäne bestimmte Begriffe besonders oft gemeinsam Auftreten, z.B. im Bereich von Hotels das Wort Frühstück und Buffet, das Frühstücksbuffet.
Als dritten Schritt wurde das *German Collocation Dictionary* zur Erstellung des Lexikons herangezogen. Dieses Lexikon gruppiert nach Wörtern welche mit gewissen Nomen häufig gemeinsam auftreten, die eine ähnliche Semantik aufweisen. Hierdurch konnten ebenfalls neue Wörter identifiziert werden, die Kombination 'sonnendurchflutet' wird von Remus et al., 2010 als Beispiel

aufgeführt.

Zur Berechnung der Gewichtung wird die Semantische Orientierung (SO) herangezogen , vgl. Turney, 2002a und Turney, 2002b. Diese wird auf das Intervall $[-1; 1]$ normalisiert. Eine mögliche Formel zur Normalisierung ist:

$$2 \times \frac{SO - min(SO)}{max(SO) - min(SO)} - 1. \tag{5.1}$$

Hierdurch ist eine Vergleichbarkeit der Intensitäten einzelner Aussagen möglich.

Bing Liu's Lexikon

Hu and Liu, 2004 erstellen ein Wörterbuch-basiertes Lexikon zur Klassifikation von Kundenbewertungen bezüglich Produktkomponenten. Die Ergebnisse werden in zusammengefasster Form für die jeweiligen Komponenten dargestellt.
Dazu werden aus den subjektiven Sätzen einer Kundenbewertung die polaritätsgebenden Adjektive extrahiert. Diese Adjektive repräsentieren erwünschte oder unerwünschte Zustände, wie schön oder hässlich. Um die unbekannte Polarität der extrahierten Adjektive zu bestimmen wird z.B. *WordNet* herangezogen. *WordNet* ist ein frei zugängliches Wörterbuch der Universität Princeton, vgl. Fellbaum, 1998.

Zur Bestimmung der Orientierung der Adjektive werden Wortfelder definiert. Die Wortfelder werden in zwei Bereiche geteilt. Zum einen in das zu polarisierende Adjektiv und dessen Synonyme, zum anderen in das entsprechende Antonym und dessen Synonyme. Zusätzlich werden nicht nur Synonyme in das Wortfeld aufgenommen, sondern auch artverwandte, oder ähnliche Begriffe.
Mit den zentralen Annahmen, dass (a) Synonyme (und artverwandte oder ähnliche Begriffe) eine gleiche Polarität aufweisen und (b) Antonyme eine gegensätzliche Polarität besitzen, lässt sich die Orientierung der Zieladjektive bestimmen. Bei einer genügend großen Informationsbasis, wie die Verwendung mehrerer Wörterbücher und andere Hilfsmittel wie *WordNet* ist es möglich die Polarität aller Adjektive zu bestimmen.
Zur Erstellung des dafür nötigen Lexikons wird ein Algorithmus verwendet der die Polarität des gesuchten Wortes bestimmt und die klassifizierten Wörter anschließend der Liste der Referenzwörter hinzufügt, woraus am Ende das Lexikon entsteht.
Zuerst wird eine Liste von Referenzwörtern mit den zugehörigen positiven

und negativen Polaritäten manuell erstellt, ein Beispiel hierfür kann in Tabelle 6.1 eingesehen werden. Die Orientierung der Referenzwörter kann durch das Heranziehen von bereits bestehenden Lexika festgestellt werden. Auch die Klassifikation der Referenzwörter über subjektives annotieren der Polarität ist möglich. Hierbei finden einfache Wörter wie 'gut' oder 'schlecht', deren Polarität allseits anerkannt ist, Verwendung.
Im zweiten Schritt werden die Referenzliste und die Liste der zu klassifizierenden Adjektive miteinander verglichen. Ist ein Wort nicht in der Referenzliste enthalten so wird dessen Polarität entsprechend dem obigen Verfahren bestimmt und danach der Ausgangsliste zugeführt. Ist das Wort enthalten wird das nächste Wort ausgewählt. Dies wird solange wiederholt bis keine neuen Wörter in die Referenzliste aufgenommen werden müssen.

Dieser im Vergleich zu manuellen Methoden einfache aber effiziente, kostengünstigere und weniger zeitaufwendige Ansatz bietet die Möglichkeit durch die Extraktion von anderen Wortarten wie Nomen oder Verben und Redewendungen das Lexikon nicht nur für Adjektive zu formulieren. Dies kann daher sehr flexibel gestaltet werden und neben der Domäne auch regionale Dialekte aufnehmen.
Bei Verwendung von mehreren Hilfsmitteln kann die Datenbasis erweitert werden, wodurch sich die Notwendigkeit des subjektiven Eingreifen verringert. Zum anderen kann die Orientierung der einzelnen Treffer miteinander verglichen werden und die am häufigsten angegebene Polarität eines Wortes als Ergebnis angesehen und aufgenommen werden.

5.3 Korpus-basierter Ansatz

Bei dieser Art ein Lexikon zu erstellen wird eine Menge von Texten Zusammengefasst und aus dieser Stichprobe mittels z.B. Bootstrapping Methoden ein Lexikon generiert, Wolfgruber, 2015. Die Texte können in verschiedenen Arten vorliegen, z.B. Hotelbewertungen, oder allgemeine Produktbewertungen. Durchaus denkbar wären aber auch Twitter Nachrichten oder SMS. Kiritchenko et al., 2014 verwenden beispielsweise Twitternachrichten zur Erstellung eines Lexikons. Diese Methode bietet die Möglichkeit Domänespezifische Lexikas zu erstellen, da die Grundmenge an Texten und die verwendeten Begriffe eine starke thematische Ausrichtung aufweisen. Durch die Verwendung unterschiedlicher Domänespezifischer Texte, kann diese Methode ein interdisziplinäres Lexikon repräsentieren.

Als zwei Vertreter seien an dieser Stelle das Lexikon nach Kiritchenko et al., 2014 und Wilson et al., 2005 erwähnt. Das Vorgehen zur Erstellung der Lexikas wird im folgenden kurz beschrieben, für einen tieferen Einblick sei hier auf die jeweiligen Arbeiten verwiesen.

Lexikon nach Kiritchenko et al.

Kiritchenko et al., 2014 erstellen ein Lexikon dass die Kontextbedingte Änderung der Polarität und / oder die Kontextbedingte Änderung der Intensität berücksichtigt. Zwei Korpora werden hierzu verwendet, zum einen ein Korpus bestehend aus Twitter Nachrichten die von April bis Dezember 2012 gesammelt wurden. Zum anderen wird der *Sentiment 140* Korpus verwendet, dieser besteht aus 1,6 Millionen Tweets die Emoticons beinhalten. Die Tweets werden abhängig von den darin enthaltenen Emoticons als positiv oder negativ klassifiziert. Die beiden Korpora werden zusammengefasst und anschliessend für jedes Wort die SO berechnet. Dieser Wert dient später als Basiswert.
Die Tweets werden dann wiederum in zwei Korpora eingeteilt, erstens in ein Lexikon das negative und zweitens in ein Lexikon das den positiven Kontext berücksichtigt. Teile des Tweets die mit einer Negation beginnen und mit einem Satzzeichen enden werden als negativ angesehen und dem negativen Korpus zugeteilt. Die negierenden Wörter wurden dazu entsprechend einer Referenzliste definiert. Die restlichen Teile des Tweets werden als positiv bewertet und dem positiven Korpus beigefügt. Die Polaritäten werden in beiden Fällen beibehalten. Anschließend wird die SO für den positiven und negativen Korpus berechnet. Die jeweilige SO für die einzelnen Wörter werden zu einem Lexikon zusammengefasst.
Die folgende Tabelle zeigt einen Ausschnitt aus diesem Lexikon.

	Basis	Zustimmend	Ablehnend
Positives Wort great	1.177	1.273	-0,367
Negatives Wort terrible	-1.766	-1.850	-0.890

Tabelle 5.2: Auszug aus dem Lexikon nach Kiritchenko et al., 2014

Durch die Berücksichtigung des Kontexts sind präzisere Aussagen über die semantische Orientierung möglich, da sich je nach Kontext die Orientierung ändern und / oder intensivieren kann.
Ein Nachteil des Verfahrens ist die starke Abhängigkeit der SO vom verwendeten Korpus, sowie die große Menge an Daten die benötigt wird.

Das MPQA Lexikon

Wilson et al., 2005 dient als Basis der Multi-perspective Question Answering Opinon Corpus (MPQA). Der MPQA stellt eine Sammlung von englischen Zeitungsberichten dar. Diesen Berichten wird eine bestimmte subjektive Einschätzung wie z.B eine Emotion, oder eine Polarität zugeordnet. Nähere Informationen sind unter dem Link `http://mpqa.cs.pitt.edu` zu erfahren. Durch das manuelle Zuweisen werden die Aussagen entsprechend der Polarität klassifiziert.

Kapitel 6

Statistisches Lernen

6.1 Unüberwachte Verfahren

Die Semantische Orientierung (SO), nach Turney, 2002a und Turney, 2002b, stellt ein frühes Verfahren der Sentiment Analyse dar und dient dem generieren eines Scores der die Polarität und deren Intesität quantitativ angibt.

Die Berechnung findet in drei Schritten statt. Zuerst werden aus einem Text Kombinationen extrahiert die beispielsweise aus zwei Wörtern bestehen. Diese Kombinationen enthalten ein Adjektiv oder Adverb, welches mit einem anderen Wort (Adjektiv, Adverb, Nomen oder Verb) kombiniert wird. Um zusätzlich die kontextbedingte Polarität zu berücksichtigen kann ebenfalls ein weiteres Wort extrahiert werden, welches Aufschluss über die Domain bzw. dem Kontext gibt.
Im zweiten Schritt wird die SO der Kombinationen jeweils einzeln berechnet. Hierzu wurden von den Autoren bestimmte Referenzwörter für eine positive und negative Polarität gewählt. Die Richtung der einzelnen Wörter und Wortkombinationen wird durch, z.B. die lexikalische Polarität festgelegt, wobei auch Polaritäten aus den oben erwähnten Lexika denkbar wären. Die folgende Tabelle 6.1 gibt die Referenzwörter die von Turney, 2002a verwendet werden an.[1].

[1]Die englischen Begiffe sind Sammelbegriffe und die deutsche Übersetzung nur einer von vielen möglichen Begriffen, andere Übersetzungen sind ebenfalls möglich.

Positive	Negative
good	bad
nice	nasty
excellent	poor
positive	negative
correct	wrong
fortunate	unfortunate
superior	inferior

Tabelle 6.1: Referenzwörter der semantischen Orientierung

Diese jeweils sieben Begriffe werden hinsichtlich deren Polarität als Referenz verwendet, sie werden nicht als Trainingsdatensatz eines Algorithmus angesehen.
Um die Stärke des semantischen Zusammenhangs zwischen zwei Wörtern zu messen wird zuerst die Pointwise Mutual Information (PMI) berechnet, vgl. Church and Hanks, 1989. Diese ist wie in Gleichung 6.1 definiert,

$$\text{PMI(Wort 1, Wort 2)} = log_2 \left\{ \frac{\text{P(Wort 1} \wedge \text{Wort 2)}}{\text{P(Wort 1)P(Wort 2)}} \right\}. \tag{6.1}$$

Dabei kann für Wort 1 und Wort 2 ein einzelnes Wort oder eine Kombination aus mehreren Wörtern eingesetzt werden. Hier wird für Wort 1 die aus dem Text extrahierte Kombination und für Wort 2 die positiven oder negativen Referenzwörter eingesetzt, wodurch sich zum einen die negative PMI, PMI(Phrase, negativ) und zum anderen die positive PMI PMI(Phrase, positiv) ergibt.
Der Zähler repräsentiert die Wahrscheinlichkeit dass Wort 1 und Wort 2 gemeinsam in einem Satz oder Dokument auftreten. Der Nenner beinhaltet jeweils die Einzelwahrscheinlichkeiten, dass Wort 1 oder Wort 2 auftreten. Bei Unabhängigkeit entspricht der Zähler dem Produkt der beiden Einzelwahrscheinlichkeiten im Nenner, womit der Bruch Eins ergibt, und durch die Berechnung des Logarithmus einen Zusammenhang von Null ergibt. Der gesamte Bruch ist ein Maß für die statistische Abhängigkeit der beiden Wörter, je größer der Zusammenhang der Wörter, desto größer die Werte in die positive oder negative Richtung.
Die Wahrscheinlichkeiten werden durch die Eingabe des einzelnen Wortes, sowie die beiden Wörter gleichzeitig in eine Suchmaschine generiert und die entsprechende Anzahl der Treffer, der relevanten Dokumente, notiert.
Die Autoren des Papers verwenden die Suchmaschine AltaVista. Diese bot eine Volltextsuche nach relevanten Seiten im Internet. AltaVista wurde im Jahr 2013 durch die Suchmaschine von Yahoo.com ersetzt. `http://www.heise.`

`de/newsticker/meldung/Yahoo-macht-Altavista-dicht-1908789.html`. Bei Aufruf der Suchmaschine wird der Anwender direkt auf die Webseite von Yahoo.com umgeleitet. AltaVista bot die NEAR Funktion an, durch welche die Suche eingegrenzt werden kann. Die eingegebenen Begriffe wurden hierbei entsprechend der Reihenfolge berücksichtigt und die Dokumente in welchen die Begriffe gemeinsam vorkommen als relevante Treffer angezeigt. Heute können andere Suchmaschinen, wie `https://www.google.de`, verwendet werden. Die Suche kann eingegrenzt werden indem die Begriffe in Anführungszeichen gesetzt werden, dabei werden nur Dokumente als relevant markiert die exakt diese Begriffe in der eingegebenen Reihenfolge beinhalten. Setzt man jeden Begriff einzeln in Anführungszeichen so werden nur Dokumente angezeigt die alle Begriffe beinhalten. Durch eine Tilde vor dem Wort werden Dokumente die den Begriff und Synonyme für diesen Begriff enthalten als relevant markiert. Durch einsetzen der Begriffe 'and' und 'or' werden Seiten ausgegeben die entweder alle Begriffe oder zumindest einen der beiden Begriffe beinhalten.

Die SO ist die Differenz der positven und der negativen PMI, vgl. Turney, 2002a und Turney, 2002b,

$$\text{SO(Phrase)} = \text{PMI(Phrase, positiv)} - \text{PMI(Phrase, negativ)}. \tag{6.2}$$

Durch einige algebraische Umformungen und die Verwendung der Treffer statt der Wahrscheinlichkeiten kann die SO vereinfacht dargestellt werden siehe Gleichung 6.3. Die SO ist positiv wenn die Assoziation zu den positiven Referenzwörtern größer als zu den negativen Referenzwörtern ist.

$$\text{SO(Phrase)} = log_2 \left\{ \frac{Treffer(\text{Phrase, positiv}) \times Treffer(\text{negativ})}{Treffer(\text{Phrase, negativ}) \times Treffer(\text{positiv})} \right\} \tag{6.3}$$

Um eine Division durch Null zu verhindern wird 0.01 zu jedem Treffer addiert. Desweiteren werden Anzahlen unter vier für $Treffer$(Phrase, positiv) und $Treffer$(Phrase, negativ) verworfen.

Zur Bewertung eines Textes wird im dritten Schritt Durchschnitt der einzelnen SO der Phrasen des Eingabetextes berechnet und somit ein Wert für die Stärke der Polarität einer Beurteilung generiert.
Folgende Tabelle wurde aus Turney, 2002a entnommen und zeigt die SO der extrahierten Phrasen einer exemplarsichen Bewertung der Bank of America.

Extracted Phrase	Semantic Orientation
online experience	2.253
low fees	J 0.333
local branch	0.421
small part	0.053
online service	2.780
printable version	-0.705
direct deposit	1.288
well other	0.237
inconveniently located	-1.541
other bank	-0.850
true service	-0.732
Average Semantic Orientation	0.322

Tabelle 6.2: Beispiel Semantische Orientierung einer Beurteilung

Die erste Spalte gibt die extrahierte Wortkombination bestehend aus zwei Wörtern an, eine geringere oder höhere Anzahl an extrahierten Wörtern wäre ebnefalls denkbar. Die dritte Spalte enthält das Ergebnis für die SO der einzelnen Phrasen. In der letzten Zeile der Tabelle befindet sich das arithmetische Mittel der SO der einzelnen Phrasen und wird als gesamt SO der Beurteilung angesehen. Die SO der betrachteten Bewertung kann als positiv eingestuft werden.
Die von Turney, 2002a und Turney, 2002b vorgeschlagene Methode der SO ist eine einfache und flexible unüberwachte Lernmethode. Diese ist mit einfachen Mitteln zu programmieren. Die Polarität der Referenzwörter kann abhängig von der Domain der Beurteilung gewählt werden und die SO somit Domänen spezifisch berechnet werden. Desweiteren ist die Methode nicht nur auf Adjektive beschränkt, die extrahierten Phrasen können dementsprechend angepasst werden, so dass auch andere Wortarten oder Redewendungen o.ä. berücksichtigt werden. Die Methode benötigt einen großen Korpus an Wörtern. Liefert aber auch mit einer geringeren Anzahl an verfügbaren Informationen robuste, akkurate Ergebnisse bei der Klassifizierung von Beurteilungen, Turney, 2002b. Turney, 2002a weisen darauf hin, dass die SO und die Fünf Sterne Kategorisierung von Beurteilungen eine positive Korrelation aufweisen, in dem Sinne je höher die SO desto höher die Anzahl der Sterne die ein Autor vergeben hat.
Die Rechenzeit und der Aufwand der Berechnung kann unter Umständen hoch sein. Je größer die zu durchsuchende Information und die Anzahl der Wörter bzw. Referenzwörter, sowie je länger die Beurteilung ist, desto höher die Zeit die von der Suchmaschine benötigt wird.

Ein weiterer erwähnenswerter Punkt ist die mögliche Verzerrung die bei der Klassifizierung einer Beurteilung auftreten kann. Da die einzelnen extrahierten Phrasen durch die Berechnung des arithmetischen Mittels gleich gewichtet werden kann im äußersten Fall eine stark negative Bemerkung bei einer sonst positiven (z.B. drei Sterne durch den Autor vergeben) Beurteilung einen großen Einfluss haben und die Bewertung fälschlicherweise als negativ eingestuft werden. Dies ist auch in die entgegengesetzte Richtung denkbar, eine negative Bewertung wird f'laschlicherweise als positiv eingestuft.
Als weiterer Punkt ist hier zu erwähnen, dass die SO maßgeblich von den Referenzwörtern abhängt. Darüber kann im Vorfeld das Ergebnis stark beeinflußt werden indem seltene Wörter als Referenz gewählt werden und somit die Trefferzahlen der relevanten Ergebnisse gesteuert werden.
Ein weiterer negativer Punkt ist die starke Abhängigkeit vom herangezogenen Sentiment Lexikas, so könnten aufgrund des fehlens eines sehr starken wichtigen Wortes Bewertungen fälschlicherweise einer Polarität zugeordnet werden. Zum anderen könnte, durch Verwendung verschiedener Lexika andere Polaritäten der Bewertung ergeben.

Zur Verbesserung könnte eine Kombination aus mehreren Lexikas verwendet werden, z.B. könnte man diese vor der Berechnung der SO kombinieren. Anschließend berechnet man den Mittelwert der jeweiligen Scores der unterschiedlichen Lexikas. Dieser Score könnte dann erst zur Bewertung der einzelnen Phrasen verwendet werden.

6.2 Überwachte Verfahren

Verfahren nach Hatzivassiloglou and McKeown, 1997

Als ein frühes Verfahren stellen Beispielsweise stellen Hatzivassiloglou and McKeown, 1997 eine Methode vor, mit deren Hilfe sich die semantische Orientierung von Adjektiven schätzen lässt. Dabei wird ein Log-Lineares Modell verwendet um zu schätzen ob die mit einer Konjunktion verbundenen Adjektive die selbe oder eine unterschiedliche Orientierung aufweisen. Auf dieses Verfahren wird nicht näher eingegangen.
Sämtliche Methoden zur Klassifikation der Hotelbewertungen ab Kapitel 9 sind ebenfalls dem überwachtem Lernen zuzuordnen.

Kapitel 7

Deskriptive Beschreibung des Datensatzes

Der in den nachfolgenden Analysen verwendete Datensatz wurde manuell erstellt, und besteht aus deutschsprachigen Hotelbewertungen, des Portals `https://www.holidaycheck.de`. *holidaycheck.de* ist nach eigener Angabe das grösste deutschsprachige Bewertungsportal und umfasst sechs Millionen Bewertungen zu 600000 Hotels. Über eine Suchoptionen kann die Auswahl der Hotels eingegrenzt werden, wie beispielsweise die Hotelkategorie, das Reiseziel, die Ausstattung des Hotels., u.v.m.
Beim bewerten eines Hotels, muss vom Gast angegeben werden ob dieser das Hotel weiter empfehlen würde. Dazu dienen zwei Buttons, Daumen hoch, Daumen runter, ein gesenkter Daumen steht für eine negative Polarität und ein erhobener Daumen gibt eine positive Polarität an. In weiteren Analysen wird diese Angabe als Variable Polarität angesehen, Desweiteren wird das Hotel über die Angabe einer bestimmten Anzahl an Sonnen evaluiert. Tabelle 7.1 gibt einen Überblick über die einzelnen Bewertungskategorien. Die Anzahl der Sonnen wird im folgenden als Bewertung bezeichnet.

Sonnen	1	2	3	4	5	6
Bewertung	sehr schlecht	schlecht	eher schlecht	eher gut	gut	sehr gut

Tabelle 7.1: Tabelle der Bewertungskategorien

Im nächsten Schritt muss ein mindestens 100 Zeichen langer Kommentar verfasst werden. Im folgenden als Text oder Kritik bezeichnet. Zusätzlich können optional Unterkategorien wie der Service oder das Zimmer, etc. ebenfalls mit der Angabe einer bestimmten Anzahl an Sonnen bewertet und mit einem separaten Kommentar versehen werden. Hierbei wird dann ein Mittelwert über

die Sonnen gebildet und dieser als Gesamtbewertung angegeben. Der Titel der Hotelbewertung muss vom Verfasser selbst erstellt werden, höchstens 50 Zeichen. Daneben werden Angaben zu seiner Person, das Geschlecht, das Alter, mit wem dieser verreist ist, die Art der Reise, die Anzahl der Kinder, die Länge des Aufenthalts, der Monat des Aufenthalts und die Nationalität erhoben.

Zur Erstellung des Datensatzes wurden folgende Variablen aufgenommen, Geschlecht, Alter, Reiseart, Monat, Jahr, Polarität, Bewertung und der Text. Als Grundlage dienten Hotels der selben Stadt innerhalb der gleichen Kategorie. Für jedes Hotel liegen meist über 95% positive Hotelbewertungen vor. Der Datensatz umfasst sämtliche Hotelbewertungen drei unterschiedlicher Hotels. Um eine höhere Anzahl an negativen Hotelbewertungen zum Training der Algorithmen zu gewährleisten wurden die noch bis zur Anzahl von 1000 Hotelbewertungen fehlenden Plätze mit negativen Hotelbewertungen anderer zufällig ausgewählten Hotels der selben Kategorie aufgefüllt.

Sämtliche Berechnungen und Grafiken werden mit der freizugänglichen Software R, *The R Project for Statistical Computing erstellt.*

7.1 Identifikation

Jeder Kritik wurde zur eindeutigen Identifikation eine Zahl $\in [1, 1000]$ zugewiesen.

7.2 Bewertung und Polarität

Grafik 7.1 zeigt die Anzahlen der Bewertungen innerhalb der jeweiligen Kategorien. Zunächst ist festzustellen dass innerhalb der Variablen Polarität und Bewertung jeweils weniger negative als positive Hotelbewertungen vorkommen. Wobei die Variable Bewertung ab der ersten Dezimalstelle abgerundet wurde. Hierfür wird die Annahme getroffen, dass der negative Aspekt einer Bewertung überwiegt. Wird beispielsweise das Zimmer mit vier Sonnen und der Service mit drei Sonnen bewertet, so ergibt sich eine durchschnittliche Bewertung von 3.5 Sonnen. Auf den Einfluss der unterschiedlichen Rundungsverfahren wird im späteren Verlauf noch einmal genauer eingegangen.

Fasst man die Kategorien '1' bis '3' der Variablen Bewertung als negativ zusammen, sowie die restlichen Kategorien als positiv, stellt man fest dass

127 negative und 873 positive Bewertungen vorliegen. Demgegenüber stehen 147 negative zu 853 positiven Hotelbewertungen der Variablen Polarität.

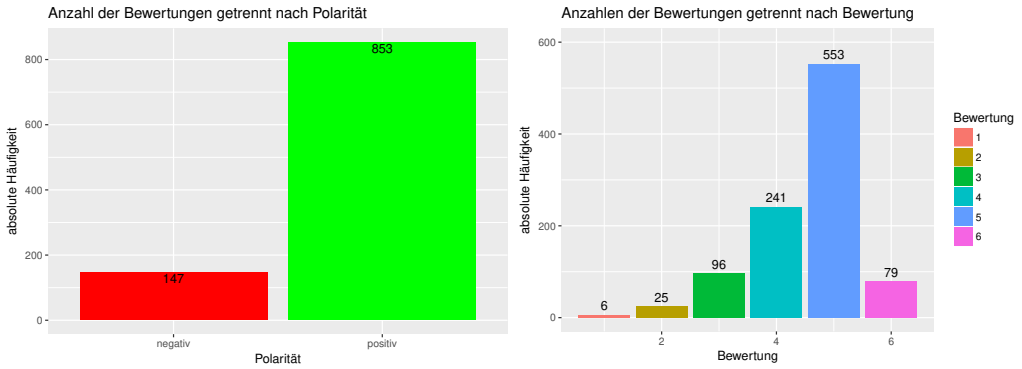

Abbildung 7.1: Absolute Häufigkeiten der negativen und positiven Hotelbewertungen im Datensatz getrennt nach den Variablen Polarität und Bewertung

Um den Umstand der unterschiedlichen absoluten Häufigkeiten innerhalb der beiden Variablen genauer zu untersuchen dient Grafik 7.2. Diese zeigt eine detaillierte Ansicht des Vergleichs der Variablen Polarität und Bewertung.

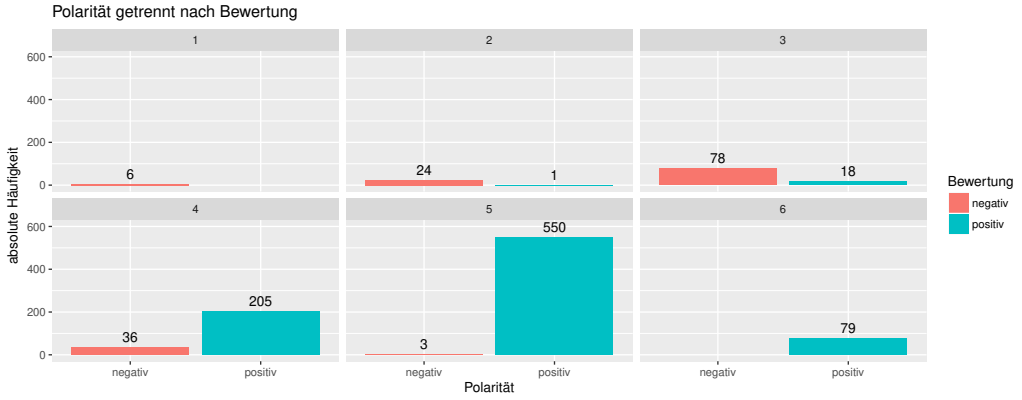

Abbildung 7.2: Vergleich der angegebenen Polaritäten und Bewertungen

In Grafik 7.2 werden die absoluten Häufigkeiten der Polaritäten innerhalb der Kategorien für die Bewertung separat aufgefürt. Im Idealfall sollten innerhalb der Kategorien '1' bis '3' keine positiven Hotelbewertungen existieren, sowie innerhalb der Kategorien '4' bis '6' keine negativ annotierten Kritiken. Diese Situation liegt in den Kategorien '1' und '6' vor. In den Kategorien '2' und '3' werden 24 bzw. 78 negativ und eine bzw. 18 positiv behaftete Hotelbewertungen abgegeben. In den Kategorien '4' und '5' sind 36 bzw. drei der

abgegebenen Hotelbewertungen negativ und 205 bzw. 550 positiv.

Ein Grund für die unterschiedliche Kategorisierung kann ein Missverständnis der Bewertungskategorien der Sonnen sein. Ein Gast könnte dies als Schulnoten interpretieren, somit entspräche die Vergabe einer Sonne der besten Note, die Vergabe von sechs Sonnen der schlechtesten Note.
Ein weiterer Aspekt kann sein, dass der Gast versehentlich den falschen Button wählt.
Auch kann diese Fehlklassifikation aufgrund der subjektiven Einstellung des Gastes entstehen. Hierfür wird folgende Vermutung aufgestellt.
Jede Person gewichtet einzelne Aspekte des Hotels anders, z.B. ist für Gast A der Umfang des Frühstücksbuffets von sehr hoher Bedeutung. Für Gast B ist die Größe der Zimmer ausschlaggebend. Wird z.B. das Frühstück mit einer '2' bewertet, hingegen die Zimmer mit einer '4', ergibt sich der Mittelwert von '3'. Für Gast A ist der Aufenthalt negativ und Gast B der Aufenthalt positiv. Wird nur auf die Bewertung Bezug genommen ist zu entnehmen dass der Aufenthalt von beiden Gästen als 'eher schlecht' wahrgenommen wurde. Jedoch entsprechen die subjektiven Wahrnehmungen einer anderen Polarität. Für Gast A ergäbe sich eine Fehlklassifikation.
Betrachtet man nun den Fall das Gäste keine besonderen Präferenzen für einzelne Aspekte des Hotels aufweisen, sondern nur einzelne Eindrücke bewerten. Beispielweise liegen nun die folgenden Bewertungen vor, '3' für das Frühstück und für die Zimmer '4', im Mittel 3.5. Es lassen sich wieder zwei Fälle unterscheiden (1) Für Gast C fließen positive Empfindungen stärker zur Meinungsbildung ein. Dieser wird den Aufenthalt als 'eher positiv' bewerten. Dies kann z.B. durch Aufrunden ab der ersten Dezimalstelle erfolgen. (2) Lässt Gast D negative Empfindungen stärker in die Meinungsbildung einfließen, wird der Aufenthalt als 'eher negativ' betrachtet. Dies kann z.B. durch das Abrunden ab der ersten Dezimalstelle erfolgen.
Aus den oben Aufgeführten Beispielen lassen sich die folgenden Annahmen herleiten, diese Annahmen basieren auf den Ausführen von Liu and Zhang, 2012.

- Präfernzen des Gastes:

  (1) Der Gast weist Präferenzen für einen bestimmten Aspekt des Hotels, wie z.B. der Sauberkeit des Hotels, der Qualität der Speisen, oder nur für Teilaspekte wie die Sauberkeit des Zimmers oder der Auswahl der Speisen am Buffet auf.

  (2) Der Gast besitzt keine Präferenzen gegenüber einzelnen Aspekten des Hotels.

- Empfindungen des Gastes bezogen auf Eindrücke:

  (1) Der Gast nimmt eine negative Empfindung stärker wahr als eine positive (pessimistisch).

  (3) Der Gast nimmt eine positive Empfindung stärker wahr als eine negative (optimistisch).

Weder die Präferenzen bezogen auf bestimmte Aspekte oder Empfindungen des Gastes sind beobachtbar, daher wird in späteren Modellen die Annahme getroffen, dass alle Aspekte mit der gleichen Gewichtung in eine Hotelbewertung einfließen, das heißt der Gast weist keine Präferenzen gegenüber eines bestimmten Aspekts auf. Zusätzlich wird angenommen der Gast sei pessimistisch. Aus Gründen der Einfachheit wird angenommen, dass keine Interaktionen zwischen den Präferenzen und der Einstellung des Gastes existieren.

Die Fälle der klassischen Rundung und dem Aufrunden ab der ersten Dezimalstelle wurden ebenfalls betrachtet. Dabei ergibt sich eine höhere Anzahl der Fehlklassifikationen. Die jeweiligen Grafiken sind dem Anhang beigefügt, wobei auf die Grafik ohne Rundung aus Gründen der Übersichtlichkeit verzichtet wurde. Die Tabelle 7.2 zeigt die Ergebnisse der Fehlklassifikation durch die unterschiedlichen Arten der Rundung. Dabei entspricht die jeweils erste Zeile der Gesamtanzahl der Bewertungen innerhalb der Bewertungskategorie. Die zweite Zeile gibt die Anzahlen der fehlklassifizierten Bewertungen an. Auf die Darstellung der Situation ohne jegliche Rundung wurde aus Gründen der Übersicht verzichtet, es sei an dieser Stelle lediglich erwähnt dass dabei Insgesamt 125 Fehlklassifikationen entstehen, dies entspricht dem höchsten Wert.

Bewertungskategorie	1	2	3	4	5	6
Abrunden	6	25	96	241	553	79
Fehlklassifiziert	0	1	18	36	3	0
klassisch	3	15	52	165	440	325
Fehlklassifiziert	0	1	2	66	14	0
Aufrunden	3	9	31	111	315	531
Fehlklassifiziert	0	1	1	76	28	2

Tabelle 7.2: Tabelle der Fehlklassifikationen aufgrund der Rundungsart

Je nach Rundungsart ändern sich die Gesamtzahlen in den Bewertungskategorien. Die höchste Gesamtanzahl an fehlklassifizierten Hotelbewertungen

(108) entsteht durch das Aufrunden. Die geringste Gesamtanzahl an Fehlklassifikationen (58) ergibt sich durch abrunden. Dies legt den Schluss nahe dass negative Empfindungen stärker wahrgenommen werden. Die fehlklassifizierten Kritiken, welche durch das Abrunden entstehen wurden dem Datensatz entnommen und bei der weiteren Modellen nicht berücksichtigt.

Zur Berücksichtigung der individuellen Einschätzung wäre eine unterschiedliche Gewichtung der Aspekte oder der Empfindungen denkbar. Dabei werden die Aspekte die durch den Gast präferiert werden mit höheren Gewichten versehen. Auch werden Empfindungen des Gastes entsprechend seiner Einstellung, pessimistisch oder optimistisch, höher oder niedriger gewichtet. Um dies anhand der Daten beobachten zu können, könnten Portale die Präferenzen und die Einstellung des Gastes als Variablen in die personenbezogenen Angaben aufgenommen werden. Dies könnte über anklicken eines jeweiligen Buttons für präferierte Aspekte und die eigene Einstellung zu Empfindungen durchgeführt werden.

7.3 Text

Der Text der Hotelbewertungen ist die Hauptvariable des Datensatzes. Dieser kann von beliebiger Länge sein und eine willkürliche Wortwahl aufweisen. Zuerst wurde den Originaltexten u.a. Zahlen, Satzzeichen und vor allem Stoppwörter entnommen. Stoppwörter sind Wörter die keine zusätzliche Information liefern, dazu zählen beispielsweise, Artikel, Pronomen, Konjugationen, Konjunktionen, etc. Aber auch Adjektive, Verben oder Nomen die sich nicht unmittelbar auf die Polarität oder ein Aspekt beziehen.

Zur grafischen Darstellung dienen *wordclouds* vgl. Grafik 7.4, diese bieten einen ersten Einblick über wichtige Schlagwörter innerhalb der Gesamtheit der betrachteten Texte, dem Korpus. Die Anzahl der durch die *wordclouds* dargestellten Wörter kann variabel eingestellt werden, hier die 250 am häufigsten verwendeten. Die linke Grafik basiert auf dem Korpus ohne Extraktion der Stoppworte. Für die rechte Grafik wurden die Stoppworte aus dem Korpus entfernt. Entsprechend der Häufigkeit im Korpus werden die Wörter der Größe nach abgebildet. Die am häufigsten erwähnten Wörter entsprechen der größten Darstellung.

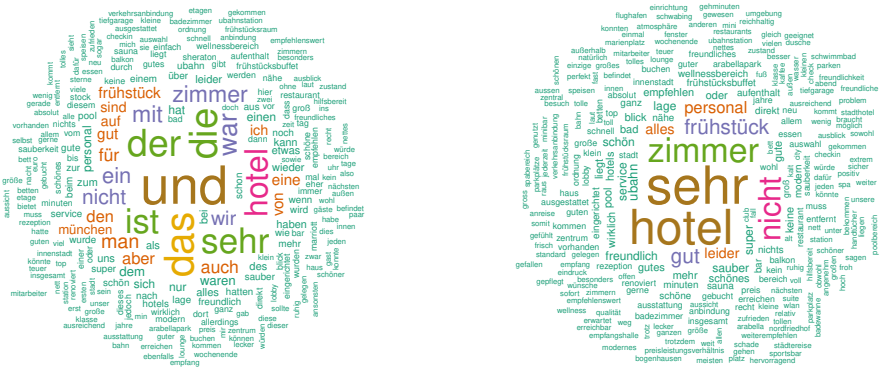

Abbildung 7.3: Wordcloud des Korpus, vor und nach Extraktion der Stoppworte, für die 250 meist verwendeten Wörter

In der linken Grafik ist zu erkennen, dass Stoppwörter wie 'und', 'das' etc., am häufigsten erwähnt werden, dadurch ist es kaum möglich eine Information bezüglich der beurteilten Aspekte bezüglich deren Polaritäten der Kritiken zu extrahieren. Die rechte Grafik zeigt eine *wordcloud* nach Extraktion von Stoppworten. Dadurch sind die in den Beurteilungen häufig angesprochenen Aspekte eines Hotes wie Zimmer, Frühstück und Personal erkennbar. Weniger häufig werden z.B. Frühstücksbuffet, -raum, Restaurant, Bad, Ausblick und Auswahl erwähnt. Hierüber lässt sich eine Hierarchie der für die Gäste wichtigen Aspekte erstellen, die Zimmer, das Frühstück und das Personal waren wichtiger als die U-Bahn Anbindung des Hotels oder der Ausblick. Adverben und Adjektive wie 'gut', 'leider', 'nicht', 'sehr', 'super', 'sauber' sind ebenfalls stark frequentierte Wörter des Korpus. Es kann vermutet werden, dass eine Vielzahl an Gästen zufrieden waren und deshalb positive Texte verfassen. Betrachtet man die Kombinationen von Aspekten mit Adverben oder Adjektiven kann man zu dem Schluss kommen, dass Floskeln wie beispielsweise, 'Das Zimmer / Personal war (nicht) sehr gut' häufig in den Texten vorkommen, oder auch die Sauberkeit ein häufig erwähntes Thema darstellt. *wordclouds* bieten somit die Möglichkeit erste Eindrücke über die Kritiken zu erlangen.
Daneben exitieren noch zwei weitere Versionen von *wordclouds*, zum einen die *comparisioncloud* und die *commonalitycloud*. Bei beiden Varianten wird der Korpus zuerst nach einem bestimmten Kriterium getrennt, hier wird nach der Polarität getrennt. Im Falle der *comparisioncloud* werden die am häufigsten verwendeten Wörter bezüglich des Kriteriums dargestellt. Die *commonalitycloud* bildet die Wörter die in beiden Kategorien häufig gemeinsam auftreten ab. Grafik 7.4 zeigt links eine *comparisioncloud* und rechts eine *commonalitycloud*, der jeweils 300 häufigsten Wörter.

Abbildung 7.4: Comparision Cloud und Commonality Cloud

Die *comparisioncloud* ist in zwei Kategorien unterteilt zum einen in positive (grün), zum anderen in negative Kritiken (rot) unterteilt.
Erkennbar ist, dass die Aspekte Frühstücksbuffet, Pool, Wellnessbereich, Sauna im positiven Kontext häufig erwähnt wurden. Auch die Lage, die U-Bahn sowie das Zentrum (Innenstadtbereich) werden thematisiert. Daneben sind Adjektive wie unter anderem 'super', 'freundlich' zu erkennen.
Im Segment der negativen Kritiken sind die Aspekte wie das Bad, die Empfangshalle, Zimmer und Bar aufgeführt. Wörter wie 'nicht', 'leider', 'kabuff' oder 'laut', sind verstärkt im Korpus aufzufinden.
Zusammenfassend kann daraus geschlossen werden, dass die meisten Gäste mit der Freundlichkeit des Personals, der Lage, dem Frühstücksbuffet, etc. zufrieden waren. Hingegen wurden z.B. die Bar, die Empfangshalle in negativen Kritiken am häufigsten erwähnt, hieraus kann geschlossen werden, dass eine Vielzahl der Gäste damit nicht zufrieden sind.
Betrachtet man die *commonalitycloud*, ist erkennbar, dass die Aspekte, Frühstück, Personal, Zimmer, Service in beiden Segmenten häufig erwähnt werden. Auch werden 'gut', 'nicht', 'sehr', 'leider', 'schönes' ebenfalls erwähnt. Hieraus lässt sich ableiten, dass die am häufigsten bewerteten Aspekte, das Frühstück, die Zimmer und das Personal sind. Ebenfalls lässt sich hieraus vermuten, dass die Wortwahl innerhalb der Kritiken ähnlich sind. Hierdurch kann die Aussage getroffen werden, dass die meisten Gäste wert auf das Frühstück, die Größe und die Sauberkeit der Zimmer, der Freundlichkeit des Personals, aber auch auf die Lage des Hotels legen.

Weitere Darstellungen von Texten bzw. von Aspekten und deren zugehörigen Polaritäten werden von James et al., 2013 vorgeschlagen. Auf diese graphischen Darstellungen wird im Rahmen der Arbeit jedoch nicht eingegangen.

Balkendiagramme

wordclouds bieten einen qualitativen Einblick, jedoch können keine detaillierten Aussagen über explizite Häufigkeiten getroffen werden. Eine Möglichkeit um erste Einblicke in die Datenlage zu erlangen bieten Balkendiagramme. Diese geben Auskunft über die Struktur der am häufigsten gewählten Wörter und lassen Aussagen über die absolute Häufigkeit der Erwähnung eines Wortes zu. Grafik 7.5 zeigt ein Balkendiagramm der 20 am meisten verwendeten Wörter des Korpus.

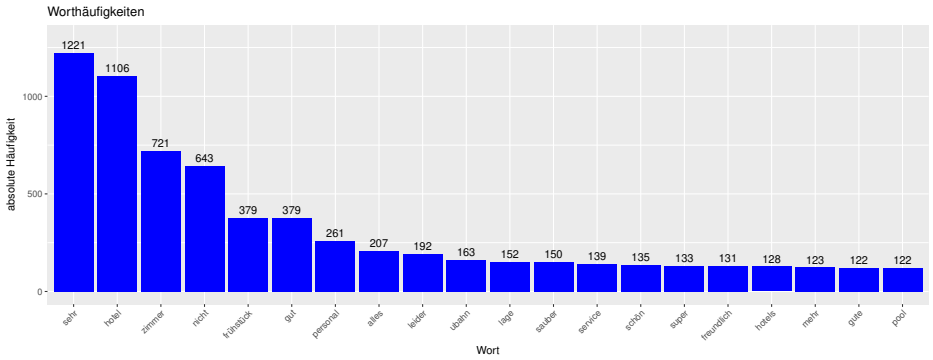

Abbildung 7.5: Balkendiagramm der 20 meist verwendeten Wörter im Korpus

Die Aussagen des Balkendiagramms decken sich mit der Interpretation der *wordclouds*.
Abbildung 7.6 zeigt jeweils die 20 am häufigsten benutzten Wörter im positiven bzw. negativem Kontext.

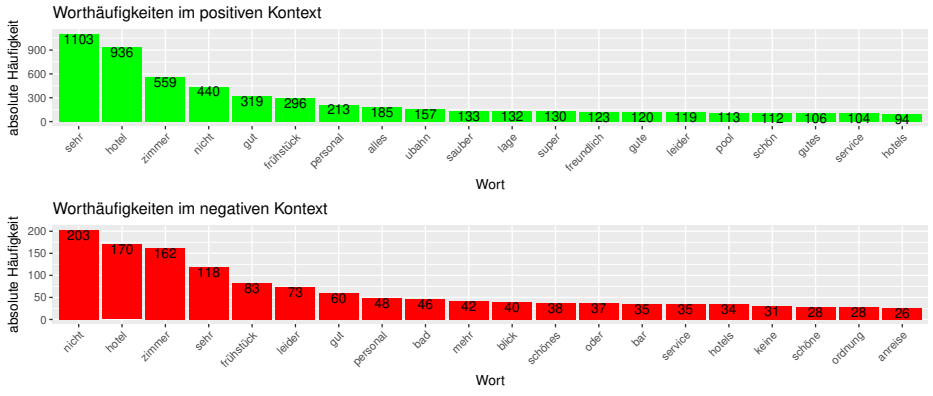

Abbildung 7.6: Balkendiagramm der 20 meist verwendeten Wörter im positiven und negativen Kontext

Die unterschiedlichen Häufigkeiten der meist genannten Wörter in den beiden Fällen der Polarität resultieren aus aus der Tatsache, dass im gesamten Korpus ca. 15% negative und ca. 85% positive Bewertungen enthalten sind. Wodurch in Summe gesehen die Wörter in positiven Texte unter normalen Umständen öfter vorkommen können.
Die am häufigsten angesprochenen Aspekte im positiven Kontext, die Zimmer das Frühstück, das Personal, die U-Bahn, die Lage, der Pool und der Service. Die Adjektive und Adverbien, z.B. 'nicht', 'gut', 'sauber', 'super', 'freundlich'. Hieraus, lässt sich ebenfalls ableiten, dass die meisten Gäste beispielsweise mit der Freundlichkeit des Personals, der Lage des Hotels, oder der nähe zur U-Bahn zufrieden waren. Das erscheinen des Wortes 'nicht' im positiven Kontext ist darauf zurückzuführen, dass in positiven Kritiken auch negative Aspekte angesprochen werden, wie 'Das Frühstück war gut, aber die Aussicht war nicht gut.' Daneben sind auch Floskeln wie 'Das Personal war nicht schlecht', oder 'Das Hotel lag nicht weit vom Zentrum entfernt.', denkbar.
Betrachtet man das Balkendiagramm des negativen Korpus können die Aspekte Zimmer, Frühstück, Personal, Bad, Bar, Service , Blick identifiziert werden. Die Adjektive und Adverbien, 'nicht', 'leider', 'gut', 'schönes', 'keine', 'schöne'. Eine mögliche Interpretation ist, das eine Vielzahl der Gäste mit dem Frühstück, dem Zimmer, dem Personal, oder der Ordnung des Hotels nicht zufrieden gewesen sind. Auffallend ist auch das Adjektive wie 'nicht', 'sehr', 'gut' häufiger verwendet wurden, als negative Adjektive wie 'schlecht'. Dies ist damit zu erklären, dass Floskeln wie 'nicht sehr gut' oder 'nicht schön' häufiger verwendet werden als das negative Pendant.
Im negativen sowie positiven Kontext herrschen ähnliche Aspekte und eine ähnliche Wortwahl vor, z.B. das Frühstück und das Personal. Dies liegt daran, dass im Allgemeinen den Gästen diese Aspekte wichtig erschienen und die Wortwahl der Texte sehr ähnlich gestaltet sind.
Die Interpretation der Balkendiagramme führen zu vergleichbaren Ergebnissen wie die Interpretationen der oben diskutierten *wordclouds*.
Wie hier festgestellt werden kann bieten Balkendiagramme den Vorteil quantitative Aussagen über die Verwendung einzelner Wörter treffen zu können. Demgegenüber steht die Tatsache dass die Übersichtlichkeit eines Balkendiagramms stark von der Größe des Mediums abhängt auf dem es betrachtet wird und somit die Anzahl der Balken nicht beliebig groß sein kann. Demzufolge ist in Fällen in denen nur ein qualitativer Überblick über die Schlagwörter des Korpus erlangt werden soll das Heranziehen von *wordclouds* ratsam.

7.4 Länge der Texte in Bezug auf die Polarität und Bewertung

Im folgenden Abschnitt wird die Anzahl der Wörter des einzelnen Textes betrachtet.

Tabelle 7.4 zeigt in detaillierter Form die jeweiligen fünf Punktezusammenfassungen welche den Berechnungen und den Grafiken der nächsten beiden Abschnitte zugrunde liegen.

Polartiät

Die Duchschnittliche Länge eines Textes beträgt ca. 63 Wörter. Abbildung 7.7 repräsentiert die Boxplots der Länge einer Kritik unterteilt nach der Polarität. Die rechte Grafik entspricht der linken Grafik, jedoch wurde aus Gründen der Übersicht die $Y - Achse$ beschränkt.

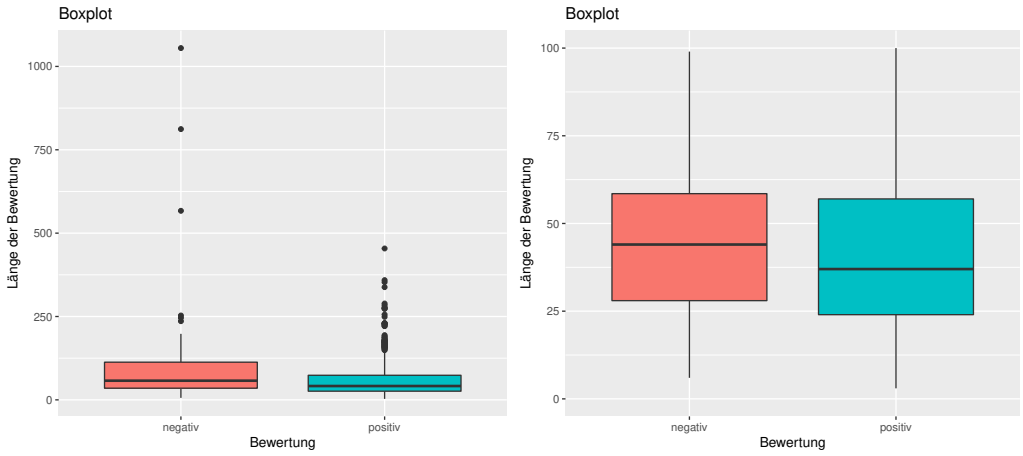

Abbildung 7.7: Boxplot der Länge einer Kritik, nach Polarität getrennt

Ersichtlich ist, dass sich keine großen Unterschiede zwischen den beiden Boxplots ergeben. Der Boxplot der positiven Kritiken wirkt im Vergleich etwas größer und der Median ist etwas niedriger als der Median der negativen Kritiken. Man könnte vermuten, dass positive Nachrichten tendenziell kürzer als negative sind.

Bewertung

In Abbildung 7.8 sind die Boxplots für die abgerundeten Werte der Bewertung zu sehen. Die rechte Grafik repräsentiert einen Ausschnitt der linken

Grafik, die $Y - Achse$ wurde zur besseren Übersicht beschränkt.

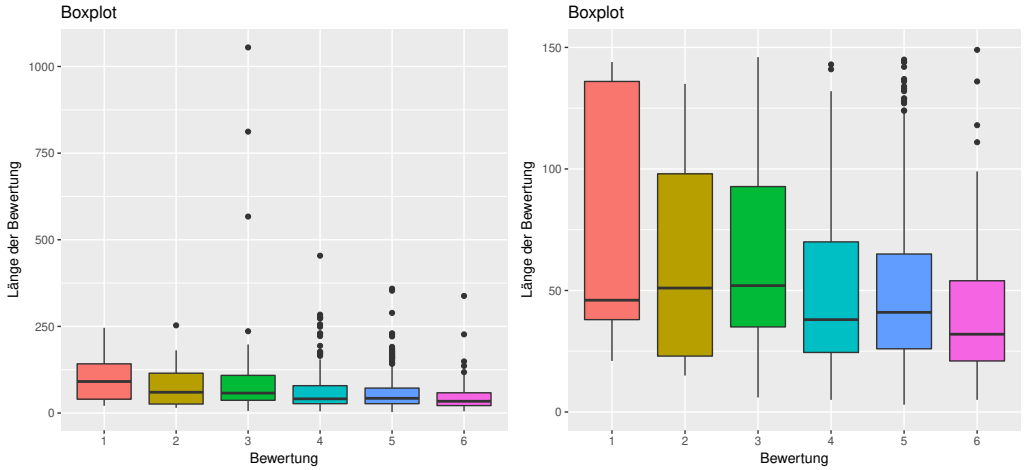

Abbildung 7.8: Boxplot der Länge einer Kritik, nach Bewertungen getrennt

Kategorie '1' der Bewertungen besitzt die größte Spannweite der Anzahlen der Wörter je Kritik, jedoch ist hier zu erwähnen dass lediglich sechs Kritiken in dieser Klasse exisitieren. Bei Betrachtung der restlichen Boxplots kann wieder vermutet werden, dass die Länge der Kritik abnimmt je besser die Bewertung ausfällt.
Aufgrund der wenigen Beobachtungen in den Klassen '1' und '6' werden die späteren Klassifikationsverfahren lediglich für die Variable Polarität durchgeführt.
Für die Variablen Polarität und Bewertung wurden jeweils getrennt Ausreisser bzw. Extremwerte identifiziert und extrahiert. Als Ausreisser werden Beobachtungen definiert die weiter als das 1.5 fache der Box-Länge vom $x_{0.25}$-Quantil bzw. vom $x_{0.75}$-Quantil entfernt liegen. Für Extremwerte gilt dies ab einer Entfernung vom dreifachen der Box-Länge, vgl. Toutenbug and Heumann, 2008. Bei der Polarität wurden neun negative Beobachtungen und 12 positive Beobachtungen aus dem Datensatz entfernt, vgl. Grafik 7.7.
Im Falle der Variablen Bewertung gibt Tabelle 7.4 einen Einblick über die Ausreisser und Extremwerte.

Bewertung	1	2	3	4	5	6
Ausreisser	0	1	1	10	29	3
Extremwerte	0	0	3	8	7	2

Tabelle 7.3: Tabelle der Anzahlen der Ausreisser und Extremwerte der Variablen Bewertung

Nach Entfernung der Ausreisser und Extremwerte ergab sich für den Datensatz von 882 Beobachtungen.
Tabelle 7.4 zeigt in der ersten Zeile die jeweiligen fünf Punktezusammenfassungen vor der Identifikation und dem Herauslassen der Ausreisser und Extremwerte. In Klammern sind die Werte nach der Extraktion aufgelistet.

Kategorie	Min	$x_{.25}$	Median	$\bar{x}$	$x_{.75}$	Max	Anzahl
1	21	40	91	105.2	142	246	6
	(21)	(40)	(91)	(105.2)	(142)	(246)	(6)
2	15	26	60	78.21	115	253	24
	(15)	(25)	(56)	(70.61)	(109.5)	(181)	(23)
3	6	36.75	57.50	99.96	108.8	1055	78
	(6)	(35.25)	(54)	(69.28)	(98.75)	(198)	(74)
4	5	27	41	65.93	79	454	205
	(5)	(25)	(38)	(49.56)	(70.5)	(155)	(187)
5	3	27	41.5	58.19	72	359	550
	(3)	(26)	(41)	(49.2)	(64)	(137)	(514)
6	5	21.5	34	48.41	58.5	338	79
	(5)	(20.25)	(30.5)	(38.59)	(51.50)	(111)	(74)
Negativ	6	35	57.5	95.42	113	1055	108
	(6)	(35)	(54)	(69.96)	(105.5)	(198)	(102)
Positiv	3	26	41.5	59.17	73.75	454	834
	(3)	(25)	(40)	(48.83)	(65)	(145)	780

Tabelle 7.4: Tabelle der fünf Punkte Zusammenfassungen

Erkennbar ist, dass sich die Anzahlen der Wörter der Kritik in den meisten Fällen nach unten korrigieren.

Kapitel 8

Modelle

Dieser Abschnitt dient der Klärung der Fragen,
(1) Beeinflußt die Länge einer Kritik die Polarität einer Hotelbewertung?
(2) Besitzt die Polarität oder die Bewertung einen Einfluss auf die Länge eines Texts?

8.1 Einfluss der Länge des Textes auf die Polarität

Zur Beantwortung der Frage, ob die Länge der Kritik einen Einfluss auf die Polarität besitzt, dient ein generalisiertes logistisches Regressionsmodell mit linearem Prädiktor für einen binomialen Response $y \sim B(n, \pi = 0.5)$, vgl. Fahrmeir et al., 2009 und Tutz et al., 2007. Das Modell ist definiert durch,

$$P(y = 1|x) = \pi(x) = \frac{exp(\eta)}{1 + exp(\eta)} = \frac{exp(\beta_0 + x_i\beta_r)}{1 + exp(\beta_0 + x_i\beta_r)}. \tag{8.1}$$

Wobei $P(y = 1|x) = \pi(x)$ der bedingte Erwartungswert einer positiven Kritik, gegeben den Beobachtungen x entspricht. Mit $\beta = \{\beta_0, \beta_1, \ldots, \beta_R\}$ den $r = 1, \ldots, R$ Regressoren,
Die Chancen sind definiert durch,

$$exp(\eta) = \frac{\pi(x)}{1 - \pi(x)}. \tag{8.2}$$

Als abhängige Variable wird die dichotome Variable *Polarität* angesehen und die Länge einer Kritik als Regressor.

Es ergab sich eine multiplikative Änderung der Chance um den Faktor $exp(\beta) = 0.984$, zu einem Signifikanzniveau von $1.08 * 10^{-8}$.

Die Chance eine positive Bewertung zu beobachten sinkt somit mit steigender Anzahl an Wörtern. Dies zu vernachlässigen, da nahezu kein Erklärungswert vorliegt.
Die Chance beträgt ohne Entfernung der Ausreißer und Extremwerte 0.9945628. Im Falle der Exklusion von ausschließlich Extremwerten 0.9898518. Beide Regressoren zeigten sich als signifikant zum Niveau von 0.00. Die Interpretationen bleiben gleich.

Desweiteren wurde ein Waldtest, mit der allgemeinen Nullhypothese vgl. Fahrmeir et al., 1996 und Fahrmeir et al., 2009,

$$H_0 : C\hat{\theta} = d \quad vs. H_1 : C\hat{\theta} \neq d, \tag{8.3}$$

durchgeführt. Die zugehörige Teststatistik lautet,

$$w = (C\hat{\theta} - d)^T (C\hat{V}C^T)^{-1}(C\hat{\theta} - d). \tag{8.4}$$

$\hat{V}$ entspricht der geschätzten Fischer Information des Schätzers $\hat{\theta}$. $C\hat{V}C^T$ repräsentiert die geschätzte Kovarianzmatrix von $C\hat{\theta} - d$. Im Spezialfall eines Regressors mit der Nullhypothese $H_0 : \beta_r = 0$ kann dies wie folgt geschreiben werden,

$$w = \frac{\hat{\beta}_r^{\,2}}{Var(\hat{\beta}_r)} \sim \chi^2_{1-\alpha,1}, \tag{8.5}$$

wobei $Var(\hat{\beta}_r)$, das entsprechende Diagonalelement der geschätzten Kovarianzmatrix ist. $\chi^2_{1-\alpha,1}$ ist das $(1-\alpha)$-Qunatil der χ^2-Verteilung.
Die Nullhypothese kann abgelehnt werden falls,

$$w^2 > \chi^2_{1-\alpha,1}. \tag{8.6}$$

Es ergibt sich ein Prüfgrößenwert von 8.11. Das 0.95-Quantil der χ^2-Verteilung, mit einem Freiheitsgrad, ist gegeben durch 3.84. Womit die Nullhypothese abgelehnt werden kann. Der Regressor kann als signifikant angesehen werden, obwohl keine deutliche Änderung der Chance zu vermerken ist. Es kann somit darauf geschlossen werden dass $P(Y = 1|x) = P(y = 0|x) = 0.5$

8.2 Einfluss der Bewertung auf die Länge des Texts

Zur Beantwortung der Frage, ob die Kategorie der Bewertung einen Einfluss auf die Länge der Bewertung besitzt wurde ein lineares Intercept Modell

angepasst. Das lineare Modell ist für eine normalverteilte Zielgröße mit Erwartungswert μ und Varianz σ^2. Die kategoriale Einflussgröße x_r mit $x_r = 1$ falls Kritik der Bewertungskategorie r angehört und 0 sonst, vgl. Tutz et al., 2007, Fahrmeir et al., 2009 sowie Sachs and Hedderich, 2009.
Die Modellgleichung lautet,

$$y_i = \beta_0 + \beta_r x_r + \epsilon_i. \qquad i = 1, \ldots, 6 \tag{8.7}$$

Wobei der Intercept β_0 dem Mittelwert der Referenzkategorie entspricht. β_r repräsentieren die Koeffizienten der Variablen Bewertung. Der Normalverteilte Fehlerterm ist gegeben durch $\epsilon \sim N(0, \sigma^2)$.
Die jeweiligen Regressoren stellen die mittleren Abweichungen der Kategorien in Bezug auf die Referenzkategorie dar.

Basierend auf den Grafiken 7.7 und 7.8 wird ein Modell angepasst, für welches Kritiken mit mehr als 300 Wörtern extrahiert wurden. Der Tabelle 8.1 sind die Ergebnisse der jeweiligen Koeffizienten und deren Signifikanzniveaus zu entnehmen.

Koeffizient	Wert	Signifikanz
Intercept	105.17	0.00
β_2	-34.56	0.02
β_3	-35.88	0.01
β_4	-55.61	0.00
β_5	-55.96	0.00
β_6	-66.57	0.00

Tabelle 8.1: Koefifizienten des linearen Modells für die Länge der Kritik in Abhängigkeit der Bewertung

Die Referenzkategorie ist Kategorie '1', und entspricht einer durchschnittlichen Anzahl von 105.17 Wörtern. Wobei hier zu erwähnen ist, dass sich lediglich sechs Bewertungen in dieser Kategorie befinden.
Die Koeffizienten der Kategorie '2' sind zu einem Niveau 0.05 und der Kategorie '3' zu einem Niveau von 0.01 signifikant. Die Koeffizienten der Kategorien '4', '5' und '6' sind zu einem Niveau von 0.00 signifikant.
Wird eine Kritik der Kategorie zwei zugeordnet so senkt im Vergleich zur Referenzkategorie sich die Anzahl der Wörter im Mittel um ca. 35. Sämtliche Koeffizienten weisen einen negativen Einfluss auf, die Reduktion nimmt mit steigender Kategorie zu. Zusätzlich fällt auf, dass die Kategorien '2' und '3' sowie die Kategorien '4' und '5' einen sehr ähnlichen Einfluss besitzen.

Zusammenfassend kann gesagt werden, dass die Länge der Kritik mit steigender Zufriedenheit des Gastes abnimmt.

Um den generellen Einfluss der verschiedenen Bewertungskategorien auf die Zielgröße zu untersuchen wird ein F-Test herangezogen, vgl. Sachs and Hedderich, 2009 und vgl. Tutz et al., 2007.
Die Prüfgröße ist gegeben durch,

$$F = \frac{MQE}{MQR}, \tag{8.8}$$

wobei MQE der mittlere quadratische Fehler zwischen den Kategorien und MQR der mittlere quadratische Fehler innerhalb der Kategorien ist.

Die generelle Definition der Hypothesen des Tests für das $(1-\alpha)$-Qunatil ist,

$$H_0 : \beta_r = 0 \quad vs. \quad H_1 : \beta_r \neq 0 \text{ für mindestens ein } \beta_r. \tag{8.9}$$

Falls,

$$F > F_{1-\alpha,R-1,N-R} \tag{8.10}$$

erfolgt die Ablehnung der Nullhypothese.

Der Prüfgrößenwert entspricht 12.08. Das 0.95 - Quantil der Fisherverteilung mit 5 Parametern und 872 Beobachtungen entspricht 2,22. H_0 kann somit verworfen werden, dies bedeutet dass mindestens ein Koeffizient einen signifikanten Einfluss besitzt.

Eine Tabelle der Ergebnisse für die Berechnungen ohne Exklusion von Ausreissern, sowie der Exklusion von ausschließlich Extremwerten befindet sich im Anhang. Ebenfalls befindet sich die Tabelle der Ergebnisse des linearen Modells mit den zusätzlich aufgenommenen Variablen des Datensatzes. Ohne Entfernung der Ausreißer und Extremwerte enthält entspricht der Prüfgrößenwert 6.838 und im Falle der Herausnahme der Extremwerte ergab sich dieser Wert als 6.721. In beiden Fällen wird die Nullhypothese, wenn auch nicht so deutlich wie im Falle der Herausnahme der Ausreisser und Extremwerte, abgelehnt.

8.3 Einfluss der Polarität auf die Länge des Texts

Zur Beantwortung der Frage ob die Variablen *Polarität* einen Einfluss auf die Länge der Kritik aufweist, wird ein Mann-Whitney-U Test durchgeführt, vgl. Sachs and Hedderich, 2009 und Tutz et al., 2007.
Der Mann-Whitney-U Test ist ein parameterfreier Test, mit der Nullhypothese, dass zwei Stichproben der selben Population entstammen und keine von beiden tendenziell größere bzw. kleinere Werte aufweist. Dieser Test wird für das arithmetische Mittel der Anzahl der Wörter der Kritiken in den Unterpopulationen der Variablen *Polarität*, negativ und positiv durchgeführt. Die jeweiligen Mittelwerte entsprechen μ_{pos} und μ_{neg}. Die Hypothesen sind in diesem Fall folgendermaßen definiert,

$$H_0 : \mu_{neg} - \mu_{pos} \leq 0 \quad vs. \quad H_1 : \mu_{neg} - \mu_{pos} > 0. \tag{8.11}$$

Da insgesamt mehr als 25 Beobachtungen vorliegen, ist die Teststatistik Z wie folgt definiert,

$$Z = \frac{U - \frac{n_{neg} n_{pos}}{2}}{\sqrt{\frac{n_{neg} n_{pos}(n_{pos} + n_{neg} + 1)}{12}}}. \tag{8.12}$$

Wobei n_{neg} und n_{pos} die Anzahlen der Beobachtungen in der jeweiligen Subpopulation sind. U repräsentiert das Minimum der Größen U_1, U_2. Diese errechnen sich aus,

$$U_l = n_{neg} n_{pos} \frac{n_l(n_l + 1)}{2} - R_l, \quad l = \{neg, pos\}. \tag{8.13}$$

R_l entspricht der Summe der jeweiligen Ränge der Subpolpulation der gepoolten Stichprobe.
Die Nullhypothese kann abgelehnt werden falls,

$$Z > z_{1-\alpha}. \tag{8.14}$$

$z_{1-\alpha}$ entspricht dem $(1-\alpha)$-Quantil der Standardnormalverteilung.
Nach Durchführung des Tests ergab sich, dass die Anzahlen der Wörter der Kritiken hinsichtlich den Kategorien der *Polarität* signifikant unterscheiden. Insbesondere ist die Anzahl der Wörter pro Kritik in negativen Kritiken größer, zu einem $p - Wert$ von 1.936×10^{-4}.
Grafik 8.1 zeigt die Dichte der Anzahl der Wörter pro Kritik.

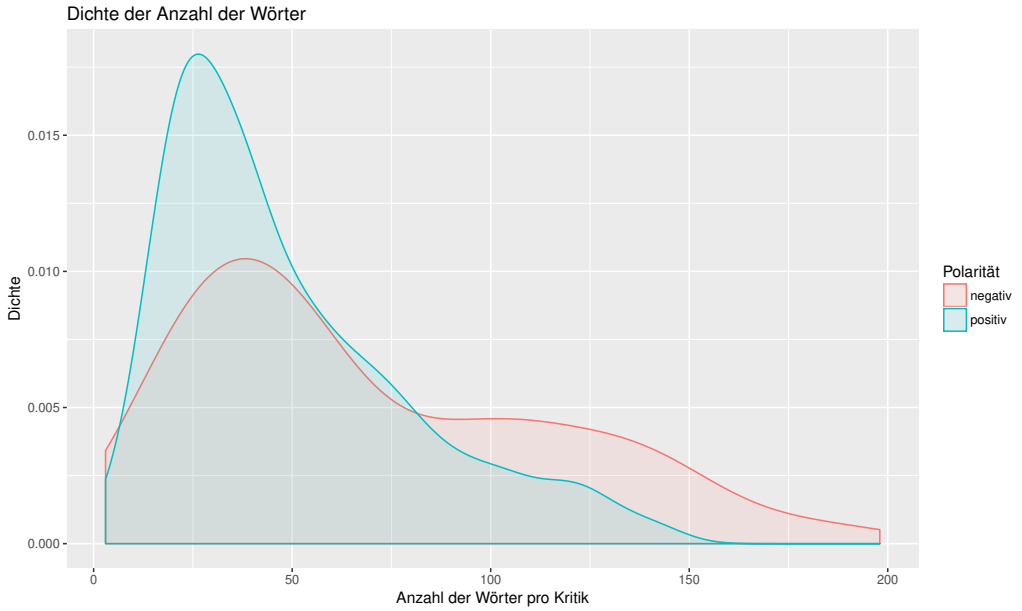

Abbildung 8.1: Dichte der Anzahl der Wörter getrennt nach Polarität

Dabei ist ersichtlich, dass im Bereich ab ca. 80 Wörtern negative Kritiken tendenziell eine höhere Anzahl an Wörtern aufweisen. Im Bereich bis 80 Wörtern liegen die Werte stets unterhalb.

Basierend auf dem Ergebnis des Mann-Whitney-U Tests und der Grafik 8.1 kann gesagt werden, dass negative Kritiken die Tendenz aufweisen eine höhere Anzahl an Wörtern zu besitzen. Jedoch sind im Falle kürzerer Kritiken diese eher positiv.

Kapitel 9

Modellbasierte Klassifikationsverfahren

In diesem Kapitel wird das Ziel verfolgt eine Hotelbewertung aufgrund des gegebenen Textes hinsichtlich der Polarität zu klassifizieren.
Zur Schätzung der Polarität finden das logistische Regressionsmodell, Support Vector Machines (SVM), der Ansatz der Maximum Entropie (MaxEnt), Verwendung. Die Modelle werden im ersten Abschnitt des Kapitels beschrieben. Im Anschluss daran werden diese zur Klassifikation verwendet und die Ergebnisse interpretiert.

9.1 Bag of Words Modelle

Bag of Words Modelle (BOW) basieren auf einer Document Term Matrix (DTM), mit $DTM \in \mathbb{R}^{n \times m}$. Wobei n der Anzahl der Hotelbewertungen entspricht. m entspricht der Anzahl der Wörter des Vokabulars. Jede Zeile repräsentiert eine Hotelbewertung und jede Spalte ein Wort. Der Zelleneintrag w_{ij} entspricht der absoluten Häufigkeit des Wortes w_j mit $j = 1, \ldots, m$ innerhalb der Hotelbewertung x_i, mit $i = 1, \ldots, n$. Die Hotelbewertungen $x_i \in X$ mit $x_i = \{w_{i1}, \ldots, w_{im}\}$ dienen als Eingabewerte der Modelle. Die Zellen der Wörter die nicht in der Kritik vorkommen werden auf '0' gesetzt. Eine andere Darstellung ist die $1 - of - V$ Kodierung, hierbei wird für das beobachtete Wort eine '1' eingetragen, unabhängig von dessen Häufigkeit.

Zusätzlich können DTM nicht nur für einzelne Wörter, sog. Unigrame, erstellt werden, sondern für N-Grame. N seht für die Anzahl der einbezogenen Wörter, beispielsweise werden bei Bigramen Wortkombinationen bestehend

aus zwei benachbarten Wörtern aufgegriffen. Hierdurch können Floskeln wie 'nicht gut' in die Modellierung einfließen. Unigrame berücksichtigen diesen Zusammenhang nicht.

Die nachfolgenden Modelle werden, bis etwas anderes erwähnt wird, basierend auf der DTM geschätzt.

9.2 Definition der Modelle

Das Logitmodell

Um die Polarität eines Textes der Hotelbewertungen zu klassifizieren wird ein generalisiertes Logitmodell herangezogen. Dabei wird für die binäre Zielvariable Polarität eine Binomialverteilung angenommen. Mit $Y \sim B(n, \pi)$, wobei $\pi = P(Y = 1|x)$ der erwarteten Auftretenswahrscheinlichkeit für eine positive Hotelbewertung, bedingt durch den Text, entspricht, siehe Fahrmeir et al., 2009 und Kapitel 8.1. Zur Klassifikation wird die penalisierte negative Log-Likelihood minimiert.

Diese ist gegeben durch,

$$\min_{\Theta} = -l(\Theta, x) + \lambda P(\Theta). \tag{9.1}$$

Θ entspricht dem interessierenden Parameter des Logit-Modells. Durch den Strafterm des Modells $\lambda P(\Theta)$ mit Kontrollparameter λ, wird die Stärke der Bestrafung zu komplexer Modelle gesteuert. Die Glättungsfunktion $P(\Theta)$ kann durch beispielsweise die Ridge-Regression oder Lasso-Regression geschätzt werden, vgl. Fahrmeir et al., 1996 und Tutz et al., 2007.

Support Vektor Machines

Support Vector Machines (SVM) können zur Klassifikation von Texten in negative oder positive Bewertungen herangezogen werden, vgl. Suykens and Vandewalle, 1999 und Pang et al., 2002.
Dazu werden die Tupel (x_i, y_i), in zwei Gruppen, hinsichtlich deren Polarität mittels einer Hyperbene getrennt, z.B einer linearen Geraden. Die Gruppen sollen so getrennt werden, dass der kleinste Abstand zwischen den Stützvektoren maximal ist und möglichst wenig Fehlklassifizierungen vorliegen.

Die Hyperebene ist definiert als,

$$y_i = sgn(< q, x_i > +b). \tag{9.2}$$

Wobei q den Normalenvektor und $< . >$ das Skalarprodukt repräsentiert. b entspricht einer Konstanten.

Die Nebenbedingung ist gegeben durch,

$$sgn = \begin{cases} 1 & < q, x_i > +b \geq 0 \\ -1 & < q, x_i > +b < 0. \end{cases} \tag{9.3}$$

Dies führt zu folgendem Optimierungsproblem,

$$\min_{q,b} ||q||_2 \text{ mit der Nebenbedingung } y_i(< q, x_i > +b) \geq 1. \tag{9.4}$$

$||q||_2$ entspricht der euklidischen Distanz.

Die Lösung findet über die Lagrange Methode mit dem Ansatz

$$L(q, b, \alpha) = \frac{1}{2}||q||_2 - \sum_{i=1}^{n} \alpha_i [y_i(< q, x_i > +b) - 1] \tag{9.5}$$

statt. α_i repräsentiert den Lagrange Multiplikator.

Die Maximum Entropie

Beim Verfahren der Maximum Entropie (MaxEnt) wird das Vorwissen aus den Daten genutzt um die Wahrscheinlichkeitsfunktion die das Wissen am besten abbildet zu schätzen. Dabei wird diejenige Wahrscheinlichkeitsfunktion mit der größten Entropie gewählt, diese sollte der Gleichverteilung möglichst ähnlich sein, vgl. Berger et al., 1996.

Das Prinip der Maximum Entropie wird im folgenden dargestellt, vgl. Berger et al., 1996.

Zuerst wird den Beobachteten Tupel eine gemeinsame Wahrhscheinlichkeit gemäß,

$$\tilde{p}(x,y) = \frac{1}{N} \times \text{Anzahl des gemeinsamen Auftretens von (x,y)} \tag{9.6}$$

zugeornet.

Mit der Indikatorfunktion,

$$f(x,y) = \begin{cases} 1 & y_i \in \{-1,1\} \text{ und x Wort } w_m \text{ enthält} \\ 0 & \text{sonst.} \end{cases} \tag{9.7}$$

Von Interesse ist der Erwartungswert bezogen auf die empirische Verteilungsfunktion $\tilde{p}(x,y)$,

$$\tilde{p}(f) = \sum_{x,y} \tilde{p}(x,y) f(x,y). \tag{9.8}$$

Der Erwartungswert des Modells ist definiert durch,

$$p(f) = \sum_{x,y} \tilde{p}(x) p(y|x) f(x,y). \tag{9.9}$$

Wobei $\tilde{p}(x)$ der empirischen Verteilung von x im Trainingsdatensatz entspricht.
Zusätzlich wird die Annahme getroffen,

$$\tilde{p}(f) = p(f) \tag{9.10}$$

Mit der maximalen Entropie,

$$p^* = \arg\max_{p \in C} \left\{ -\sum_{x,y} \tilde{p}(x) p(x|y) log(p(x|y)) \right\}. \tag{9.11}$$

Mit

$$C = \{p \in P | p(f_i) = \tilde{p}(f_i) \text{ für } i = 1, \ldots, N\}, \tag{9.12}$$

wobei C einer Teilmenge der unbedingten Verteilungsfunktionen entspricht.

Zur Lösung des Optimierungsproblems wird die Lagrange Methode herangezogen. Das Maximum ist gegeben durch,

$$p^*(y|x) = \frac{exp\left\{\sum_{i=1}^n \lambda_i f_j(x,y)\right\}}{\sum_y exp\left\{\sum_{i=1}^n \lambda_i f_j(x,y)\right\}}. \tag{9.13}$$

9.3 Anwendung

Trainings- und Testphase

Um den Trainingsdatensatz zu generieren werden zufällig die Tupel (x_i, y_i) ohne Zurücklegen gezogen. Die restlichen im Datensatz verbliebenen Hotelbewertungen werden als Testdatensatz angesehen. Die Orientierung der Texte ist bekannt und wird durch die Variable Polarität definiert. Aus den Texten der Trainings- und Testdaten wird eine gemeinsame DTM erstellt, so dass den Modellen das gleiche Vokabular zugrunde liegt. Dies ist für alle Methoden gleich.

Interpretation und Evaluation

Um die Ergebnisse eines Modells bzw. die Veränderung der Ergebnisse in bestimmten Situationen interpretieren zu können werden die Gesamttrefferrate, die Sensitivität, Spezifität und der F1-Score betrachtet. Die Modelle werden mithilfe der Fläche unterhalb der Receiver Operating Characteristic Kurve (AUC) evaluiert und miteinander verglichen, vgl. Sachs and Hedderich, 2009.
Da die Ziehung der Trainingsdaten vom Zufall abhängt wird der Vorgang des Ziehens und des Schätzens eines Modells wiederholt durchgeführt Um stabilere Ergebnisse zu erreichen werden die Analyseparameter gemittelt.

Definition der Analyseparameter

Ausgehend von einer Vierfeldertafel findet die Analyse statt.

		Prognose	
		negativ	positiv
Annotation	negativ	RN	FP
	positiv	FN	RP

Tabelle 9.1: Vierfeldertafel

RN und RP repräsentiert die Anzahl der richtig spezifizierten Texte, hier stimmt die durch das Modell prognostiezierte Orientierung einer Hotelbewertung mit der tatsächlichen Orientierung im Datensatz überein. FN und FP repräsentieren die absolute Häufigkeit der falsch spezifizierten Texte. Im Idealfall sollten keine fälschlicherweise negativ (FN) bzw. positiv (FP) prognostizierten Werte vorliegen.

Die absoluten Zellhäfigkeiten der Kontingenztafel werden im folgenden mit h_{ij} mit $i, j \in \{1, 2\}$ bezeichnet, wobei $h_{i.}$ der Zeilensumme und $h_{.j}$ der Spaltensumme entspricht, vgl. Tutz et al., 2007.
Die Übereinstimmung der Prognose mit der tatsächlichen Annotation wird als Treffer bezeichnet.

Die Gesamttrefferrate (GT) ist gegeben durch,

$$GT = \frac{h_{11} + h_{22}}{n}, \quad i \neq j. \tag{9.14}$$

Die GT ist der Anteil der Treffer innerhalb des Testdatensatzes und n die Anzahl der Tupel (x_i, y_i) im Testdatensatz.

Die Sensitivität (S) gibt die Wahrscheinlichkeit der richtig spezifizierten negativen Texte an, also die Wahrscheinlichkeit dass eine negative Prognose auch tatsächlich negativ ist.

$$\mathrm{S} = \frac{h_{11}}{h_{1.}}. \tag{9.15}$$

Die Spezifität (Sp) gibt die Wahrscheinlichkeit der richtig spezifizierten positiven Texte an,

$$\mathrm{Sp} = \frac{h_{22}}{h_{2.}}. \tag{9.16}$$

Die Präzision ist der Anteil der richtig klassifizierten Texte, innerhalb einer bestimmten Orientierung. Die Präzision der negativen Prognosen ist gegeben durch,

$$\mathrm{Pr}_{neg} = \frac{h_{11}}{h_{.1}}. \tag{9.17}$$

Die Präzision der positiven Texte (Pr_{pos}) wird durch das Verhältnis $h_{22}/h_{.2}$ betrachtet.

Das F1-Maß ist ein Kennwert für die durchschnittliche Meßgenauigkeit des Modells und kann entsprechend dem harmonischen Mittel interpretiert werden.

$$F1 = \frac{S \times Pr}{S + Pr}. \tag{9.18}$$

Wobei Pr der Präzision entspricht.
Es gilt,

$$\mathrm{S},\ \mathrm{Sp},\ \mathrm{Pr},\ \mathrm{F1} \in [0, 1], \tag{9.19}$$

je größer der Wert eines Maßes desto besser.

Zufallsziehung des Trainingsdatensatzes

Aus dem ursprünglichen Datensatz werden zufällig ohne Zurücklegen 500 Tupel gezogen und als Trainingsdaten verwendet. Zur Klassifikation wird ein Logitmodell mit $\lambda = 0.0001$ und Lasso Regression für den Penalisierungsterm verwendet.
Sämtliche Ergebnisse im Rahmen dieser Arbeit sind auf zwei Stellen nach dem Komma gerundet.

Wdh.	GT	S	Sp	Pr neg	Pr pos	F1 neg	F1 pos	AUC
1	0.88	0.19	0.96	0.38	0.91	0.25	0.93	0.68
10	0.86	0.16	0.96	0.32	0.90	0.21	0.92	0.69
25	0.86	0.15	0.96	0.32	0.90	0.20	0.93	0.70
100	0.86	0.15	0.96	0.32	0.90	0.20	0.93	0.69
200	0.86	0.15	0.96	0.32	0.90	0.20	0.93	0.69

Tabelle 9.2: Ergebnisse des Logit Modells bei verschiedenen Anzahlen von wiederholten Ziehungen des Trainingsdatensatzes

Tabelle 9.2 zeigt die durchschnittlichen Ergebnisse des Logit Modells, für verschiedene Trainingsdatensätze. Erkennbar ist hier, dass sich die Ergebnisse ab einer Anzahl von zehn Wiederholungen im Durchschnitt kaum unterscheiden. Aus diesem Grund wird im Folgenden von zehn Wiederholungen ausgegangen. Die verschiedenen Anzahlen der wiederholten Ziehung und Schätzung wurden ebenfalls für die beiden anderen Verfahren betrachtet, dabei zeigt sich das gleiche Bild.
Die nachfolgenden Interpretationen basieren auf durchschnittlichen Werten, dies wird jedoch nicht mehr explizit erwähnt.

Modell	GT	S	Sp	Pr neg	Pr pos	F1 neg	F1 pos	AUC
SVM	0.89	0.19	0.99	0.76	0.90	0.28	0.94	0.86
MaxEnt	0.90	0.42	0.97	0.61	0.93	0.49	0.95	0.78

Tabelle 9.3: Ergebnisse der SVM und MaxEnt10 Wiederholungen

Tabelle 9.3 präsentiert die Ergebnisse der SVM mit einer linearen Trennfunktion und MaxEnt für je 10 wiederholte Ziehungen des Trainingsdatensatzes.
Die Ergebnisse der Tabellen 9.2 und 9.3 können folgendermaßen interpretiert werden.
Die Gesamttrefferrate (GT) gibt den Anteil der Übereinstimmung der prognostizierten und tatsächlich annotierten Klasse einer Kritik an. Für alle drei Modelle ist in mindestens 86% der Fälle eine Übereinstimmung zu erkennen.

Wobei das Logit Modell am schlechtesten klassifiziert und MaxEnt den besten Wert liefert. Dies spricht grundsätzlich für eine Gute Anpassung.
Betrachtet man die Ergebnisse der Spezifität (Sp), der Präzision der positiven Prognosen (Pr pos) und den F1 - Score der positiven Kritiken (F1 pos), ist zu sehen dass die Werte mindestens 0.9 betragen. Damit kann mit ziemlich hoher Sicherheit behauptet werden dass bei vorliegen einer positiven Prognose auch tatsächlich eine positive Hotelbewertung vorliegt.
Bei Betrachtung der Sensitivität (S) ist zu sehen dass die Ergebnisse starken Schwankungen unterliegen. Für das Logit Modell ergibt sich ein Wert von 0.15 und die SVM weist einen Wert von 0.19 aus, was als eher schlechte Werte angesehen werden können. Im Falle der MaxEnt liegt dieser Wert bei 0.42 welcher als mittelmäßig zu bewerten ist. Im Falle des Logit Modells liegt die Präzision (Pr neg), bei 32%, welche ebenfalls als eher schlecht einzustufen ist. Diese wird von SVM (0.76) und MaxEnt (0.61) deutlich übertroffen und zeugt in beiden Fällen von eher zuverlässigen Prognosen. Jedoch zeigt sich bei Betrachtung des F1-Scores der negativen Kritiken (F1 neg) für das Logit Modell und der SVM eine schlechte Genauigkeit vorliegt. Im Falle der MaxEnt liegt eine mittelmässige Genauigkeit vor.
Desweiteren wird als Qualitätsmaß des Klassifikators die Fläche unter der ROC Kurve (AUC) betrachtet, für alle drei Modelle lässt sich basierend auf der AUC sagen, dass eine mittelmässig gute Leistung des Klassifikators zu beobachten ist.
Zusammenfassend kann behauptet werden, dass MaxEnt im Vergleich zu den anderen beiden Modellen am besten zur Klassifikation von Hotelbewertungen geeignet ist. Die Performance ist im Bereich der negativen Hotelbewertungen am besten.

Durch die getrennte Betrachtung der Leistung der Modelle innerhalb der positiven und negativen Kritiken ist festzustellen, das die richtige Klassifikation von negativen Hotelbewertungen als Herausforderung angesehen werden kann. Denn die Güte der Modelle wird durch die grundsätzlich gute Performance bei der Erkennung und richtigen Klassifikation der positiven Hotelbewertungen getragen. Betrachtet man die Leistung der Modelle bezogen auf die Klassifikation von negativen Bewertungen ist auffällig dass diese weit unterhalb der Qualität innerhalb der positiven Kritiken liegt. Neben der Ähnlichkeit des Wortlauts der positiven und negativen Bewertungen kann ein möglicher Grund der geringe Anteil von negativen Hotelbewertungen im Ausgangsdatensatz sein. Dadurch gelangt eine geringere Anzahl der negativen Hotelbewertungen in den Trainingsdatensatz. Hierdurch lernt der Klassifikator eher die vorliegenden Daten als positiv einzustufen.

Zusätzlich sei erwähnt dass für alle Modelle andere Settings wie $\lambda = 0.1$, oder eine quadratische Trennfunktion betrachtet wurden. Es konnte aber keine Verbesserung der Ergebnisse festgestellt werden.

Desweiteren hängen die Ergebnisse von den gewählten Daten, sowie den Ziehungen der Trainingsdatensätze ab, für andere Daten und andere Stichprobenziehungen können sich unterschiedliche Ergebnisse ergeben.

Größe des Trainingsdatensatzes

An dieser Stelle wird der Frage nachgegangen, wie sich die Leistung der Modelle hinsichtlich der Klassifikation von Hotelbewertungen mit steigender Anzahl an Tupeln im Trainingsdatensatz verändert.

Modell	Anzahl	GT	S	Sp	Pr neg	Pr pos	F1 neg	F1 pos	AUC
Logit	200	0.87	0.09	0.97	0.30	0.89	0.14	0.93	0.62
	300	0.87	0.13	0.97	0.34	0.90	0.19	0.93	0.66
	700	0.87	0.19	0.96	0.39	0.90	0.25	0.93	0.72
SVM	200	0.89	0.16	0.98	0.63	0.90	0.23	0.94	0.81
	300	0.89	0.17	0.99	0.71	0.90	0.24	0.94	0.84
	700	0.90	0.27	0.98	0.76	0.91	0.38	0.95	0.85
MaxEnt	200	0.89	0.29	0.97	0.56	0.91	0.37	0.94	0.73
	300	0.89	0.34	0.97	0.56	0.92	0.41	0.94	0.75
	700	0.91	0.47	0.96	0.63	0.93	0.53	0.95	0.79

Tabelle 9.4: Unterschiedliche Umfänge des Trainingsdatensatzes

Die Tabelle 9.4 zeigt die Ergebnisse der Analyseparameter unterschiedlicher Anzahlen von Trainingsdaten. Die Ergebnisse des Umfangs von 500 Tupel im Trainingsdatensatz sind den Tabellen 9.2 und 9.3 zu entnehmen.

Bei Betrachtung der Tabellen ist erkennbar dass sich die Performance innerhalb der positiven Bewertungen kaum ändert. Die Leistung der Klassifikatoren nimmt mit steigender Größe des Trainingsdatensatzes zu. Dies ist auf die verbesserte Leistung im Bereich der negativen Hotelbewertungen zurückzuführen. In allen drei Fällen steigt die Sensitivität, die Präzision (Pr_{neg}) und der F1 - Score ($F1_{neg}$).
Gut erkennbar ist hier ebenfalls dass die Leistung von MaxEnt innerhalb der negativen Kritiken immer am höchsten ist. Dies ist auf die Berücksichtigung der A Priori Wahrscheinlichkeiten der Daten zu beziehen, welche z.B. beim Logit Modell nicht berücksichtigt werden.
Die SVM sind hinsichtlich der AUC am besten einzustufen.

Beeinflußung der Stichprobenziehung

Der folgende Abschnitt behandelt die Frage ob sich die Performance der Modelle hinsichtlich der Treffer verbessern lässt. Dazu wird die Anzhal der negativen Kritiken im Trainingsdatensatz schrittweise erhöht.
Tabelle 9.5 zeigt die Ergebnisse der unterschiedlichen Anteile der negativen Bewertungen aus dem Originaldatensatz in Prozent, $a \in \{25, 33, 50, 66, 75\}$, die in den Trainingsdatensatz aufgenommen werden. Diese Anteile werden in der Spalte Anzahl als absolute Häufigkeiten angegeben. Im folgenden werden 300 Tupel in den Trainingsdatensatz aufgenommen.

Modell	Anzahl	GT	S	Sp	Pr neg	Pr pos	F1 neg	F1 pos	AUC
Logit	27	0.86	0.10	0.97	0.34	0.88	0.16	0.92	0.64
	36	0.87	0.13	0.96	0.30	0.90	0.18	0.92	0.65
	54	0.89	0.20	0.95	0.26	0.93	0.22	0.94	0.67
	72	0.90	0.23	0.94	0.19	0.95	0.20	0.95	0.69
	81	0.91	0.26	0.94	0.16	0.97	0.20	0.95	0.69
SVM	27	0.88	0.14	0.99	0.67	0.89	0.22	0.94	0.83
	36	0.90	0.18	0.99	0.68	0.91	0.26	0.89	0.83
	54	0.92	0.30	0.98	0.49	0.94	0.41	0.96	0.81
	72	0.93	0.36	0.97	0.42	0.96	0.35	0.96	0.79
	81	0.94	0.40	0.96	0.29	0.97	0.36	0.97	0.77
MaxEnt	27	0.89	0.30	0.98	0.67	0.91	0.41	0.94	0.75
	36	0.90	0.37	0.97	0.61	0.92	0.45	0.95	0.75
	54	0.91	0.46	0.96	0.49	0.95	0.47	0.95	0.74
	72	0.91	0.58	0.93	0.33	0.97	0.42	0.95	0.70
	81	0.91	0.59	0.93	0.26	0.98	0.36	0.95	0.68

Tabelle 9.5: Unterschiedliche Anteile der negativen Texte im Trainingsdatensatz

Bei allen drei Modellen ist mit steigender Anzahl an negativen Hotelbewertungen eine Verbesserung der Sensitivität um mehr als das Doppelte zu vermerken. Die Spezifität sinkt mit steigender Sensitivität, diese Abnahme ist nicht so drastisch einzustufen als die Verbesserung der Sensitivität. Diese Wechselwirkung ist auf den Fehler erster und zweiter Art zurückzuführen, hierbei kann eine Verbesserung des einen nur auf Kosten der Verschlechterung des anderen herbeigeführt werden, vgl. Tutz et al., 2007.
Die Präzision nimmt innerhalb der negativen Hotelbewertungen um fast die Hälfte ab. Die Präzision innerhalb der positiven Bewertungen steigt bei allen Modellen. Durch die Klassifikatoren werden mit steigender Anzahl an negativen Bewertungen im Trainingsdatensatz vermehrt negative Prognosen getroffen, die aber ungenauer werden.

Der F1 - Score der positiven Kritiken nimmt konstant zu. Hingegen steigt dieser Wert für die negativen Kritiken bis zu einer Anzahl von 54 negativen Kritiken an und nimmt dann wieder ab. Jedoch ist das Endniveau des Logit Modells und der SVM nicht so niedrig wie das Ausgangsniveau. Im Falle der MaxEnt liegt die Genauigkeit des Klassifikatiors unterhalb dem Ausgangswert. Dies ist auf die stärkere Abnahme der Präzision als die Zunahme der Sensitivität zurückzuführen.
Bei Betrachtung der AUC kann gesagt werden, dass die Leistung des Logit Modells als Klassifikator mit steigender Anzahl negativer Hotelbewertungen erhöht werden kann. Die Güte der beiden anderen Modelle als Klassifikatoren ist als stagnierend zu bewerten.

9.4 Modellierung mit Bigramen

In den vorhergehenden Abschnitten werden Modelle für Unigrame betrachtet. Unigrame besitzen den Nachteil, dass Wortkombinationen, wie beispielsweise 'nicht gut' nicht berücksichtigt werden. Um dies in die Modelle einfließen zu lassen werden Bigrame betrachtet. Basierend auf den Erkenntnissen der vorherigen Abschnitte wird eine Trainingsdatensatzgröße von 500 Beobachtungen verwendet. Im Falle der beeinflussten Stichprobenziehung werden 300 Beobachtungen in die Trainingsdaten aufgenommen.

Modell	GT	S	Sp	Pr neg	Pr pos	F1 neg	F1 pos	AUC
Logit	0.88	0.05	0.98	0.21	0.89	0.09	0.93	0.68
SVM	0.90	0.16	0.98	0.54	0.91	0.24	0.95	0.86
MaxEnt	0.89	0.24	0.97	0.48	0.91	0.31	0.94	0.76

Tabelle 9.6: Ergebnisse der Bigram Modelle

Durch die Aufnahme von Bigramen kann die GT und F1 pos des Logit Modells sowie der SVM leicht verbessert werden, bei fast gleichbleibenden Werten für die AUC. Innerhalb der positiven Kritiken verbessert sich die Leistung des Logit Modells und der SVM. Im Bereich der negativen Hotelkritiken ist eine Verschlechterung zu erkennen. Die Leistung der MaxEnt verschlechtert sich grundsätzlich durch die Betrachtung der Bigrame.

Die Ergebnisse der Änderung der Anteile der negativen Bewertungen im Trainingsdatensatz können Tabelle 9.7 entnommen werden.
Bei der Berücksichtigung von Bigramen kann die Gesamttrefferrate gesteigert werden. Die Spezifität nimmt ebenfalls mit steigender Sensitivität ab, jedoch

ist die Abnahme der Spezifität nicht so dramatisch wie die Verbesserung der Sensitivität.
Innerhalb der Klassifikation der positiven Hotelbewertungen nehmen die Werte der Präzision, des F1-Scores zu. Wodurch man zu dem Schluss kommt, dass im Falle des Vorliegens einer positiven Kritik die Güte des Klassifikators zunimmt.
Betrachtet man die Kennzahlen innerhalb der negativen Kritiken ist zu vermerken, dass die Präzision mit steigendem Anteil negativ annotierten Kritiken teilweise sehr stark abnimmt jedoch nimmt der F1-Score zu. Dieser Sachverhalt ist ebenfalls darauf zurückzuführen, dass die Zunahme der Sensitivität schneller als die Abnahme der Präzision stattfindet. Dies führt zu der Aussage dass die durchschnittliche Genauigkeit der Klassifikation besser wird.
Die Fläche unter der Kurve steigt einzig beim Logit Modell. Jedoch zeigt sich die Güte der Klassifikatoren in allen drei Fällen in einem mittelmäßig bis gutem Bereich.

Modell	Anzahl	GT	S	Sp	Pr neg	Pr pos	F1 neg	F1 pos	AUC
Zufall	300	0.89	0.08	0.98	0.30	0.90	0.14	0.94	0.64
	27	0.87	0.08	0.98	0.37	0.88	0.13	0.93	0.64
	36	0.88	0.13	0.97	0.36	0.90	0.19	0.93	0.65
Logit	54	0.90	0.17	0.96	0.29	0.93	0.22	0.94	0.69
	72	0.91	0.20	0.95	0.20	0.95	0.20	0.95	0.70
	81	0.91	0.23	0.94	0.15	0.97	0.18	0.95	0.69
Zufall	300	0.90	0.11	0.98	0.45	0.91	0.16	0.94	0.82
	27	0.88	0.11	0.99	0.64	0.89	0.20	0.94	0.85
	36	0.89	0.16	0.99	0.61	0.90	0.25	0.94	0.84
SVM	54	0.92	0.30	0.98	0.48	0.94	0.41	0.96	0.82
	72	0.93	0.37	0.97	0.41	0.96	0.35	0.96	0.80
	81	0.94	0.41	0.96	0.34	0.97	0.34	0.97	0.78
Zufall	300	0.89	0.16	0.98	0.47	0.91	0.23	0.94	0.71
	27	0.89	0.24	0.99	0.74	0.90	0.36	0.94	0.74
	36	0.91	0.32	0.98	0.68	0.92	0.43	0.95	0.75
MaxEnt	54	0.92	0.45	0.96	0.53	0.95	0.48	0.96	0.75
	72	0.92	0.57	0.94	0.37	0.97	0.44	0.96	0.71
	81	0.91	0.58	0.93	0.27	0.98	0.37	0.95	0.69

Tabelle 9.7: Ergebnisse der Bigram Modelle für bei Beeinflussung der Ziehung der Tupel des Trainingsdatensatzes

Basierend auf den Ergebnissen kann gesagt werden, dass die Qualität der Klassifikatoren durch die Betrachtung von Bigramen grundsätzlich gesteigert werden kann. Wenn auch eine geringe Verschlechterung der Performance innerhalb der Klassifikation von negativen Hotelbewertungen zu beobachten ist.

Kapitel 10

Algorithmusbasierte Klassifkationsverfahren

Bei den algorithmusbasierten Klassifikationsverfahren der folgenden Abschnitte wird im Gegenzug zu modellbasierten Klassifikationsverfahren die Ziehung mit Zurücklegen der Hotelbewertungen zugelassen.

10.1 Definition der Verfahren

Klassifikationsbäume

Klassifikationsbäume (KB) sind für kategoriale Zielvariablen mit endlich vielen Klassen definiert. Durch KB sollen Regeln geschaffen werden um, basierend auf den Prädiktorvariablen, Prognosen für die Zielvariable zu generieren. Die Regeln zur Prognose werden durch Rekursive Partitionierung des Datensatzes in möglichst homogene Regionen, bezüglich der Prädiktorvariablen geschaffen. Dabei wird der Datensatz solange partitioniert bis keine Verbesserung bezüglich eines bestimmten Optimimierungskriterium zu beobachten ist.
Ein KB wird nach folgendem Muster erstellt vgl. Cutler et al., 88.

- Nacheinander wird für jede Prädiktorvariable schrittweise die optimale binäre Trennung gesucht. Dies kann beispielsweise durch die Minimierung der Fehlklassifikationen (THAID), die Minimierung der Entropie der Fehlklassifikationen (C4.5) oder die Minimierung des Gini Index (CART) stattfinden. Jede Prädiktorvariable des Datensatzes wird durch einen Knoten repräsentiert, dieser gibt die optimale Trennung an.

- Nachdem für jede Prädiktorvariable die optimale Trennung gefunden wurde und jede Variable durch einen Knoten repräsentiert wird, ist die Partitionierung bezogen auf das Optimierungskriterium möglichst homogen.

- Der entstandene Baum wird als voll ausgewachsen bezeichnet.

- Zur Variablenselektion, das Trimmen, kann eine Kreuzvalidierung durchgefürt werden. Dabei wird der original Datensatz in $k = 1, \ldots, K$ gleich große Teildatensätze untereilt. $k-1$ Datensätze werden als Trainingsdatensatz zur Modellanpassung herangezogen. Der verbleibende Datensatz dient als Testdatensatz. Dieser Prozess wird k mal wiederholt bis jeder Datensatz einmal als Testdatensatz definiert wurde. Die Bewertung der Güte des Fits wird durch den Mean Square Error (MSE) gemessen. Es wird solange getrimmt bis keine Verbesserung zu beobachten ist.

Random Forests

Das Random Forests Verfahren definiert für den Datensatz eine beliebige Anzahl, z.B. 200, KB. Durch Kombination der Prognosen der Bäume für jede Beobachtung des Datensatzes wird dann das Ergebnis ausgegeben. Dies geschieht auf Basis des folgenden Algorithmus, vgl. Cutler et al., 88:

1. Generieren von $b = 1, \ldots, B$ Bootstrap Teildatensätze durch ziehen mit Zurücklegen. Beobachtungen die nicht im Trainingsdatensatz enthalten sind werden als Out of Bag Beobachtungen bezeichnet und dienen als Testdatensatz.

2. Für jeden Teildatensatz wird ein KB mit $m < M$ zufällig gewählten Prädiktoren definiert. Die Anzahl der zufällig gewählten Variablen kann z.B. durch $\sqrt{M}$ bestimmt werden. Die Generierung der Bäume geschieht nach dem oben beschriebenen Verfahren. Die Bäume werden voll ausgewachsen und es findet kein Pruning statt.

3. Jede Out of Bag Beobachtung wird durch jeden Baum klassifiziert. Die Prognose der jeweiligen Klasse entspricht dem häufigsten Ergebnis das durch den gesamten Wald ausgegeben wird.

Die Akkuratheit der Ergebnisse kann durch Cross Validation angegeben werden.

Bootstrap Aggregating

Dieses Verfahren wurde von Leo Breiman entwickelt, vgl. Breiman, 1996. Die Idee ist, die Prognoseergebnisse zu verbessern in dem nicht nur ein identischer Baum, sondern viele unterschiedliche Bäume zur Klassifikation verwendet werden.

Sei der Datensatz durch die Tupel $(y_i, x_i), i = 1, \ldots, N$ gegeben, mit M unabhängigen Beobachtungen. Wobei y_i dem Klassenlabel der Beobachtung i entspricht, mit Klassen $j = \{1, \ldots, J\}$. x_i repräsentiert die Beobachtungen für $m = 1, \ldots, M$ Prädiktorvariablen.
Ziel ist es basierend auf einem Trainingsdatensatz L einen Prädiktor $\rho(x, L)$ zu finden, der die Daten des Trainingsdatensatzes und Testdatensatzes T am besten beschreibt.
Die Prognose kann verbessert werden indem eine unterschiedliche Anzahl $\{L_k\}$ von $k = 1, \ldots, K$ Lernstichproben gezogen und für jede ein $\rho(x, L_k)$ geschätzt wird. Die Sequenz $\{\rho(x, L_k)\}$ wird als Schätzer der Klassenlabel betrachtet.
Folgender Algortihmus wird beim Bagging angewendet,

1. Der original Datensatz wird zufällig in einen Trainings- (L) und Testdatensatz (T) geteilt.

2. Basierend auf L wird ein KB mithilfe des CART Verfahrens entsprechend dem obigen Algorithmus generiert.

3. Aus dem Trainingsdatensatz wird eine Bootstrap Stichprobe L_B mit zurücklegen gezogen. Für diese Stichprobe wird ein Baum generiert und L als Testdatensatz verwendet um den getrimmten Teilbaum zu finden. Dieser Vorgang wird $t = 1, \ldots, T$ mal wiederholt, hierdurch werden $\phi_1(x), \ldots, \phi_T(x)$ Unterbäume generiert. Jeder dieser Bäume wird zur Prognose der jeweiligen Klasse der Beobachtung n verwendet.

4. Die prognostizierte Klasse für die Beobachtung i, ist diejenige für welche die größte Häufigkeit an Prognosen durch die Bäume beoabachtet wird.

- Dieser Vorgang kann beliebig oft wiederholt werden. Wobei bei jeder Iteration die gewählte Klasse notiert werden muss. Als Gesamtergebnis dient dann die Klasse die am häufigsten gewählt wurde.

Boosting

Boosting basiert auf der Idee durch die Kombination mehrerer schwacher Lernalgorithmen einen starken Lernalgorithmus mit höherer Treffsicherheit zu generieren. Schwache Lernalgorithmen können beispielsweise Entscheidungsbäume oder Regressionsmodelle sein, es wird angenommen dass sie nur wenig besser als ein Zufallsprozess sind. Dies geschieht durch eine unterschiedliche Gewichtung der Strategien. Die Strategien werden entsprechend der Präzision gewichtet, dabei fließen präzisere Schätzungen mit einem höheren Gewicht ein. Prinzipiell soll die Differenz zwischen dem Verlust der idealen Strategie und dem Verlust der durch die Allokation der Ressourcen auf die verschiedenen Strategien entsteht möglichst gering sein vgl. Freund and Schapire, 1997 und Schapire, 1990.

Formal lässt sich dies darstellen als,

$$L_A - \min_v L_v = \sum_{t=1}^{T} p_v^t l_v^t - \min_v \sum_{t=1}^{T} l_v^t, \tag{10.1}$$

wobei L_A dem kumulativen Verlust des Algorithmus A der ersten $t = 1, \ldots, T$ Iterationen entspricht. p_v^t entspricht der Allokation der Ressourcen der Strategie $v = 1, \ldots, V$ bei Iteration t, mit $p_v^t \geq 0$ und $\sum_{v=1}^{V} p_v^t = 1$. l_v^t ist der Verlust der durch Strategie v der Iteration t entsteht, mit $l_v \in [0, 1]$.
Hierfür wurde ursprünglich der sogenannte AdaBoost Algorithmus von Freund and Schapire, 1997 vorgestellt und kann zur Klassifikation der Hotelbewertungen ebenfalls verwendet werden.

Für die Klassifikation der Hotelbewertungen wurde der LogitBoost Algorithmus nach Friedman et al., 2000 verwendet. In RTexttools ist eine modifizierte Form der Funktion nach Dettling and Bühlmann, 2003 implementiert.

Der kombinierte Klassifikator ist gegeben durch,

$$C^{(m)}(X) = sign\left(\sum_{m=1}^{M} \alpha_m f^{(m)(X)}\right),$$

mit den aggregierten Gewichten α_m. Desweiteren sei ein Trainingsdatensatz mit $i = 1, \ldots, N$ Tupeln (x_i, y_i) gegeben. Sei f^m ein schwacher Lernalgortihmus der Iteration $m = 1, \ldots, M$. Im folgenden werden binäre Entscheidungsbäume mit zwei finalen Knoten verwendet.
Mit der Funktion $F^{(m)}(x) = \sum_{m=1}^{M} f_m(x)$ als Summe der einzelnen schwachen Lernalgorithmen $t = 1, \ldots, T$.

Die zu optimierende Verlustfunktion wird anders als bei AdaBoost durch die log-Likelihoodfunktion der Binomialverteilung definiert.

Der Algorithmus ist wie folgt definiert,

1. Inintialisierung:
   Definiere:
   - $F(x) \equiv 0$
   - $p^0(x) = 0.5$, mit $p(x) = \hat{P}(Y = 1|X = x)$
2. LogitBoost Iterationen

   (I) Berechnung des working response und der Gewichte für die Beobachtungen $i = 1, \ldots, N$

   $$w_i^{(m)} = p^{(m-1)}(x_i)\left(1 - p^{m-1}(x_i)\right),$$

   $$z_i^{(m)} = \frac{y_i - p^{(m-1)}(x_i)}{w_i^{(m)}}$$

   II Fitten eines Regressionsbaums durch die gewichtete kleinste Quadrate Methode

   $$f^{(m)} = \text{argmin}_f \sum_{i=1}^{n} w_i^{(m)}(z_i^{(m)} - f(x_i))^2$$

3. Aktualisieren des Klassifikators

   $$F^{(m)} = F^{(m-1)}(x_i) + 0.5 f^{(m)}(x_i)$$

   $$C^{(m)} = sign(F^{(m)}(x_i)),$$

   als Klassifikatorfunktion.

   $$p^{(m)}(x_i) = \left[1 + exp(-2F^{(m)}(x_i))\right]^{-1}$$

Diejenige Kategorie für welche die höchste Wahrscheinlichkeit angegeben wird entspricht der Kategorie die durch den Algorithmus vorgeschlagen wird.

Desweiteren wird von Dettling and Bühlmann, 2003 angegeben das die Funktion $F(x) = 0.5\log\left[\frac{p(x)}{1-p(x)}\right]$, zum besseren Verständnis als Schätzung der halbierten log Chancen angesehen werden kann. Ebenso dass der Algorithmus ohne weitere Feinabstimmung zu guten Ergebnissen kommt und meist schon nach 100 Iterationen das Optimum erreicht, die Anzahl der Iterationen wurde manuell auf 500 festgesetzt.

10.2 Anwendung

Zufallsziehung des Trainingsdatensatz

Der Trainingsdatensatz enthält in jeder wiederholten Ziehung 500 zufällig ausgewählte Beobachtungen. Innerhalb der wiederholten Ziehung wird dann jeder Algorithmus angewendet. Für den Algorithmus der Random Forests wurde der Vorgang des Schätzens ein, zehn, 100, 200 mal wiederholt. Die Ergebnisse der unterschiedlichen Anzahl an Iterationen sind der Tabelle 10.1 zu entnehmen. Es hat sich gezeigt, dass die Ergebnisse ab 10 Iterationen einer geringen Schwankung unterliegen, weswegen 10 wiederholte Ziehungen durchgeführt werden.

Iterationen	GT	S	Sp	Pr neg	Pr pos	F1 neg	F1 pos	AUC
1	0.87	0.08	0.97	0.25	0.90	0.13	0.93	0.71
10	0.86	0.18	0.95	0.32	0.90	0.23	0.92	0.69
100	0.86	0.20	0.95	0.34	0.90	0.24	0.92	0.70
200	0.86	0.19	0.94	0.34	0.90	0.23	0.92	0.70

Tabelle 10.1: Ergebnisse der unterschiedlichen Anzahlen von Iterationen der Random Forests

Verfahren	GT	S	Sp	Pr neg	Pr pos	F1 neg	F1 pos	AUC
Bagging	0.88	0.10	0.98	0.36	0.89	0.15	0.93	0.70
Boosting	0.87	0.49	0.92	0.44	0.93	0.46	0.92	0.76

Tabelle 10.2: Ergebnisse der zufälligen Ziehung von Beobachtungen für den Trainingsdatensatz der Algorithmusbasierten Verfahren

Tabelle 10.2 ziegt die Ergbenisse des Baggings und des Boostings.

Die GT der drei Modelle stellt sich als sehr ähnlich heraus. Auch ist grundsätzlich eine gute Performance der Verfahren zu beobachten. Die AUC

ist in allen drei Fällen mit einer mittelmässigen Anpassung zu bewerten, am besten schneidet hierbei das Verfahren des Boostings ab.
Im Bereich der Klassifikation der positiven Kritiken ist die Leitsung der drei Verfahren als gut einzustufen, wodurch ebenfalls mit ziemlicher Sicherheit bei vorliegen einer positiven Prognose diese mit der tatsächlichen Annotation übereinstimmt.
Betrachtet man die Leistung der Verfahren innerhalb der negativen Klasse, fällt auf, dass die Sensitivität sehr unterschiedlich ist. Im Falle des Baggings bietet diese einen Wert von 0.1 an, das Mittelfeld bildet RF, 0.18. Dies ist als schlecht einzustufen. Der höchste Wert ist beim Boosting mit 0.49 zu erkennen, hier kann von Mittelmaß gesprochen werden. Die Präzision und der F1-Score der RF und des Baggings sind als eher schlecht zu bewerten. Wohingegen die Werte des Boostings besser sind und sich wieder als mittelmäßig heraustellen.
Das Boosting erweist sich in diesem Fall als beste Methode, besonders im Bereich der negativen Kritiken.

Verglichen mit den Ergebnissen der modellbasierten Klassifikationsverfahren kann gesagt werden dass sich keine großen Unterschiede ergeben. Die RF weisen eine ähnliche Leistung wie das Logit Modell auf.

Da das Bagging als Spezialfall der RF gesehen werden kann und sich die Leistung der RF innerhalb der Klassifikation der negativen Hotelkritiken als besser heraustellt wird auf die Berücksichtigung des Baggings im Folgenden verzichtet.

Beeinflusste Stichprobenziehung

Ebenfalls wird in diesem Abschnitt der Effekt der Beeinflussung der Stichprobenziehung betrachtet. Tabelle 10.3 präsentiert die Ergebnisse.

Tree	GT	S	Sp	Pr neg	Pr pos	F1 neg	F1 pos	AUC
Zufall	0.85	0.24	0.93	0.32	0.90	0.26	0.92	0.67
27	0.85	0.13	0.96	0.31	0.89	0.18	0.92	0.65
36	0.86	0.19	0.95	0.31	0.90	0.23	0.92	0.69
54	0.85	0.38	0.90	0.25	0.94	0.29	0.92	0.70
72	0.83	0.46	0.85	0.16	0.96	0.23	0.90	0.66
81	0.82	0.50	0.83	0.12	0.97	0.19	0.90	0.64
Boosting								
Zufall	0.87	0.40	0.93	0.45	0.92	0.42	0.93	0.77
27	0.87	0.31	0.95	0.48	0.91	0.38	0.93	0.76
36	0.89	0.41	0.94	0.47	0.93	0.44	0.93	0.77
54	0.86	0.55	0.88	0.32	0.96	0.40	0.92	0.71
72	0.84	0.61	0.85	0.20	0.97	0.30	0.91	0.69
81	0.81	0.66	0.82	0.14	0.98	0.23	0.89	0.65

Tabelle 10.3: Sensitivität und Spezifität der Beeinflussten Stichprobenziehung

Zuerst werden zehn Tainingsdatensätze durch ziehen ohne zurücklegen generiert. Danach werden für die Random Forests je 200 Bäume verwendet und beim Boosting 500 Iterationen angewendet. Zum Vergleich sind die Kennzahlen der zufälligen Ziehung des Trainingsdatensatzes mit 300 Beobachtungen in der jeweils ersten Zeile aufgeführt.
Es ist zu erkennen, dass die Ergebnisse hierfür im Mittelfeld angesiedelt sind. Ebenfalls ist zu erkennen dass mit steigender Anzahl an negativen Hotelbewertungen im Trainingsdatensatz, die Sensitivität steigt und die Spezifität abnimmt.
Die Gesamttrefferrate und AUC nehmen mit steigendem Anteil an negativen Kritiken tendenziell ab, diese bewegen sich aber immer noch in einem Bereich der die Behauptung zulässt, dass eine mittelmäßig bis gute Eignung der beiden Verfahren vorliegt. Für die Kennzahlen die zur Bewertung der positiven Kritiken dienen ist eine Abnahme zu beobachten. Als einziges ist ein Anstieg der Präzision der positiven Kritiken zu vermerken. Jedoch kann hier immer noch von einer guten Leistung der beiden Verfahren zur Klassifikation von positiven Hotelbewertungen gesprochen werden.
Im Falle der Klassifikation der negativen Kritiken ist zu sehen, dass die Sentitivität gesteigert werden kann. Jedoch nimmt gleichzeitig die Präzision und der F1-Score ab. Diese Ergebnisse lassen den Schluss zu, dass die Sensitivität zwar verbessert werden kann, jedoch auch die Präzision gesenkt wird, was dazu führt, dass die Genauigkeit der Prognose abnimmt.
Durch dieses Vorgehen kann die Leistung der Klassifikatoren, im Vergleich zur Zufallsziehung gesteigert werden.

Das Boosting eignet sich aufgrund der Ergebnisse besser zur Klassifikation als RF.

Vergleicht man die Modelle der beiden Kapitel miteinander kann die Aussage getroffen werden, dass MaxEnt und Boosting, besonders durch die Leistung bei der Klassifikation von negativen Kritiken zu favorisieren sind.

10.3 Modellierung mit Bigramen

Im folgenden werden Bigrame statt Unigrame verwendet. Bei einem Vergleich der Tabellen 10.2 und 10.1 mit der untenstehenden Tabelle 10.4 fällt auf, dass sich die Performance durch Berücksichtigung der Bigrame der RF verbessert, jedoch verschlechtert sie sich im Falle des Boostings. Innerhalb der Klassifikation der negativen Kritiken sogar dramatisch. Dies ist auf die gesteigerte Menge an nicht informativen Variablen zurückführbar.

Es sei an dieser Stelle erwähnt, dass die Aufnahme hier im Gegensatz zu anderen Modellen auf Floskeln die mindestens viermal im Korpus beobachtet werden beschränkt wurde. Als Folge daraus ist eine Reduktion der DTM zu erkennen, dies musste vorgenommen werden, um RF schätzen zu können. Für das Boosting wurde dies in beiden Fällen vorgenommen, es stellte sich heraus, dass dies keine Änderung der gerundeten Ergebnisse bewirkt. Die letzte Zeile der Tabelle zeigt die Ergebnisse der unbeschränkten Aufnahme, (Boosting I).

Tree	GT	S	Sp	Pr neg	Pr pos	F1 neg	F1 pos	AUC
RF	0.88	0.23	0.96	0.39	0.96	0.28	0.93	0.66
Boosting	0.87	0.36	0.93	0.40	0.92	0.37	0.93	0.74
Boosting I	0.87	0.36	0.93	0.40	0.92	0.37	0.93	0.73

Tabelle 10.4: Zufallsziehung mit Bigramen

Die Ergebnisse der beeinflussten Stichprobenziehung sind der Tabelle 10.5 zu entnehmen.
Bei einem Vergleich mit Tabelle 10.3 ist zu sehen, dass keine deutliche Veränderung durch die beeinflusste Stichprobenziehung zu erkennen ist. Bei einem Vergleich der beiden Tabellen 10.3 und 10.5 ist keine deutliche Veränderung der Ergbenisse zu beobachten.

Modell	Anzahl	GT	S	Sp	Pr neg	Pr pos	F1 neg	F1 pos	AUC
Zufall	300	0.88	0.23	0.96	0.39	0.91	0.28	0.93	0.66
	27	0.85	0.14	0.96	0.31	0.88	0.18	0.92	0.65
	36	0.86	0.18	0.94	0.29	0.90	0.22	0.93	0.68
RF	54	0.85	0.37	0.90	0.25	0.94	0.30	0.92	0.70
	72	0.83	0.45	0.86	0.16	0.96	0.23	0.91	0.68
	81	0.82	0.49	0.84	0.12	0.97	0.19	0.90	0.65
Zufall	300	0.88	0.29	0.95	0.41	0.92	0.33	0.93	0.73
	27	0.87	0.29	0.96	0.50	0.90	0.36	0.93	0.77
	36	0.88	0.40	0.94	0.44	0.93	0.42	0.93	0.76
Boosting	54	0.86	0.56	0.88	0.31	0.96	0.39	0.92	0.70
	72	0.85	0.60	0.86	0.21	0.97	0.31	0.92	0.69
	81	0.79	0.68	0.79	0.13	0.98	0.21	0.87	0.63

Tabelle 10.5: Ergebnisse der Bigram Modelle für die Beeinflussung der Ziehung der Tupel des Trainingsdatensatzes

Kapitel 11

Wortvektoren

Neben der flexiblen Gestaltungsmöglichkeit innerhalb der neuronalen Netzwerke hinsichtich der Anzahl der vedeckten Schichten oder den Aktivierunsfunktionen ist es möglich Wortvektoren als Eingabe der NN zu verwenden. BOW Modelle mit einer *1-of-V* Kodierung weisen die Nachteile auf, dass die Position des Wortes innerhalb des Satzes nicht berücksichtigt wird. Die Position innerhalb des Satzes kann aber von essentieller Bedeutung für die Polarität eines Satzes sein. Sätze bestehend aus den selben Wörtern, aber unterschiedlichen Polaritäten weisen die gleiche Repräsentation auf. Desweiteren werden entweder Unigrame, Bigrame oder Trigrame, etc. mit einbezogen, die jedoch durch die Kodierung immer den selben Einfluß besitzen, z.B. gehen im Falle von Bigramen die Beobachtungen 'sehr gut' und 'war gut' in gleicherweise in die Modellierung ein.
Ein weiterer Aspekt ist, dass sich durch die Reduktion der Dimension der DTM, z.B. durch die Exklusion von Wörtern die nicht häufiger als fünf mal in einem Text vorkommen oder durch Extraktion von Stoppwörtern, Informationen verloren gehen. Beispielsweise werden bestimmte Kombinationen aus Wörtern oder Synoymen nicht mehr berücksichtigt, welche die selbe Bedeutung besitzen und auch die selbe Wirkung auf die Polarität einer Aussage aufweisen wenn diese nicht häufig genug im Korpus auftreten. Erweitert man durch die Betrachtung von weiteren N-Gramen, wie Trigramen, bläht man unter Umständen die Dimension unnötig auf, da selten Floskeln wie 'sehr sehr gut' zu beobachten sind.
Im Gegensatz zu BOW beziehen Wortvektoren in deren Darstellungsform der Wörter syntaktische und semantische Beziehungen ein. Zum Beispiel sollen die synthaktischen Beziehungen, a verhält sich zu b, wie c zu d ungefähr gleich sein. Das heisst der Abstand zwischen beispielsweise *gut* und *besser* soll dem Abstand von *schlecht* zu *schlechter* möglichst gleich sein. Weiter sollen semantische Beziehungen zwischen Wörtern durch $v(K\ddot{o}nig) - v(Mann) +$

$v(Frau) = v(Königin)$ ergeben, das Ergebnis sollte zumindest in der Nähe des Vektors $v(Königin)$ sein. Um diese Vorteile zu nutzen werden in neueren Methoden WV als Eingabe in die NN verwendet.
Vgl. Mikolov et al., 2013a, Mikolov et al., 2013b und Pennington et al., 2014.

11.1 Das Global Vectors Modell

WV könenn durch das Global Vectors Modell nach Pennington et al., 2014 erstellt werden, hierzu wird das gewichtete kleinste Quadrate Modell verwendet.

Die Basis des GloVe Modell bildet die Eingabematrix X mit den Einträgen X_{ij}, diese entsprechen der absoluten Häufigkeit des gemeinsamen Auftretens der Wörter i und j. Diese ist symmetrisch, die Zeilen und Spalten entsprechen den einzelnen Wörtern des Vokabulars. Die Summe $\sum_k X_{ik}$ entspricht der Anzahl wie oft ein Wort k im Kontext des Wortes i auftritt. Um innerhalb der Matrix X die Tatsache zu berücksichtigen, dass mit steigender Entfernung der Wörter immer weniger Einfluss aufeinander nehmen wird der Faktor $1/d$ eingefürt. d entspricht der Fenstergröße der betrachteten Kontextwörter und steigt mit grösser werdendem Abstand, z.B. entspricht der Wert $d = 2$ dem Gewicht der übernächsten Wörter des aktuellen Wortes.
Die bedingte Wahrscheinlichkeit dass das Wort j im Kontext von Wort i erscheint ist gegeben durch $P_{ij} = P(j|i) = X_{ij}/X_i$.
Der Ausgangpunkt des Modells ist die Relation,

$$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}. \tag{11.1}$$

Dabei repräsentieren $w_i, w_j, \tilde{w}_k \in \mathbb{R}^d$, der Wörter i, j und des Testwortes k. d entspricht der Dimension der Fenstergröße der betrachteten Kontextwörter. Die Darstellung der Beziehungen zwischen den Wörtern i, j, k wird von den Autoren anhand dem folgenden Beispiel motiviert.
Seien die Zielwörter $i = Eis$ und $j = Dampf$ gegeben, dann sollte die Relation durch die Hinzunahme des Wortes $k = stabil$ groß sein, wohingegen die Relation durch die Hinzunahme des Wortes $k = gasförmig$ klein sein soll. Große Werte sprechen demnach für eine größere Beziehung des Wortes Eis zu dem Wort $stabil$. Kleine Werte sprechen für eine größere Beziehung des Wortes $Dampf$ zu dem Wort $gasförmig$. Für Wörter wie $k = Wasser$ welche gleichzeitig einen Bezug zu beiden Zielwörtern repräsentieren oder $k = Mode$ welche keinen Bezug zu beiden Wörtern aufweisen sollte die Relation nahe eins sein.

Das gewichtete kleinste Quadrate Modell welches zur Schätzung der unbekannten Funktion $F()$ verwendet wird lässt sich wie folgt darstellen,

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \tilde{w}_k + b_i + \tilde{b}_k - log(X_{ij}))^2, \tag{11.2}$$

wobei V dem Umfang des Vokabularsentspricht. $w_i, \tilde{w}_j$ entsprechen den Wortvektoren, diese entsprechen der Zeile bzw. Spalte der Matrix X. $b_i, \tilde{b}_k$ entsprechen dem Interecept. $log()$ dient dazu, dass sehr häufige Beobachtungen keinen zu großen Einfluss im Vergleich, zu weniger häufigen Beobachtungen besitzen. Die stetige Gewichtsfunktion $f(X_{ij})$ soll folgende Anfoderungen erfüllen 1. $f(0) = 0$. 2. $f(x)$ nicht fallend, so dass seltenen Beobachtungen keine zu großen Gewichte zugeordnet werden. 3. $f(x)$ sollte für große Werte von x keine zu großen Gewichte definieren.
Für $f(x)$ wird folgendes vorgeschlagen,

$$f(x) = \begin{cases} (x/x_{max})^\alpha & \text{falls } x < x_{max} \\ 1 & \text{sonst} \end{cases} . \tag{11.3}$$

Nachfolgend werden zur Generierung der WV die Parameter entsprechend der Arbeit von Pennington et al., 2014 gewählt. $\alpha = 3/4$ und $x_{max} = 100$. Zur Optimierung wird das AdaGrad verfahren nach Duchi et al., 2011 verwendet, mit Startlernrate $\gamma = 0.05$. Für die WV wird die Dimension von 50 definiert und 100 Iterationen zugelassen. Wörter die mindestens vier mal im Korpus beobachtet werden, werden zur Schätzung von WV verwendet. Die Fenstergröße wird mit 21 Wörtern festgesetzt.

11.2 Erstellung der Wortvektoren

Pennington et al., 2014 zeigen in Ihrer Arbeit, dass sich die Leistung des GloVe Modells durch ein Anstieg des Vokabulars verbessert. Aus diesem Grund wird zusätzlich zum vorliegenden Korpus der Hotelkritiken ein Korpus aus der Leipziger Korpus Sammlung der Universität Leipzig verwendet, vgl. Goldhahn et al., 2012. Die beiden Korpora werden miteindander kombiniert und zur Schätzung von WV verwendet. Der Datensatz der Universität Leipzig umfasst 2791246 Sätze und besteht aus zufällig zusammengetragenen Sätzen aus Nachrichten oder dem Internet. Unvollständige Sätze, sowie Inhalte einer anderen Sprache werden entfernt. Der Datensatz aus dem Jahre 2015 kann unter dem Link: `http://wortschatz.uni-leipzig.de/en/download` heruntergeladen werden.

Eine weitere Möglichkeit bieten Datensätze wie eine Momentaufnahme des Internetlexikons Wikipedia, denkbar sind auch Bücher als Korpus.
Vor der Schätzung der WV werden Satzzeichen und Zahlen aus dem Korpus entfernt. Auch wird die Großschreibung elimieniert, so dass alle Wörter kleingeschrieben im Text erscheinen. Um zu gewährleisten, dass soviele Wörter wie möglich im Korpus enthalten sind wurde auf eine Extraktion der Stoppwörter verzichtet. Es ergaben sich 254915 verschiedene Wortvektoren.

Die beiden nachfolgenden Grafiken 11.1 und 11.2 zeigen die Schlagwörter 'Montag' und 'Vater' sowie deren sechs bzw. 28 Wörter welche die geringste Distanz zu dem jeweiligen Schlagwort innerhalb der gesamten WV aufweisen. Die Distanz zwischen den einzelnen WV wird die Kosinus Ähnlichkeit berechnet. Zur grafischen Darstellung werden jeweils die ersten beiden Dimensionen der WV verwendet.

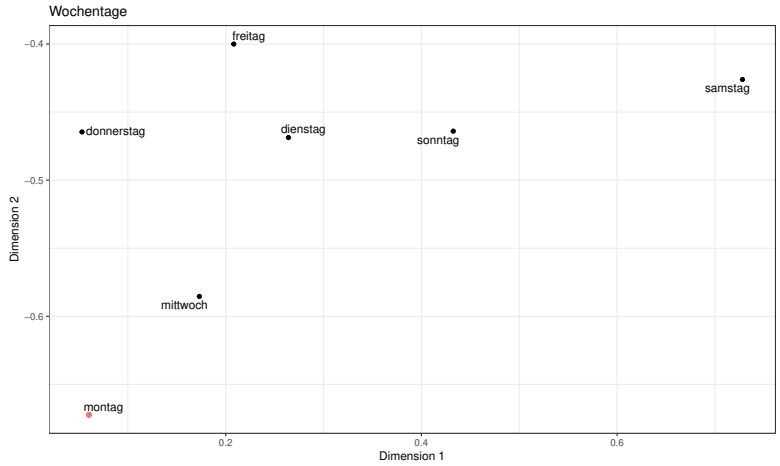

Abbildung 11.1: Umgebung des Wortes Montag

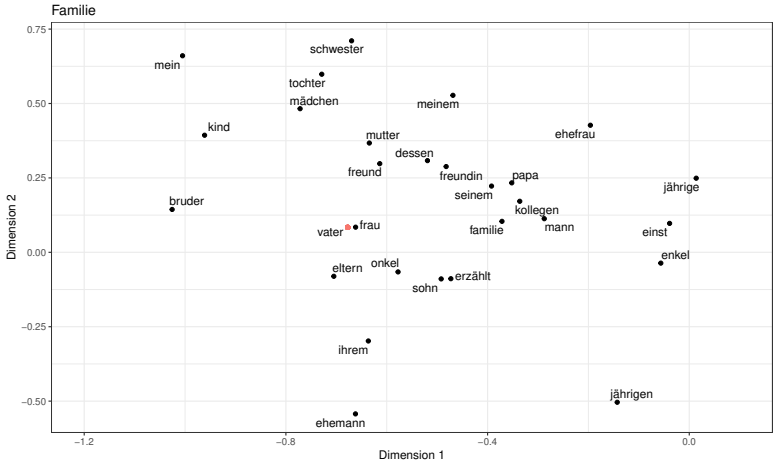

Abbildung 11.2: Umgebung des Wortes Vater

In Grafik 11.1 werden die Wochentage alls nächstgelegene Wörter identifiziert.
Abbildung 11.2 zeigt fast nur Wörter die in Bezug zu Familie stehen, wie 'Tochter', 'Mutter', oder 'Ehefrau'. Ebenfalls erkennbar sind Wörter die das soziale Umfeld beschrieben, z.B. 'Freund', oder 'Kollegen'. Dies veranschaulicht, dass durch Wortvektoren Beziehungen zwischen den einzelnen Begriffen erlernt werden können.

Grafik 11.3 bildet die 40 Wörter der nächsten Umgebung des Wortes 'Frühstücksbuffet' ab.

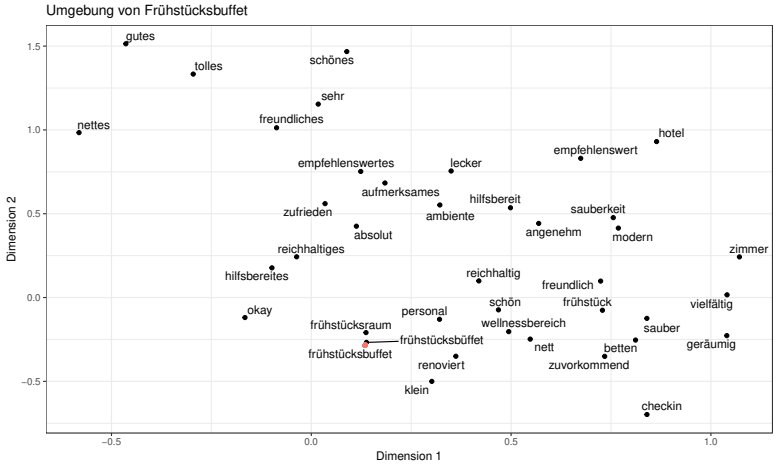

Abbildung 11.3: Umgebung des Wortes Frühstücksbuffet

Erkennbar ist hier, die Nähe von 'Frühstücksbuffet' und dessen alternativer Schreibweise 'Frühstücksbüffet', sowie dem Begriff 'Frühstücksraum'. Desweiteren zeigt sich, dass sich Sinnesabhängige Abschnitte formieren. So erscheinen im näheren Umfeld des Wortes 'Zimmer' Wörter die intuitiv mit diesem in Verbinung gebracht werden, z.B. 'Betten', oder 'geräumig'. Die letztgenannten drei Wörter weisen ebenfalls eine weitere Distanz zum 'Frühstücksbuffet' auf als sie untereinander aufweisen.

Greift man den Punkt nochmals auf, dass die WV basierend auf der Matrix des gemeinsamen Paarweisen auftretens der Wörter gebildet werden, so ist zu erwähnen, dass die Darstellung der Beziehungen zwischen den Wörtern stark von der Fenstergröße d abhängt. Kombinationen außerhalb des Fensters werden nicht berücksichtigt und somit nicht in die Modellierung aufgenommen. So werden die Wörter 'Personal' und deren Kontextwörter häufiger gemeinsam oder 'näher' in Kombination mit 'Frühstücksbuffet erwähnt, als z.B. die Kombinationan aus 'Personal' und 'Zimmer'. Hierdurch kann die Überlagerung der einzelnen Aspekte erklärt werden.

11.3 Weitere Verfahren zur Schätzung von Wortvektoren

Von Mikolov et al., 2013a und Mikolov et al., 2013b wird zur Generierung der WV das Continuous Bag-of-Words Modell (CBOW) und das Continuous Skip-gram Modell (SKG) vorgeschlagen. Diese beiden Modelle verwenden lokale Fenster zur Schätzung von WV. Dabei wird jeweils ein neuronales Netzwerk mit einer verdeckten Schicht herangezogen. Beim CBOW werden N umliegende Fensterwörter verwendet um den WV des aktuell betrachteten Wortes zu generieren. Beim SKG werden basierend auf dem aktuell betrachteten Wort die WV der N Fensterwörter prognostiziert.
Ein weiteres Verfahren ist das Latent Semantic Analysis (LSA) Verfahren nach Deerwester et al., 1990, dieses Verfahren nutzt die globale Matrix Faktorisierung zum erlernen von WV.
Auf diese Verfahren wird im Rahmen dieser Arbeit nicht näher eingegangen, für einen tieferen Einblick sei hier auf Mikolov et al., 2013a und Mikolov et al., 2013b, sowie Deerwester et al., 1990 verwiesen.

Pennington et al., 2014 stellen zum einen fest, dass bei einem Vergleich der verschiedenen Ansätze keine großen Unterschiede beobachtbar sind, wenn die

Verfaren das gemeinsame Auftreten von Wörtern als Basis verwenden. Jedoch ergibt sich durch die globale Betrachtung des GloVE Modells, im Vergleich zur lokalen Betrachtung des SKG Modells und des CBOW Modells, eine bessere Leistung der Darstellung der Zusammenhänge. Beispielsweise ergibt sich für das GloVe Modell im Vergleich zum SKG und CBOW, bei gleicher Trainingszeit eine höhere Akkuratheit bei der Modellierung der Zusammenhänge. Zum anderen

Im folgenden wird eine Möglichkeit um die unterschiedlichen WV, verschiedener Modelle, zu kombinieren vorgeschlagen.
Basierend auf identischen Korpora und den selben Eingabematrizen werden die WV durch die unterschiedlichen Modelle geschätzt. Die finalen Prognosen der jeweiligen Wörter könnten in eine Matrix, die jeweils ein Wort repräsentiert überführt werden. Durch z.b. eine Matrix Faktoriesierung könnte somit ein einzelner Vektor für ein Wort generiert werden.

Die Methoden der folgenden Kapitel greifen auf WV als Eingabewerte zurück.

Kapitel 12

Neuronale Netzwerke

12.1 Definition Neuronaler Netzwerke

Neuronale Netzwerke (NN) dienen der Approximaton einer unbekannten nicht linearen Funktion, $f(X)$. Dabei werden die Eingabewerte durch die Schichten des Netzwerks geschleust um die Zeilvariable zu prognostizieren. Jede Schicht besteht aus einer Ansammlung von Neuronen, welche durch linear Kombinationen miteinander Verbunden sind. Die Features einer nachgelagerten Schicht werden mithilfe einer Aktivierungsfunktion aktiviert.
NN verfolgen die Intention durch iterative Justierung der Gewichte des Netzwerks, die Abweichung zwischen den Prognosen und den tatsächlichen Beobachtungen zu minimieren, vgl. Hastie et al., 2008.

Zur Illustration eines neuronalen Netzwerks mit einer verdeckten Schicht und einer beliebigen Aktivierungsfunktion dient Grafik 12.1, diese Grafik wurde Hastie et al., 2008 entnommen.

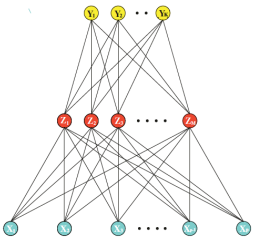

Abbildung 12.1: Illustration eines Neuronalen Netzwerks mit einer verdeckten Schicht

Im folgenden wird ungeachtet der Notation in dieser Arbeit die Notation der jeweiligen betrachteten Arbeiten aufgegriffen.

12.2 Ein einfaches Neuronales Netzwerk

Zur allgemeinen Erklärung dient das '*vanilla neural net*', auch bekannt als '*single hidden layer backpropagation network*' oder '*single layer perceptron*', vgl. Hastie et al., 2008.

Hierfür sei der Inputvektor $X \in \mathbb{R}^m$, die Modellparameter bzw. die Gewichte $\alpha = (\alpha_0, \alpha_1, \dots \alpha_m)$ mit $m = 1, \dots, M$ und $\beta = (\beta_0, \beta_1, \dots, \beta_k)$ mit $k = 1, \dots, K$ gegeben. Zusätzlich sind die tatsächlichen Werte Y_k bekannt. Im Falle der Regression entspricht $k = 1$. Im Falle der Klassifikation entspricht $k = 1, \dots, K$ der Warscheinlichkeit der k-ten Klasse.

Der '*Forward Pass*'

Der '*Forward Pass*' dient der Berechnung der Ausgabewerte des NN und lässt sich wie folgt darstellen.

$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X), \;\; m = 1, \dots, M,$$

$$T_k = \beta_{0k} + \beta_k^T Z, \;\; k = 1, \dots, K, \tag{12.1}$$

$$f_k(X) = g_k(T).$$

Mit

$$g_k(T) = \frac{exp\{T_k\}}{\sum_{l=1}^{K} exp\{T_l\}}$$

Die Gewichte $w = (\alpha, \beta)$ werden für die erste Iteration in der Regel zufällig generiert, z.B. durch Ziehung aus der Gleichverteilung. Eine Schätzung durch die kleinste Quadrate Methode oder Maximum Likelihood ist möglich, wenn die Softmax Aktivierungsfunktion und Kreuzentropie als Verlustfunktion herangezogen wird, vgl. Hastie et al., 2008.
Die Features Z_m der verdeckten Schicht werden mithilfe einer Funktion $\sigma()$ der linear Kombinationen aktiviert. Als Aktivierungsfunktion fungiert die *sigmoid* Funktion mit $\sigma(v) = 1/(1 + exp(-v))$ und $\sigma(v) \in [0, 1]$. Sämtliche

Features werden zu einem Vektor $Z = (Z_1, \ldots, Z_m)$ zusammengefasst und stellen die verdeckte Schicht dar. Die Netzwerkarchitektur kann mit selbigem Vorgehen auf eine beliebige Anzahl an verdeckten Schichten ausgeweitet werden. Als Aktivierungsfunktion können ebenfalls andere Funktionen wie die tangens hyperbolicus Funktion herangezogen werden. Im einfachsten Fall entspricht diese der Identitätsfunktion und kann als lineare Regression angesehen werden.

In der letzten Schicht werden die Prognosen des Netzwerks durch die Softmax Funktion $g_k(T)$ aktiviert, mit den linear Kombinationen T, $T_k = (T_1, \ldots, T_K)$. Zur Prognose der Ausgabewerte des Netzwerks dient die Softmax Funktion, das Ergebnis entspricht der Wahrscheinlichkeitsverteilung über die k Klassen der Zielvariablen Y_k. Als Prognose des Netzwerks wird die Klasse mit der größten Wahrscheinlichkeit gewählt.

Die sehr flexible Netzwerkarchitektur kann mit auf eine beliebige Anzahl an verdeckten Schichten ausgeweitet werden. Verschiedene Schichten können unterschiedliche Aktivierungsfunktionen besitzen.

Der '*backward Pass*'

Der Einstieg in den '*backward Pass*' erfolgt durch die Berechnung der Kostenfunktion, oder der Verlustfunktion. In diesem Fall wird der quadrierte Fehler $R(\Theta)$, wobei Θ den Parametern des Netzwerks entspricht herangezogen. Bei der Klassifikation ist auch die Wahl der Kreuzentropie als Verlustfunktion möglich.

Der Fehler des '*vanilla neural net*' nach Hastie et al., 2008 ist definiert durch,

$$R(\Theta) = \sum_{i=1}^{N} R_i = \sum_{i=1}^{N} \sum_{k=1}^{K} (y_{ik} - f_k(x_i))^2. \tag{12.2}$$

Die Minimierung von $R(\Theta)$ erfolgt durch die '*back propagation*', dazu dienen Verfahren wie das Gradientenverfahren. Der Gradient kann durch die Kettenregel hergeleitet werden. Die Ableitungen sind darstellbar als,

$$\frac{\partial R_i}{\partial \beta_{km}} = -2\{y_{ik} - f_k(x_i)\} g_k'(\beta_k z_i) z_{mi}, \tag{12.3}$$

$$\frac{\partial R_i}{\partial \alpha_{ml}} = -\sum_{k=1}^{K} 2(y_{ik} - f_k(x_i)) g_k'(\beta_k^T z_i) \beta_{km} \sigma'(\alpha_m^T x_i) x_{il}. \tag{12.4}$$

Mit diesen Ableitungen lassen sich die Gewichte für die $(r+1)$ Iteration aktualisieren,

$$\beta_{km}^{(r+1)} = \beta_{km}^{(r)} - \gamma_r \sum_{i=1}^{N} \frac{\partial R_i}{\partial \beta_{km}^{(r)}} \tag{12.5}$$

und

$$\alpha_{ml}^{(r+1)} = \alpha_{ml}^{(r)} - \gamma_r \frac{\partial R_i}{\partial \alpha_{ml}^{(r)}}. \tag{12.6}$$

γ_r entspricht der Lernrate. Diese kann als konstant gewählt werden, jedoch kann diese auch für jede Epoche (Iteration) optimal gewählt werden, vgl. Hastie et al., 2008.

Die Größen

$$\frac{\partial R_i}{\partial \beta_{km}} = \delta_{ki} z_{mi} \tag{12.7}$$

und

$$\frac{\partial R_i}{\partial \alpha_{ml}} = s_{mi} x_{il}, \tag{12.8}$$

mit

$$s_{mi} = \sigma'(\alpha_m^T x_i) \sum_{k=1}^{K} \beta_{km} \delta_{ki},$$

δ_{ki} und s_{mi} können als Fehler der aktuellen Iteration Interpretiert werden.

Ausgehend von den Fehlern werden die Gewichte rekursiv angepasst. Dieser Vorgang wird solange wiederholt bis ein Stoppkriterium erreicht ist, oder die tatsächlichen Beobachtungen mit den Prognosen des Netzwerks übereinstimmen.

12.3 Komplexität von Neuronalen Netzwerken

Die Komplexität von NN steigt mit dem Umfang der Netzwerkarchitektur. Betrachtet man ein Wort als Eingabe, $x \in \mathbb{R}^{d \times 1}$ und Gewichte, $w \in \mathbb{R}^{d \times 1}$ so ergibt sich durch die linear Kombination $w^t x$ ein Skalar als Feature Z.

Erhöht man die Anzahl der Wörter ergibt sich eine Eingabe der Dimension $X \in \mathbb{R}^{n \times d}$ dann entsteht ein Feature Vektor der Dimension $Z \in \mathbb{R}^{n \times 1}$. Wird zusätzlich die Anzahl der Einheiten der verdeckten Schicht um p erhöht ergibt sich eine Feature Matrix $Z \in \mathbb{R}^{n \times p}$. Dieses Beispiel verdeutlicht den raschen Anstieg der Komplexität von NN, vgl. Mikolov et al., 2013a.
Erhöht man weiter die Anzahl der verdeckten Schichten, bei gleicher Struktur, wird dieser Rechenaufwand ebenfalls erhöht. Dem kann entgegen gewirkt werden in dem z.B. die Dimension der Eingabevektoren bzw. die Dimension der Gewichtsvektoren verringert wird. Die Verkleinerung des Umfangs des Vokabulars, wie die Extaktion der Stoppwörter bewirkt ebenfalls eine Reduktion der Komplexität. Durch diesen Vorgang werden vor der Schätzung wenig Informative Wörter entfernt.
Ebenfalls besitzt die Wahl der Aktivierungsfunktion einen Einfluß. Wird wie im Beispiel oben, die Identitätsfunktion als Aktivierungsfunktion gewählt entsteht kein weiterer Rechenaufwand. Durch die Wahl anderer Aktivierungsfunktionen wie z.B. der Logistischen Funktion ist jeweils ein weiterer Rechenschritt nötig. Aktivierungsfunktionen können aber auch den Effekt besitzen, dass sie eine Reduktion der Dimensionen bewirken. Nimmt man die z.B. die Softmax Funktion an generiert diese einen k Dimensionalen Vektor entsprechend der Dimension der Zielvariablen, unabhängig von der Anzahl und Dimension des Funktionsarguments. Im Falle dass die Dimension der verdeckten Schicht größer als die Dimension des k dimensionalen Ausgabevektors ist entsteht eine Verringerung der Dimension. Ein lineares Regressionsmodell generiert einen Skalar als Ausgabe, ebenfalls unabhängig von der Dimension des Arguments.
Die Wahl des Lösungsverfahrens zur Optimierung der Verlustfunktion, bzw. des Prognosefehlers, ist für die Rechenzeit des Netzwerks mitverantwortlich. Je nach Optimierungsverfahren oder der Wahl der Hyperparamter, kann sich die Zahl der Iterationen, bzw. Epochen bis zur Konvergenz erhöhen oder verringern. An dieser Stelle sei angemerkt, dass bei allen Optimierungsverfahren immer das Problem auftreten kann, dass nur ein lokales Optimum gefunden wird. Diesem Problem wird begegnet indem, zur Optimierung verschiedene Startwerte gesetzt werden und die Ergebnisse anschliessend verglichen werden sollten. Daher sollte jedes Netzwerk wiederholt geschätzt werden. Auch die Wahl der Lernraten besitzt einen Einfluss, kleine Lernraten erfordern mehr Iterationen zur Konvergenz und können in lokalen Optimas hängen bleiben. Zu hohe Lernraten bergen die Gefahr Optimas zu Überspringen.
Aus diesen oben aufgeführten Punkten sollte immer die Wechselwirkung zwischen Verbesserung der Prognose durch eine veränderte und / oder umfangreichere Modell Architektur des Netzwerks beachtet werden und verglichen werden ob ein weniger komplexes NN ähnlich gute Ergebnisse liefert.

Durch die Komplexe Struktur der NN neigen diese zur Überanpassung. Dies kann durch eine Wahl zu vieler verdeckten Einheiten oder zu vieler verdeckten Schichten begünstigt werden. Neben anderen Verfahren die zur Regularisierung von NN herangezogen werden können, wird von Srivastava et al., 2014 das Dropout Verfahren vorgeschlagen. Dabei werden Einheiten des NN zufällig während des Trainings weggelassen. Während der Trainingsphase werden für jeden neuen Fall zufällig Einheiten herausgelassen und ein 'neues' NN angepasst. Im Verlauf der der Testphase wird das 'vollbestzte' NN herangezogen und skalierte approximative durchschnittliche Gewichte verwendet, vgl. Srivastava et al., 2014. Auf dieses Verfahren wird im späteren Verlauf der Arbeit nocheinmal eingegangen.

12.4 Neuronales Netzwerk mit Bag of Words

Um die neuronalen Netzwerke auf die Thematik der Klassifikation von Hotelbewertungen anzuwenden wird die Eingabematrix, für Unigrame, des neuronalen Netzwerks mit einer verdeckten Schicht entsprechend den vorherigen Kapiteln gewählt. Als Aktivierungsfunktion dient die Logistische Funktion, sowie die Softmax Funktion. Als Verlustfunktion wird der quadrierte Fehler verwendet, zur Optimierung dient das Gradientenverfahren, die Lernrate ist mit 0.05 definiert.

Die Ergebnisse können der Tabelle 12.1 entnommen werden. Die ersten beiden Zeilen der Tabelle zeigen die Resultate der zufälligen Stichprobenziehung von 300 bzw. 500 Tupeln, die restlichen Zeilen der Tabelle enthalten die Ergebnisse der beeinflussten Ziehung von negativen und positiven Kritiken die in den Trainingsdatensatz aufgenommen werden. Hierbei umfasst der Trainingsdatensatz 300 Kritiken. Die absoluten Anzahlen der negativen Kritiken die in den Trainingsdatensatz aufgenommen werden sind ab der dritten Zeile angegeben.

	GT	S	Sp	Pr neg	Pr pos	F1 neg	F1 pos	AUC
300	0.86	0.32	0.94	0.42	0.91	0.34	0.92	0.63
500	0.88	0.41	0.94	0.47	0.92	0.43	0.93	0.68
27	0.87	0.25	0.96	0.44	0.90	0.35	0.93	0.61
36	0.87	0.34	0.94	0.40	0.92	0.36	0.93	0.64
54	0.88	0.45	0.91	0.33	0.95	0.37	0.93	0.68
72	0.88	0.57	0.89	0.25	0.97	0.34	0.93	0.73
81	0.86	0.53	0.88	0.16	0.98	0.25	0.92	0.70
2	0.87	0.40	0.93	0.43	0.92	0.40	0.93	0.66

Tabelle 12.1: Ergebnisse des neuronalen Netzwerke

Bei Betrachtung der Tabelle ist erkennbar dass sich die Kennzahlen des NN ähnlich zu den Interpretationen der vorherigen Methoden verhalten.
Im Falle der zufälligen Generierung der Trainingsdaten zeigt sich die Gesamttrefferrate (0.86; 0.88). Diese verhält sich ähnlich zu den Werten der vorherigen Methoden und kann als gut eingestuft werden. Bei Betrachtung der Kennzahlen innerhalb der negativen Kritiken ist festzustellen dass die Sensitivität einen mittelmäßigen Wert annimmt. Die Präzision spiegelt wieder, dass nur fast jede zweite Prognose mit der tatsächlichen Polarität der Kritik übereinstimmt. Der F1 Score spricht für eine mittlere Messgenauigkeit der Methode.
Die Kennzahlen der Prognose der positiven Kritiken lassen den Schluss zu, dass diese relativ zuverlässig Klassifiziert werden. Die Werte der AUC sprechen ebenfalls gegen ein vorherrschen eines Zufallsprozesses.
Bei einem Vergleich der beeinflussten Stichprobenziehung ist zu sehen, dass die Gesamttrefferrate ebenfalls gute Werte aufweist, dies entspricht den Resultaten der vorhergehenden Methoden. Die Sensitivität steigt mit zunehmenden Verhältnis von negativen zu positiven Kritiken. Diese fällt jedoch ab 72 negativen Kritiken ab, wobei der Wert immer noch mehr als doppelt so hoch als bei 27 negativen Kritiken ist. Die Präzision (Pr neg) und der F1 Score (F1 neg) nehmen mit der Zunahme an negativen Kritiken in den Trainingsdaten ab. Dies weist darauf hin dass bei einer höheren Relation zwischen negativen und positiven Hotelbewertungen die Zuverlässigkeit bei der Klassifikation von negativen Hotelbewertungen abnimmt. Die Werte der AUC nehmen im Verlauf der Steigerung der Anzahl an negativen Kritiken zu. In allen Fällen sprechen diese ebenfalls gegen einen Zufallsprozess.
Die Kennzahlen der Klassifikation von positiven Kritiken können ebenfalls als gut eingestuft werden. Dies lässt auch im Falle der beeinflussten Stichprobenziehung die Aussage zu, dass dieses NN einen guten Klassifikator darstellt.
Erkennbar ist jeoch dass sich trotz der Verbesserung der GT durch die beeinflusste Stichproben Ziehung die Sensitivität, Spezifität, die präzision der

negativen Prognosen, sowie der F1 Score der negativen Kritiken im Vergleich zu zufälligen Stichprobenziehung verschlechtern. Dies führt zu der Aussage, dass durch die Beeinflussung der Stichprobenziehung keine Verbesserung des Modells erreicht werden kann.

Stellt man erneut einen Vergleich mit den Methoden der vorherigen Kaptiel an, so ist der Schluss zulässig dass dieses NN mit einer einfachen Architektur, auch im Falle der beeinflussten Stichprobenziehung, die besseren Ergebnisse liefert: Jedoch sind die Methoden MaxEnt und Boosting hervorzuheben die bei der Klassifikation von Kritiken der Leistung des NN ähneln. Das Boosting übertrifft diese sogar. Desweiteren steigt die AUC mit steigender Anzahl an negativen Bewertungen und entfernt sich immer mehr dem Zufallsprozesses. Bei den Methoden der vorherigen Kapitel sinkt der Wert der AUC und nähert sich einem Zufallsprozess immer mehr an.
Da sich keine deutliche Verbesserung des NN, insbesondere bei der Klassifikation der negativen Bewertungen, durch die beeinflusste Stichprobenziehung ergab, wird im Folgenden auf dieses Verfahren verzichtet.

Zusätzlich wurde untersucht ob sich die Leistung des NN durch geänderte Hyperparmeter verändert. Dies wurde vorgenommen indem jeweis ein Hyperparameter geändert wurde und die anderen festgehalten wurden. Betrachtet wurden zum einen die Lernraten $\gamma = (0.01, 0.1)$ und zum anderen die Anzahl der verdeckten Schichten $\nu = (1, 2, 3, 4, 5)$. Als Referenz diente das Modell mit einer zufälligen Auswahl von 500 Tupeln. Das NN mit zwei verdeckten Schichten ist zu erwähnen. Die Ergebnisse liegen jedoch etwas unterhalb des NN mit einer verdeckten Schicht und können der letzten Zeile der Tabelle 12.1 entnommen werden. Für die restlichen Änderungen ergab sich eine Verschlechterung der Leistung des NN. Dies lässt den Schluß zu, dass dieses NN bereits einen guten Klassifikator definiert und eine Erhöhung der Komplexität des NN unnötig wäre.

Zusätzlich ist jeweils ein NN mit obiger Netzwerk Architektur und Hyperparametern mit 300 Tupeln für Bigrame betrachtet worden. Die Tabelle 12.2 zeigt die Ergebnisse.

	GT	S	Sp	Pr neg	Pr pos	F1 neg	F1 pos	AUC
300	0.86	0.35	0.93	0.39	0.92	0.45	0.92	0.64
500	0.87	0.39	0.94	0.44	0.92	0.40	0.93	0.66
2	0.87	0.42	0.93	0.44	0.93	0.42	0.93	0.67

Tabelle 12.2: Ergebnisse des neuronalen Netzwerks für Bigrame

Durch die Betrachtung von Bigramen, anstelle von Unigramen, stellte sich eine Verschlechterung der Leistung des NN dar. Da sich durch Bigrame die Anzahl der Eingabewerte vergrössert und zusätzlich die Zahl der Features die keine Information liefern ansteigt, wird die Verschlechterung auf die Netzwerk Architektur zurückgeführt. Diese ist womöglich nicht in der Lage, den gesteigerten Umfang der Eingabewerte ausreichend zu approximieren. Es werden NN mit $\nu = (2, 3, 4, 5)$ verdeckten Schichten betrachtet. Es stellte sich heraus, dass ein NN mit zwei verdeckten Schichten eine Verbesserung der Leistung erbringt. Jedoch ist diese, trotz der Verbesserung, schlechter als die Leistung bei Unigramen. Durch die Verbesserung wird der Schluss gezogen, dass die höhere Anzahl an Eingabevariablen für den selben Satz eine klomplexere Architektur verlangt. Das Ergebnis kann in Zeile drei der Tabelle 12.2 eingesehen werden. Aufgrund dieses Ergebnisses kann gefolgert werden, dass eine zusätzliche Betrachtung von Bigramen keine Verbesserung der Klassifikation verursacht.

12.5 Neuronales Netzwerk mit Wortvektoren

Um die Wortvektoren in die Klassifikation der Hotelbewertungen mit einzubeziehen werden im Weiteren, WV statt die BOW Darstellung verwendet.

Die Wortvektoren des GloVe Modells werden verwendet. Ein NN mit einer verdeckten Schicht, dem stochastischen Gradientenverfahren mit Lernrate von 0.05 wird herangezogen. Um eine Überanpassung zu vermeiden wird die L2 Regularisierung verwendet. Der Trainingsdatensatz wird durch eine Zufallsauswahl von 500 Tupeln aus dem ursprünglichen Datensatz generiert. Die Tabelle 12.3 zeigt die Ergebnisse.

	GT	S	Sp	Pr neg	Pr pos	F1 neg	F1 pos	AUC
500	0.89	0.46	0.94	0.51	0.93	0.49	0.94	0.70

Tabelle 12.3: Ergebnisse des neuronalen Netzwerks mit Wortvektoren

Aus der Tabelle ist ersichtlich, dass durch die Einbeziehung der Wortvektoren, im Vergleich zu dem NN mit BOW die Gesamtleistung des Klassifikators gesteigert werden kann. Die Allgemeine Verbesserung wird auf die Leistung des Modells im Bereich der Klassifikation der negativen Hotelkritiken zurückgeführt. Die Leistung des Boosting und MaxEnt seinen an dieser Stelle wieder erwähnt. Diese stehen den Ergebnissen des hier betrachteten NN nicht nach. Betrachtet man den Aufwand der zur Erstellung des NN betrieben werden

muss kann gesagt werden dass unter Umständen das Boosting bzw. die MaxEnt bevorzugt werden können.

Kapitel 13

Weitere Neuronale Netzwerke

Dieses Kapitel soll einen Ausblick auf weitere NN geben.

13.1 Convolutional Neural Network

Convolutional Neural Networks (CNN) nutzen sich über die vorlgelagerte Schicht verschiebende Filter um lokale Features zu generieren . Diese Features werden in eine Pooling Schicht projiziert. Im Anschluss daran werden diese Features der Ausgabeschicht übergeben woraus die Ausgabewerte des Netwzwerks generiert werden. CNN können als Erweiterung der NN angesehen werden.

Verfahren nach Kim, 2014

Nachfolgend wird ein CNN auf Kim, 2014 erklärt.
Die Eingabematrix repräsentiert die n Wörter des aktuell betrachteten Satzes. Die einzelnen Wörter werden durch k-dimensionale WV nach dem CBOW Verfahren generiert, vgl. Mikolov et al., 2013a. Dazu werden die vordefinierten Wortvektoren von Word2Vec verwendet. Diese werden als statisch während des gesamten Trainings angesehen. Um im nächsten Schritt die lokalen Features c_i zu erstellen wird ein rollierender Filter bestehend aus h Wörtern verwendet und auf alle möglichen Fenster der vorgelagerten Schicht angewendet. Dabei generiert ein Filter ein Feature. Die daraus entstehenden Features lassen sich durch die Gleichung

$$c_i = f(wx_{i:i+h-1} + b) \tag{13.1}$$

darstellen. b entspricht dem Intercept. $f()$ repräsentiert eine nicht-lineare Funktion, z.B. den tangens hyperbolicus. w stellt den Gewichtsvektor dar

und $x_{i:i+h-1}$ die Wörter des aktuell betrachteten Fensters. Die Menge der daraus entstehenden Features entspricht der sogenannten Feature Map c, mit

$$c = \{c_1, c_2, \ldots, c_{n-h+1}\} \in \mathbb{R}^{n-h+1}. \tag{13.2}$$

Aus dieser Feature Map wird der maximale Wert, $\hat{c} = max\{c\}$ des jeweiligen Filters ausgewählt, dieser repräsentiert das Feature mit dem höchsten Informationsgehalt. Um eine grössere Anzahl an Features zu erreichen werden von den Autoren verschiedene Filter mit unterschiedlichen Anzahlen von Fenstergrössen h verwendet. Diese Features werden dann in die vollständig verknüpfte Ausgabeschicht übergeben. Dabei bedeutet vollständig dass jedes Element mit der Aktivierungsfunktion verbunden ist. Durch die Softmax Funktion wird dann die Wahrscheinlichkeitsfunktion der jeweiligen Labels ausgegeben. Um die Überanpassung zu verhindern wird das Dropoutverfahren verwendet, Srivastava et al., 2014. Um den finalen Feature Vektor $z = \{\hat{c}_1, \ldots, \hat{c}_m\}$ der Ausgabeschicht zu generieren werden zufällig ein Anteil von p Einheiten der verdeckten Schicht ausgelassen. Hieraus ergibt sich dann die Gleichung für die Zielvariable y,

$$y = w(z \times r) + b, \tag{13.3}$$

wobei der Vektor $r \in \mathbb{R}^m$ zur 'Maskierung' dient, dessen Einträge entsprechen Bernoulli Zufallsvariablen mit der Warscheinlichkeit p das eine eins auftritt. Durch die elementweise Multiplikation mit dem Vektor z fallen dann entsprechend die Features heraus. Die Gradienten werden nur für die noch existierenden Features berechnet.
Während der Testphase werden die Gewichte durch $\hat{w} = pw$ skaliert es existiert kein Wegfall von Gewichten. Zusätzlich wird für jede Iteration des Gradienten Verfahren angenommen dass $||w||_2 = s$ falls $||w||_2 > s$, wobei der Hyperparameter s eine obere Grenze der Aktualisierung der Gewichte darstellt und $||w||_2$ der l_2-Norm. Dies kann die Leistung eines CNN verbessern, da durch eine Begrenzung hohe Werte für die Lernraten gewählt werden können ohne dass sich die Gewichte aufblähen, vgl. Srivastava et al., 2014 und Srivastava et al., 2012. Ebenfalls verringert sich der Rechenaufwand mit sinkendem p.
Die Hyperparameter des Modells werden wie folgt gewählt, Fenstergröße $h = 3, 4, 5$ der drei Filter mit jeweils 100 Feature Maps. Die Fenstergröße ist äquivalent, dass N-Grame bestehend aus drei, vier oder fünf Wörtern betrachtet werden. Die Ausfallrate entspricht $p = 0.5$. $s = 3$ und die mini-batch Größe für das Gradienten Verfahren enspricht 50. Für das Gradienten Verfahren wurde die Adadelta Regel nach Zeiler, 2012 gewählt. Laut Kim, 2014

liefert das AdaGrad Verfahren nach Duchi et al., 2011 ähnliche Ergebnisse bei einer höheren Anzahl an Iterationen. Als Aktivierungsfunktion wird die ReLu Funktion gewählt, mit $f(x) = log(1 + exp(x))$.

Kim, 2014 wendet das oben erklärte Verfahren an um die Polarität (negativ/positiv) von Filmbewertungen zu schätzen. Dabei besteht jede Filmbewertung aus einem Satz, vgl. Pang et al., 2002. Durch dieses Modell kann eine Gesamttrefferrate von 81% erzielt werden.
Kim, 2014 verwenet noch drei weitere Modelle um die Polarität der Filmbewertungen zu prognostizieren, (1) ein Modell dessen Wortvektoren zuerst zufällig generiert werden und während des Trainings modifiziert werden, hierdurch ergibt sich eine Trefferrate von 76.1%. (2) Für dieses Modell werden ebenfalls die Vordefinierten WV verwendet und dann für jeden Korpus präzise abbgestimmt. Dieses Modell kann die höchste Gesamttrefferrate von 81.5% aufwiesen. (3) Ein Modell bestehend aus zwei Mengen an WV als Input. Jeder Filter wird auf jede der beiden Mengen angewendet. Wobei die Backpropagation nur auf eine der beiden Mengen angewendet wird. In der anderen Menge werden die WV speziell angepasst, eine Gesamtrefferrate von 81.1% kann hierfür ausgewiesen werden. Auf dieses Verfahren wird nicht näher eingegangen, da hierdurch keine neuen Erkenntnisse gewonnen werden können und sich die Gesamttrefferrate der Modelle nicht wesentlich ändert im Vergleich zu dem Modell mit statischen WV.

Die Ergebnisse lassen den Schluss zu, dass durch CNN relativ gute Ergebnisse im Bereich der Klassifikation von Kritiken erzielt werden kann.

Verfahren nach Dos Santos and Gatti, 2014

Im folgenden wird ein von Dos Santos and Gatti, 2014 vorgeschlagenens Deep CNN zur Prognose der Polarität $\tau \in T$ eines Satzes kurz erklärt, dabei wird jedoch nicht auf die detaillierte Darstellung eingegangen, für eine tiefergehende Ansicht sei an dieser Stelle auf das Paper verwiesen.
Dos Santos and Gatti, 2014 ziehen ein Depp CNN heran um die Polarität von Sätzen, bestehend aus N Wörtern, zu schätzen. Dabei wird die Schätzung der WV $u_n = \left[r^{wrd}; r^{chr}\right]$ des Wortes $n = 1, \ldots, N$ nach Mikolov et al., 2013a als erste Schicht angesehen. Wobei der WV aus zwei Teilvektoren besteht. Zum einen aus dem WV $r^{wrd} \in \mathbb{R}^{d^{wrd}}$ und zum anderen aus der vektoriellen Darstellung der Buchstaben, $r^{chr} \in \mathbb{R}^{cl_u^0}$. Auf die Darstellung des zweiten Teilvektors wird hier nicht näher eingegangen. Zur Schätzung der WV wird das SKG Modell mit der Fenstergröße neun verwendet. Als Korpus wird eine Momentaufnahme des englischen Wikipedia Korpus von 2013 herangezogen.

Die nachgelagerte Schicht repräsentiert ein einschichtiges CNN, hierbei wird der Vektor $z_n = (u_{n-(k-1)/2}, \ldots, u_n, \ldots, u_{n+(k-1)/2})^T$, mit $z_n \in \mathbb{R}^{(d^{wrd}+cl_u^0)k}$ und k einer beliebigen Fenstergröße durch rollierende Filter generiert.
Die Features $[r^{sent}]_j = \max_{1 \leq n \leq N} [W^1 z_n + b^1]$ werden in den Vektor r_x^{sent} zusammengefasst. Wobei die Gewichtsmatrix durch $W^1 \in \mathbb{R}^{cl_u^1 \times (d^{wrd}+cl_u^0)k}$ gegeben ist. Mit cl_u^1 der Dimension der Satz Ebenen Features. $b^1 \in \mathbb{R}^{cl_u^1}$ repräsentiert den Intercept.
Anschliessend wird dieser Vektor einem zweischichtigen NN übergeben. Dieses errechnet anschließend den Score $s(x) = W^3 h(W^2 r_x^{sent} + b^2) + b^3$ der Labels. Die Gewichtsmatrizen sind gegeben durch $W^3 \in \mathbb{R}^{hl_u \times cl_u^1}$ und $W^3 \in \mathbb{R}^{|T| \times hl_u}$, wobei hl_u der Anzahl der Einheiten der verdeckten Schicht entspricht. $b^2 \in \mathbb{R}^{hl_u}$ und $b^3 \in \mathbb{R}^{|T|}$ entsprechen auch hier wieder dem Intercept. Die Aktivierungsfunktion entspricht der tangens hyperbolicus Funktion. Die Ausgabewerte des NN werden durch die Softmax Funktion errechnet. Der Ausgangspunkt des Trainings des Netzwerks ist die Log-Likelihood Funktion. $log(p(\tau|x, \Theta)) = s_\theta(x)_\tau - log(\sum_{\forall i \in T} exp(s_\theta(x)_i))$, mit Parametern des Netzwerks θ . Zur Optimierung der negativen Log-Likelihood Funktion $\theta \rightarrow \sum_{(x,y) \in D} -log(p(y|x, \theta))$ wird das stochastische Gradientenverfahren herangezogen. Wobei y der Polarität des aktuell betrachteten Satzes entspricht des zum Training herangezogenen Korpus D entspricht.

Dieses Modell wird auf den Korpus der Stanfort Sentiment Treebank angewendet. Dieser Korpus enthält Sätze aus Filmkritiken und deren Polarität. Für die jeweilige Bewertung ist die Polarität angegeben, dabei wird für die Schätzung die Klasse der neutralen Bewertungen entfernt und die Kategorien der positiven bzw. negativen Bewertungen jeweils zu einer Klasse zusammengefasst. Der Testdatensatz bestand immer aus vollständigen Sätzen. Zum Training des Netzwerks werden einmal alle Kritiken zugelassen welche auch unvollständige Sätze enthalten hier konnte eine Gesamttrefferrate von 85.5% erreicht werden. Andererseits werden nur vollständige Sätze in den Trainingsdatensatz aufgenommen hier ergibt sich eine Gesamttrefferrate von 82.0%. Dos Santos and Gatti, 2014 beobachten dass ein besseres Ergebnis erzielt werden kann wenn auch unvollständige Sätze verwendet werden. Dies kann an der Tatsache liegen, dass unvollständige wichtige Informationen enthalten, z.B. ein Wort dass nur in unvollständigen Sätzen beobachtet wird und ein starkes Feature zur Klassifikation darstellt. Wird dieses Wort nicht aufgenommen kann dies nicht durch das Netzwerk erlernt werden und somit nicht in der Testphase verwendet werden.

Desweiteren wird das Modell von Dos Santos and Gatti, 2014 mit einem nach

Socher et al., 2013 SVM Modell zur Klassifikation der Kritiken verglichen. Dieses beinhaltet ebenfalls unvollständige Sätze im Trainingsdatensatz. Hierdurch wird eine Gesamttrefferrate von 79.4% erzielt. Dieses Ergebnis lässt die Behauptung zu dass das verwenden von komplexeren Methoden zur Klassifikation eine Steigerung der Gesamttrefferrate von ca. 3% bzw. 6% bewirken kann.

Vergleicht man die Netzwerke nach Kim, 2014 und Dos Santos and Gatti, 2014. So kann gesagt werden dass durch Anwendung von unterschiedlichen Ansätzen unterschiedlich gute Ergebnisse erzielt werden. In diesem Fall kann durch eine komplexere Modellstruktur ein besseres Ergebnis erzielt werden. Jedoch stellt sich Verbesserung des Ergebnisses im Falle der Aufnahme von ausschließlich vollständigen Sätzen als sehr gering dar. In der Situation dass ebenfalls unvollständige Sätze involviert werden ist die Verbesserung um fast 5% zu erreichen.

13.2 Recursive Neural Network

Socher et al., 2013 stellen zwei weitere Netzwerke zur Klassifikation von Hotelkritiken vor, das Recursive Neural Network (RNN) und das Recursive Neural Tensor Network (RNTN). Zur Klassifikation werden Binärbäume verwendet. Beispielhaft zeigt die nachfolgende Grafik 13.1 einen Ausschnitt eines RNN, die Grafik wurde Socher et al., 2013 entnommen. Die Blätter sind durch die vektorielle Darstellung der Wörter eines Satzes gegeben. Die jeweiligen Knoten oder Elternvektoren werden durch die Komposition von je zwei Vektoren gebildet. Dabei nimmt die Kompositionsfunktion die Stelle der Aktivierungsfunktion ein. Jeder Knoten kann als einzelne Schicht angesehen werden, dabei ist die Anzahl der Schicht von der Anzahl der Wörter des jeweilig betrachteten Satzes abhängig. Es werden solange Pärchen gebildet, bis schließlich keine Pärchen mehr gebildet werden können. Auf die letzte Schicht wird dann wiederum durch die Anwendung der Softmax Funktion die Prognose des Netzwerks ausgegeben.

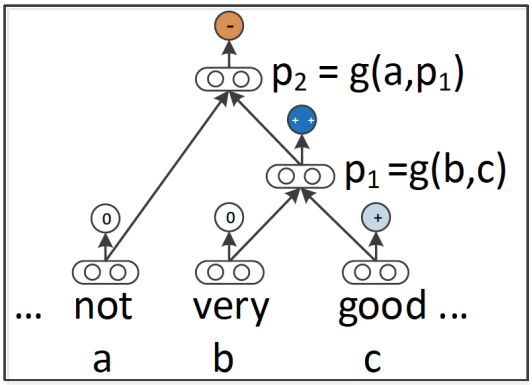

Abbildung 13.1: RNN nach Socher et al., 2013

Die Vektoren der verdeckten Schichten werden dann wie folgt berechnet,

$$p_1 = f\left(W \begin{bmatrix} b \\ c \end{bmatrix}\right), \text{und } p_2 = f\left(W \begin{bmatrix} a \\ p_1 \end{bmatrix}\right) \tag{13.4}$$

Die Paramertermatrix ist gegeben durch $W \in \mathbb{R}^{d \times 2d}$. d enspricht der Dimension der WV. Zusätzlich wird jeder verdeckte Vektor durch die Softmax Funktion bewertet.
Das Modell wird mithilfe von Recursive Autoencoders (RAE) trainiert, vgl. Socher et al., 2013, Socher et al., 2011a, Socher et al., 2011b und Huang, 2011. Nachfolgende Grafik dient der kurzen Erklärung der Anwendung der RAE an einem Binärbaum, diese Grafik wurde Socher et al., 2011a entnommen.

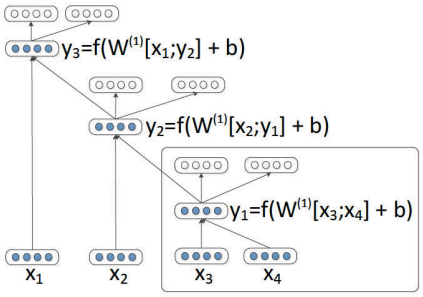

Abbildung 13.2: Beispiel der RAE an einem Binärbaum

An jedem Knotenpunkt wird die selbe Gewichtsmatrix $W^{(1)}$ verwendet um die Elternvektoren zu berechnen. f entspricht einer Aktivierungsfunktion und b dem Biasterm der jeweiligen Schicht. Das Verahren sei innerhalb des Umrandeten Bereichs kurz erklärt, dies stellt das Standard Verfahren dar, vgl. Socher et al., 2011a. Der Elternvektor y_1 ist durch die Komposition der beiden Kinder x_3 und x_4 gegeben. Und der Elternvektor y_2 durch die Komposition der beiden Kinder y_1 und x_2. Und Rückgerichtet soll y_1 durch das Paar x_1 und y_2 angegeben werden können. Die blau ausgefüllten Schichten repräsentieren dabei die verdeckten Schichten und die nicht ausgefüllten Schichten die Rekonstruktionsschicht. Basierend auf der gegebenen Baumstruktur wird der Elternvektor durch,

$$y_1 = f\left(W^{(1)}\left[c_1; c_2\right] + b^{(1)}\right), \tag{13.5}$$

berechnet. $(c_1, c_2) = (x_3, x_4)$ entsprechen den Kindern.

Und die Rekonstruktion durch

$$\left[c_1^{'}; c_2^{'}\right] = f\left(W^{(2)}p + b^{(2)}\right), \tag{13.6}$$

wobei p den Elternvektor bezeichnet, berechnet.

Der Rekonstruktionsfehler von $[c_1; c_2]$ und $\left[c_1^{'}; c_2^{'}\right]$ ist gegeben durch die euklidische Distanz,

$$E_{rec}([c_1; c_2]) = \frac{1}{2}|| \left[c_1; c_2\right] - \left[c_1^{'}; c_2^{'}\right] ||^2. \tag{13.7}$$

Darauffolgend werden dann die Kinder der nachfolgenden Schicht neu definiert mit $(c_1, c_2) = (x_2, y_1)$. Diese Schritte wiederholen sich für alle Paare des

Baumes, bis die Fehler aller inneren Knoten berechnet sind. In der Trainingsphase werden dann die aufsummierten Rekonstruktionsfehler der einzelnen Knoten minimiert. Der Fehler kann ebenfalls durch z.B. die Kreuzentropie berechnet werden, da an jedem inneren Knoten die Klassen-Wahrscheinlichkeit durch die Softmax Funktion ausgegeben wird, vgl. Socher et al., 2013. Dieses Verfahren erfordert eine feste Struktur des Baumes.
Im Falle des verwendeten Modells wird eine feste Baumstruktur angenommen und der Rekonstruktionsfehler ignoriert.
Dieses Modell erreichte bei der binären Klassifikation der Filmewertungen eine Gesamttrefferrate von 82.4%.

13.3 Recursive Neural Tensor Network

Das zweite Modell das durch Socher et al., 2013 vorgestellt wird ist das RNTN, dieses ist eine Erweiterung des RNN.
Bei RNN werden die Interaktionen der paarweisen Nachbarn über die Aktivierungsfunktion aufgenommen. Das RNTN lässt durch die zusätzliche Aufnahme einer für alle Knoten gleichen tensorbasierten Kompositionsfunktion mehr Interaktionen zwischen den Wörtern zu. Dieser Unterschied lässt sich wie folgt darstellen.

$$p_1 = f\left(\begin{bmatrix} b \\ c \end{bmatrix}^T V^{[1:d]} \begin{bmatrix} b \\ c \end{bmatrix} + W \begin{bmatrix} b \\ c \end{bmatrix}\right), \tag{13.8}$$

$$p_2 = f\left(\begin{bmatrix} a \\ p_1 \end{bmatrix}^T V^{[1:d]} \begin{bmatrix} a \\ p_1 \end{bmatrix} + W \begin{bmatrix} a \\ p_1 \end{bmatrix}\right). \tag{13.9}$$

Wobei $V^{[1:d]} \in \mathbb{R}^{2d \times 2d \times d}$ dem Tensor der die multiblen bilinearen Formen enthält entspricht. Durch diesen können die Interaktionen der Wörter untereinander 'direkt' modelliert werden und nicht nur der Effekt der vorgelagerten Wörter auf den Elternvektor, wie beim RNN. Das RNN ist ein Spezialfall des RNN mit $V^{[1:d]} = 0$.
In der Trainingsphase wird die Kreuzentropie minimiert. Da an jedem inneren Knoten die Wahrscheinlichkeit der Klassenzugehörigkeit durch die Softmax Funktion ausgegeben wird, vgl. Socher et al., 2013.

Mithilfe des RNTN kann eine Gesamttrefferrate von 85.4% bei der Klassifikation (positiv / negativ) von Filmbewertungen erzielt werden.
Es stellt sich heraus, dass durch diese beiden Ansätze die Leistung bei der

Klassifikation von Texten eine weitere Verbesserung der Gesamttrefferrate erzielt werden kann.

Kapitel 14

Diskussion

In den ersten Kapiteln werden sprachwissenschaftliche Einflüsse auf die Orientierung einer Aussage beleuchtet. Dabei zeigt sich dass neben Nomen und Verben besonders Adjektive für die Polarität verantwortlich sind. Auch kann bei gleicher Kernaussage eines Satzes festgestellt werden das ein anderer Wortlaut, oder eine geänderte Satzstellung die Intensität sowie die Richtung der Orientierung des Satzes drastisch beeinflussen können. Zu betonen ist die Änderung der Intensität und der Richtung einer Aussage durch hinzufügen eines negierenden Wortes. Weiter stellt sich heraus, dass der Kontext bzw. die Domain die gleiche Aussage in einem anderen Licht erscheinen lässt.
Um diesen Herausforderungen zu begegnen werden in dieser Arbeit eine Reihe an verschiedenen Ansätzen vorgestellt.

Zur Sentiment Analyse werden als frühe Methoden Sentiment Lexika und die semantische Orientierung herangezogen. Im Rahmen dieser Arbeit zeigt sich, dass die Erstellung eines Lexikons sehr zeitaufwendig ist und z.T. subjektives Eingreifen erforderlich ist. Die semantische Orientierung ist von der gewählten Suchmaschine abhängig.
Beispielsweise ist die Erstellung der Referenzliste oder die Polarisierung einzelner Wörter von einzelnen Personen abhängig und kann je nach Person variieren. Dies kann die resultierenden Lexika und die semantische Orientierung stark beeinflussen. Wird z.B. ein Wort der Referenzliste anders polarisiert kann dies zur Konsequenz haben, dass eine ganze Reihe Wörter durch ein automatisiertes Verfahren zur Erstellung eines Lexikons anders eingestuft wird. Erstellt man nun zwei Lexika basierend auf Referenzlisten mit gleichen Wörtern jedoch einer unterschiedlichen Polarität entstehen zwei Lexika die für die gleichen Wörter unterschiedliche Polaritäten angeben. Berechnet man nun die sematische Orientierung der Einträge können hierdurch andere Ergebnisse entstehen. Bei der Analyse einer Reihe identischer Texte kann dies

zu starken Abweichungen der Einstufung der Polarität führen.
Eine Möglichkeit zur Reduktion des subjektiven Einflusses kann ein kombiniertes Lexikon sein. Dabei wird basierend auf mehreren Referenzlisten für jede Liste nach mehreren Verfahren jeweils ein Lexikon erstellt. Bei einem Vergleich der Einträge der Lexika wird dann der Eintrag gewählt für den sich die häufigsten Übereinstimmungen ergeben und dem kombiniertem Lexikon zugeteilt. Darauf basierend kann dann die semantische Orientierung der einzelnen Einträge berechnet werden. Oder aus diesen Einträgen eine Referenzliste erstellt werden, die zur Berechnung der semantischen Orientierung dient.
Auch führt die Wahl verschiedener Korpora zu unterschiedlichen Lexika und semantischen Orientierungen der Einträge. Dies kann umgangen werden, indem verschhiedene Datenbanken oder Korpora zur Erstellung eines Lexikons herangezogen werden. Dies umgeht gleichzeitig das Problem, dass Fehlinterpretationen der Polarität eines Textes stattfinden, durch z.B. nicht existieren eines relevanten Begriffs. Ebenfalls steigert die Hinzunahme von weiteren Daten die domainspezifische Flexibilität der Lexika. Auch hier ist eine Erstellung eines kombinierten Lexikons eine Lösung.

Spätere Verfahren verwenden statistische Modelle und die BOW Darstellung der Texte. Durch die BOW Darstellung und der Verwedung von DTM als Eingabewerte für die unterschiedlichen statistischen Modelle, sowie das überwachte Lernen reduziert sich der subjektive Einfluss. Die Texte werden nicht mehr von einer Person klassifiziert, sondern die Information wird den Daten entnommen. Da die Autoren der Kritiken die Polarität selbst festlegen.
Das verwenden von DTM steigert zusätzlich die Flexibilität der Domainspezifität, indem als Korpus zum einen die Kritiken selbst herangezogen werden und durch Hinzunahme von weiteren Kritiken aus anderen Themenbereichen, erweitert werden können. Auch bietet die Verwendung von weiteren Datensätzen bestehend aus Nachrichten o.ä. eine Möglichkeit das Vokabular zu erweitern und die Einsetzbarkeit auf andere Bereiche zu erhöhen. Desweiteren wird dadurch der Verlust an Informationen innerhalb der Daten verrringert, da das Vokabular ansteigt und somit mehr Wörter erkannt und in die Analyse miteinbezogen werden können.
Die Einbeziehung der N-Grame bei der Erstellung der DTM bietet die Möglichkeit den Einfluss von Wortkombinationen in die Modellierung mit aufzunehmen.
Zudem ist ein Vorteil dass durch Hinzunahme von Hilfsvariablen z.B. das Alter der Autoren oder anderen Merkmalen die Untersuchung der Einflüsse auf die Polarität erweitert werden kann. Hierdurch könnten sich die Schätzun-

gen der statistischen Methoden verbessern. Weiter können durch Variablen Selektion irrelevante Variablen ausgeschlossen werden.
Die in dieser Arbeit als modell- und algorithmusbasierte bezeichneten Verfahren erwiesen sich als mittelmäßige bis gute Klassifikatoren. Jedoch ist zu erwähnen, dass diese ihr Potenzial nicht voll ausschöpfen können. Dies wird auf die Ähnlichkeit der Kritiken unterschiedlicher Orientierungen zurückgeführt. Beispielsweise unterscheiden sich sonst gleiche Kritiken nur in einem Wort, welches eine Umkehrung der Polarität provoziert. Weiter werden in einer Kritik unterschiedliche Polaritäten zu unterschiedlichen Aspekten angegeben, somit spielen die Präferenzen der Autoren eine große Rolle. Dadurch können zwei identische Kritiken eine unterschiedliche Polarität aufweisen. Weiter sind hier die bereits thematisierten sprachwissenschaftlichen Phänome zu erwähnen, welche die Trennung in zwei eindeutig unterscheidbare Lager erschweren. Vielmehr liegt die Vermutung nahe, dass die Kritiken mehrere Felder innerhalb der Daten bilden welche Feldintern, nach positiv und negativ getrennt werden müssten. Hierzu sind diese Modelle jedoch nicht nicht in der Lage. Auch wird die Likelihood mit steigender Gleichheit der Daten sehr flach. Hierdurch kann die Identifikation eines globalen Maximums erschwert werden oder nur schwache Effekte geschätzt werden.
Auch ist die große Anzahl an Variablen der Modelle zu erwähnen, aufgrund der Fülle an Möglichkeiten ein und dieselbe Aussage zu treffen beinhalten die Modelle oft zu viele Variablen. Dies fürt dazu dass unter umständen sehr viele schwache Effekte vorliegen. Eine Mögliche Lösung könnte sein die Parameter zu einem starken Effekt zusammenzufassen, indem für Synonyme eine repräsentative Bezeichnung verwendet wird. Auch beziehen die Modelle wenig informative Effekte mit ein, die aufgrund der Häufigkeit der Beobachtungen als starke Effekte erscheinen, dies wird z.B. durch die Extraktion von Stoppwörtern umgangen.
Desweiteren tragen die bereits diskutierten Nachteile durch die BOW Basis der Modelle dazu bei dass diese bei der Trennung der Daten nicht ihr volles potenzial ausschöpfen können.

Als aktuellste Verfahren zur Klassifikation sind die NN in Kombination mit Wortvektoren zu erwähnen. Wortvektoren sind im Gegensatz zu BOW in der Lage Beziehungen in den Daten aufzufassen und in die Modellierung miteinfließen zu lassen. Zum anderen sind NN aufgrund deren Flexibilität ein starkes Handwerkzeug zur Klassifikation von Kritiken. Vorteilhafte Ergebisse werden bei der Kombination von Wortvektoren und NN durch die Annahme der Approximation einer nicht linearen Funktion erzielt. Weiter sind NN fähig die Felder innerhalb der Daten zu erkennen und während der Modellierung zu berücksichtigen. Dies wird z.B. durch die Feature Extrakti-

on der CNN in die Modellierung einbezogen.
Ein besonderes Maß an Aufmerksamkeit ist dem raschen Ansteigen der Komplexität der NN zu widmen. Die hohe Anzahl an Daten die verwendet werden müssen um Wortvektoren und NN zu trainieren ist ebenfalls nicht ausser Acht zu lassen. Beispielsweise sind Wortvektoren nur in der Lage die semantische und synthaktische Beziehungen akurrat zu erkennen wenn eine große Menge an Eingabewerten vorliegt. Die Schätzungen von Wortvektoren sind sehr von der Datenbasis abhängig. Das kann dazu führen dass bei unzureichender oder veränderten Datenbasis Zusammenhänge nicht ausreichend angegeben werden. Werden diese Wortvektoren dann zur Klassifikation durch NN verwendet kann dies, abhängig von der Domain, zu Fehlklassifikationen führen. Auch ist die Gefahr zu erwähnen dass durch Lösungsverfahren nur ein lokales Optimum gefunden wird was zu Fehlklassifikationen führen kann. Je flacher die Likelihood, desto höher die Gefahr.
Die gewählte Netzwerk Architektur sowie die Kombination aus Aktivierungsfunktionen, Optimierungsverfahren und die Wahl der Hyperparameter, wie z.B. die Wahl der Dimension der Wortvektoren oder die Lernrate können ebenfalls zu sehr unterschiedlichen Ergebnissen führen. Im Extremfall dazu, dass die ideale Kombination nicht gefunden wird. Dies gilt auch für die Schätzung der Wortvektoren,z.B. durch die Wahl der idealen Anzahl an Kontextwörter.

Als eine mögliche Erweiterung der NN könnte das Zulassen von höher dimensionierten Funktionen als Verbindungen der Neuronen einen Vorteil verschaffen.

Um ein Resume, basierend auf den Ergebnissen dieser Arbeit, zu ziehen, kann gesagt werden das die Anwendung der NN in Verbindung mit WV, die MaxEnt und das Boosting am besten zur Klassifikation der Hotelbewertungen geeignet sind. Besonders sind die Fortschritte im Bereich der Klassifikation der negativen Kritiken, hervorzuheben.
Die vorteilhaften Ergebnisse können durch verschiedene Ansätze der NN wie CNN noch verbessert werden.

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt, noch nicht in einer anderen Prüfungsbehörde vorgelegt und noch nicht veröffentlicht habe.

München, den	Greber, Andre

Literaturverzeichnis

C. Aggarwal and C. Zhai. *Mining Text Data*. Springer, 2012.

A. L. Berger, S. A. Della Pietra, and V. Della Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 1996.

L. Breiman. Bagging predictors. *Machine Learning*, 1996.

K. W. Church and P. Hanks. Word assoziation norms, mutual information and lexicography. *Proceedings of the 27th Annual Conference of the ACL*, 1989.

D. R. Cutler, T. C. Edwards, K.H. Beard, A. Cutler, K.T. Hess, J.C. Gibson, and J. J. Lawler. Random forests for classification in ecology. *Ecology*, 88.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 1990.

M. Dettling and P. Bühlmann. Boosting for tumor classification with gene expression data. *Bioinformatics*, 2003.

C. N. Dos Santos and M. Gatti. Deep convolutional neural networks for sentiment analysis of short texts. *COLING*, 2014.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research 12*, 2011.

L. Fahrmeir, Hamerle. A., and G. Tutz. *Multivariate statistische Verfahren*. de Gruyter, 1996.

L. Fahrmeir, T. Kneib, and S. Lang. *Regression - Modelle, Methoden und Anwendungen*. Springer Verlag, 2009.

C. Fellbaum. Wordnet: an electronic lexical database. *EUA*, 1998.

Y. Freund and R. E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Science*, 1997.

J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 2000.

D. Goldhahn, T. Eckart, and U. Quasthoff. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. *LREC*, 2012.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. Springer, 2008.

V. Hatzivassiloglou and K. R. McKeown. Predicting the semantic orientation of adjectives. *Proceedings of the 35th Annual Conference of the ACL and the 8th Conference of the European Chapter of the ACL*, 1997.

M. Hu and B. Liu. Mining and summarizing costumer reviews. *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004.

E. Huang. Paraphrase detection using recursive autoencoder. *Source:[http://nlp. stanford. edu/courses/cs224n/2011/reports/ehhuang. pdf]*, 2011.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, 2013.

Y. Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

S. Kiritchenko, X. Zhu, and S. M. Mohammad. Sentiment analysis of short informal texts. *Journal of Articel Intelligence Research*, 50:723–762, 2014.

U. Ligges. *Programieren mit R*. Springer Verlag, 2003.

B. Liu and L. Zhang. A survey of opinion mining and sentiment analysis. *Aggarwal C. C., Zhai C. (Eds.), Mining Text Data, pp. 415-463. Sringer US*, 2012.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv:1301.3781*, 2013a.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems.*, 2013b.

B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information and Retrieval*, 2 (1-2):1–135, 2008.

B. Pang, L. Lee, and S. Vaithyanathan. Thumps up? sentiment classification using machine learning techniques. *Proceedings of EMNLP*, 2002.

J. Pennington, R. Socher, and C.D. Manning. Glo ve: Global vectors for word representation. *EMNLP*, 2014.

L. Polanyi and A. Zaenen. Contextual valence shifters. *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*, 2004.

R. Remus, U. Quasthoff, and G. Heyer. Sentiws - a publicly available german-language resource for sentiment analysis. *Proceedings of Language Resources and Evaluation*, 2010.

Lothar Sachs and Jürgen Hedderich. *Angewandte Statistik - Methodensammlung mit R.* Springer Verlag, 2009.

R. E. Schapire. Strength of the weak learnability. *Machine Learning*, 1990.

A. Schiller, S. Teufel, C. Thielen, and C Stöckert. *Guidlines für das Tagging deutscher Textcorpora mit STTS.* 1999.

R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. *Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics*, 2011a.

R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. *NIPS. Vol. 24.*, 2011b.

R. Socher, J. Perelygin, J. Wu, J. Y. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, 2013.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaption of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 2014.

J. A. K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 1999.

H. Toutenbug and C. Heumann. *Deskriptive Statistik - Eine Einführung in Methoden und Anwendungen mit R und SPSS*. Springer Verlag, 2008.

P. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. *Proceedings of Annual Meeting if the Association for Computational Linguistics (ACL 2002)*, 2002a.

P. Turney. Unsupervised learning of semantic orientation from a hundered-billion-words corpus. *National Research Council, Institute for Information Technology, Technical Report ERB-1094*, 2002b.

Tutz, Künstler, Pigeot, and Fahrmeir. *Statimiktik – Der Weg zur Datenanalyse*. Springer Verlag, 2007.

H. Wickham. *ggplot2 Elegant Graphics for Data Analysis*. Springer Verlag, 2009.

T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005.

M. Wolfgruber. *Sentiment Analyse mit lokalen Grammatiken - Wissenschaftlicher Ansatz zur Extraktion von Sentiments in Hotelbewertungen*. Universitätsbibliothek der Ludwig-Maximilians-Universität, 2015.

M. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

C. Züll and P. Mohler. Der general inquirer iii: Ein dinosaurier für die historische forschung. *ZUMA-Arbeitsbericht 1989/10*, 1989.